Metadata Standardization and Collection Based on the Ontology for Biomedical Investigations (OBI)

Jie Zheng and Chris Stoeckert

University of Pennsylvania

November 3, 2014

Outline

- Introduction of OBI
  - OBI scope and purpose
  - High level overview
  - Semantic representation of assay and investigation
- Application of OBI
  - NIAID Metadata standardization
  - ICEMR project: protein array data
  - ICEMR project: PRISM studies

How can an ontology help with data integration and retrieval?

- Heterogeneous data integration
  - requires consistent annotation
- Retrieval of data of the same kind across multiple databases
  - requires shared semantics
  - requires consistent and unambiguous annotation
- Enhance data retrieval using logical inferences
  - Automatic inference that a person who is *Plasmodium* parasite positive based on a laboratory test and a body temperature > 38^{o}C degree has malaria
  - Automatic inference that a man who is 6 feet tall and 210 pounds is overweight based on BMI
  - These inferences are simple to humans. But the logical rules need to be specified in order to make the inferences using a computer.

The logical rules are in ontologies and NOT in data dictionaries (terminology)

Ontology for Biomedical Investigations (OBI)

Ontology for Biomedical Investigations

- OBI is about capturing all aspects of a biological and clinical investigation (investigation, assay, specimen, protocol, device, data, data analysis, etc.)
- Things to know about OBI
  - a member of the OBO Foundry
  - interoperability with other ontologies following OBO Foundry principles, such as the Gene Ontology (GO)
  - uses the Basic Formal Ontology (BFO) as its top level ontology
  - uses the Information Artifact Ontology (IAO) for general information entities
- Details on OBI can be found at:
  - http://obi-ontology.org
  - J Biomed Semantics. 2010. Modeling biomedical experimental processes with OBI, Ryan R Brinkman, Mélanie Courtot, Dirk Derom, Jennifer M Fostel, Yongqun He, Phillip Lord, James Malone, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Larisa N Soldatova, Christian J Stoeckert, Jr., Jessica A Turner, Jie Zheng, and the OBI consortium

OBO Foundry

Perspective

Nature Biotechnology **25**, 12
Published online: 7 November

The OBO Foundry: support biomedical

Barry Smith[1], Michael Ash
William Bug[5], Werner Ceu
Ireland[9], Christopher J Mu
Philippe Rocca-Serra[9], Ala
Richard H Scheuermann[14]
Lewis[10]

The value of any kind of form that allows it to be integration is through th common controlled voca success of this approach itself creates obstacles (OBO) consortium is pur Existing OBO ontologies coordinated reform, and of an evolving set of sha The result is an expanding family of ontologies designed t interoperable and logically well formed and to incorporate accurate representations of biological reality. We describe this OBO Foundry initiative and provide guidelines for those who might wish to become involved.

OBO Foundry ontologies

Title	Domain	Prefix	File
Biological process	biological process	GO	go.obo
Cellular component	anatomy	GO	go.obo
Chemical entities of biological interest	biochemistry	CHEBI	chebi.obo
Molecular function	biological function		
Ontology for biomedical investigations	experiments		
Phenotypic quality	phenotype	PATO	quality.obo
Plant Ontology	anatomy and development	PO	plant_ontology.obo?view=co
PRotein Ontology (PRO)	proteins	PR	pro.obo
Xenopus anatomy and development	anatomy	XAO	xenopus_anatomy.obo
Zebrafish anatomy and development	anatomy	ZFA	zfa.obo

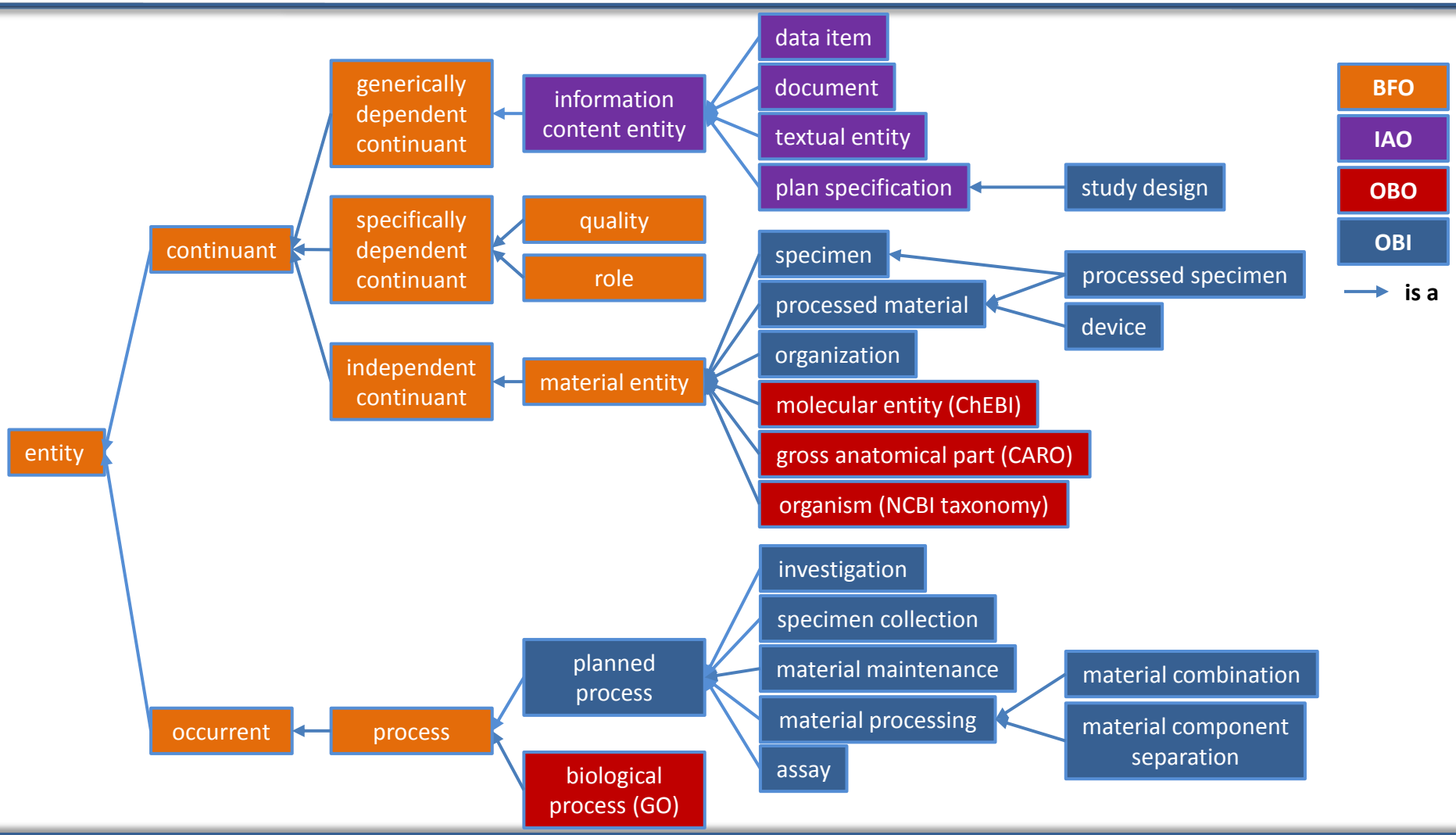
Ten OBO Foundry ontologies

OBO Foundry candidate ontologies and other ontologies of interest

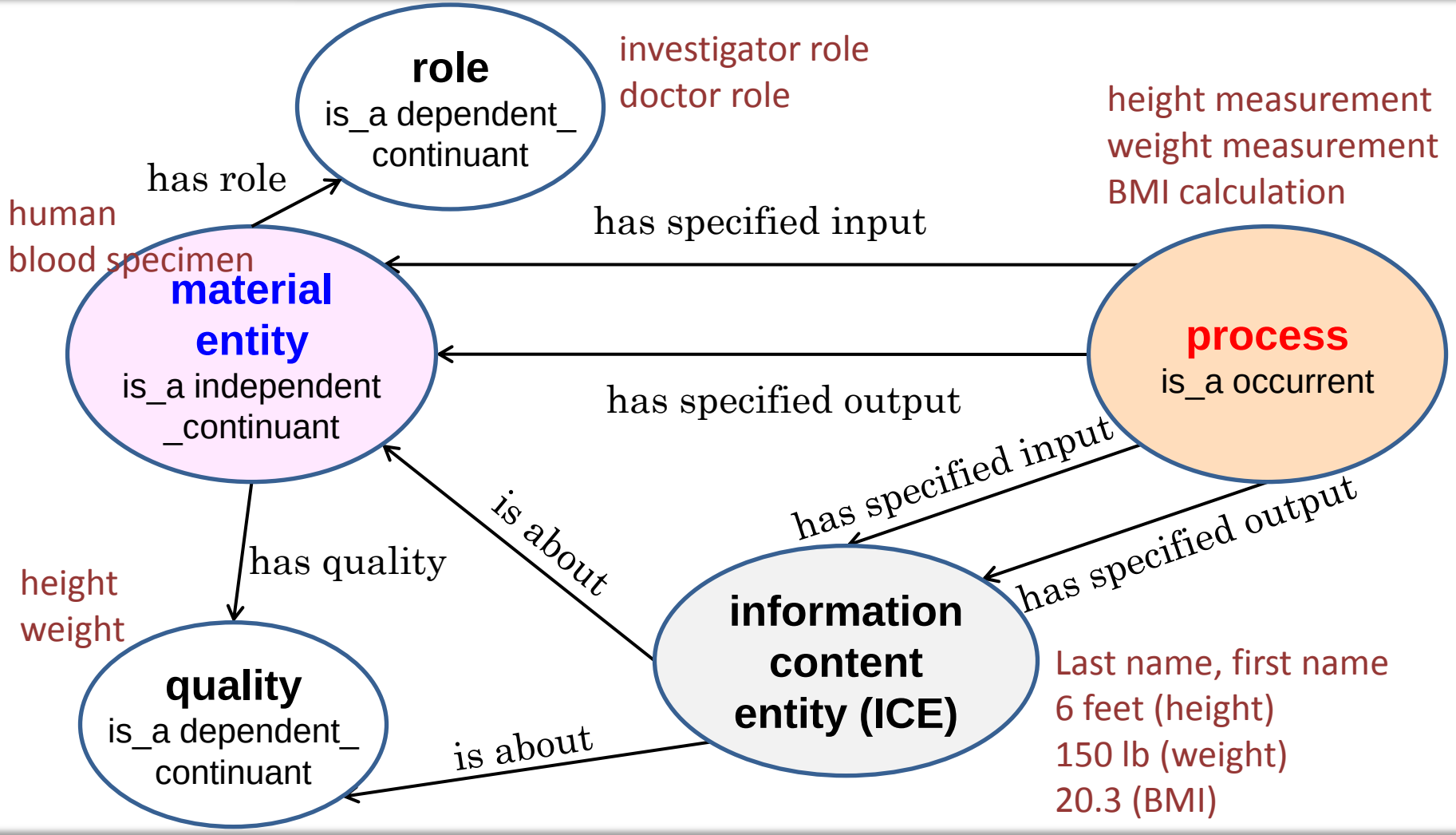
Title	Domain	Prefix	File
Adverse Event Reporting Ontology	health	AERO	aero.owl
Anatomical Entity Ontology	anatomy	AEO	aeo.obo
Ascomycete phenotype ontology	phenotype	APO	ascomycete_phenotype.obo
Basic Formal Ontology	upper	BFO	1.1
Beta Cell Genomics Ontology		BCGO	bcgo.owl
Biological Collections Ontology		BCO	bco.owl
Biological imaging methods	experiments	FBbi	image.obo
Biological Spatial Ontology	anatomy	BSPO	bspo.obo
BRENDA tissue / enzyme source	anatomy		
C. elegans development	anatomy		

Over hundred of ontologies in OBO Library

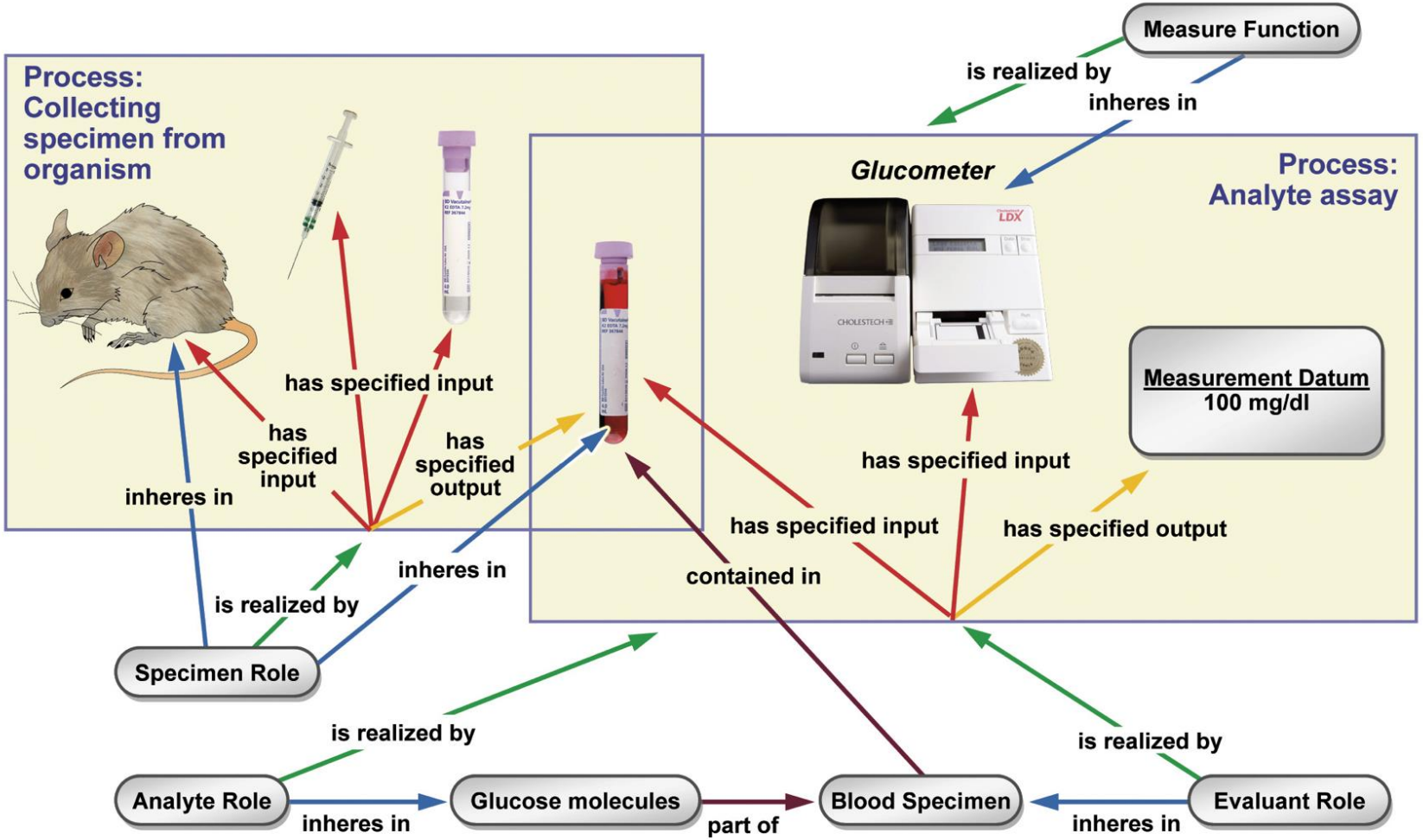
High level structure of OBI



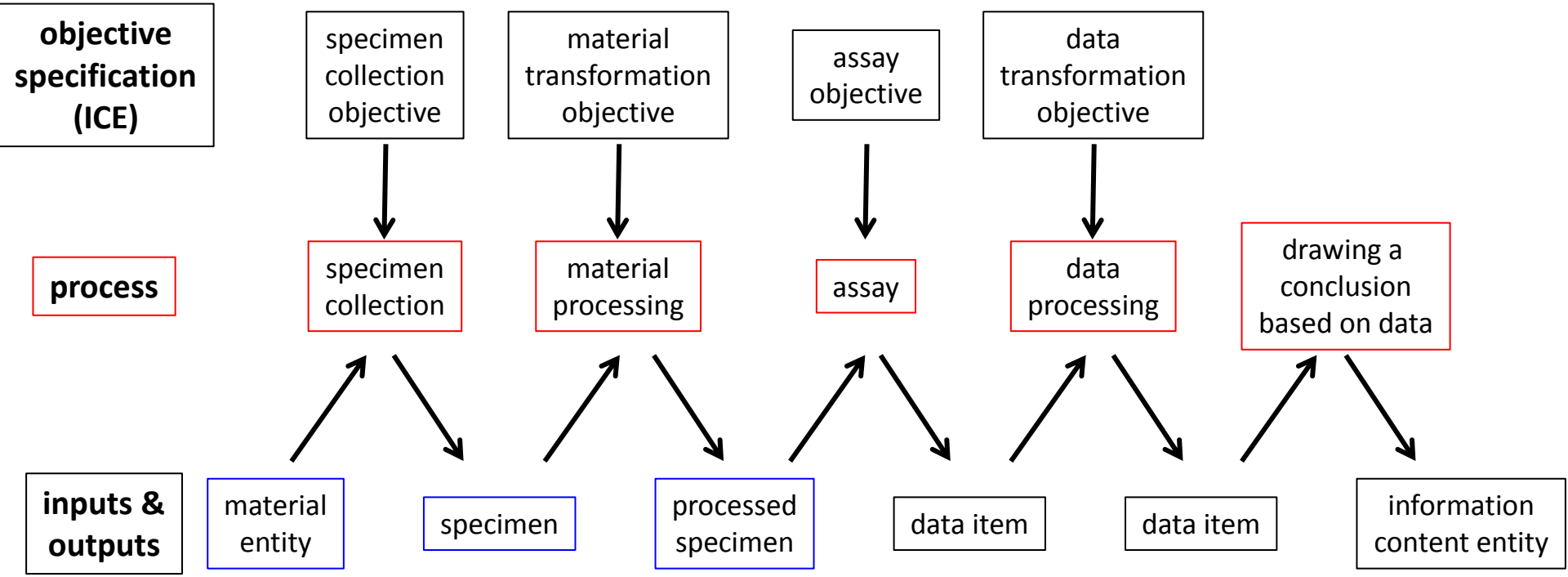
Main components of OBI and their relations



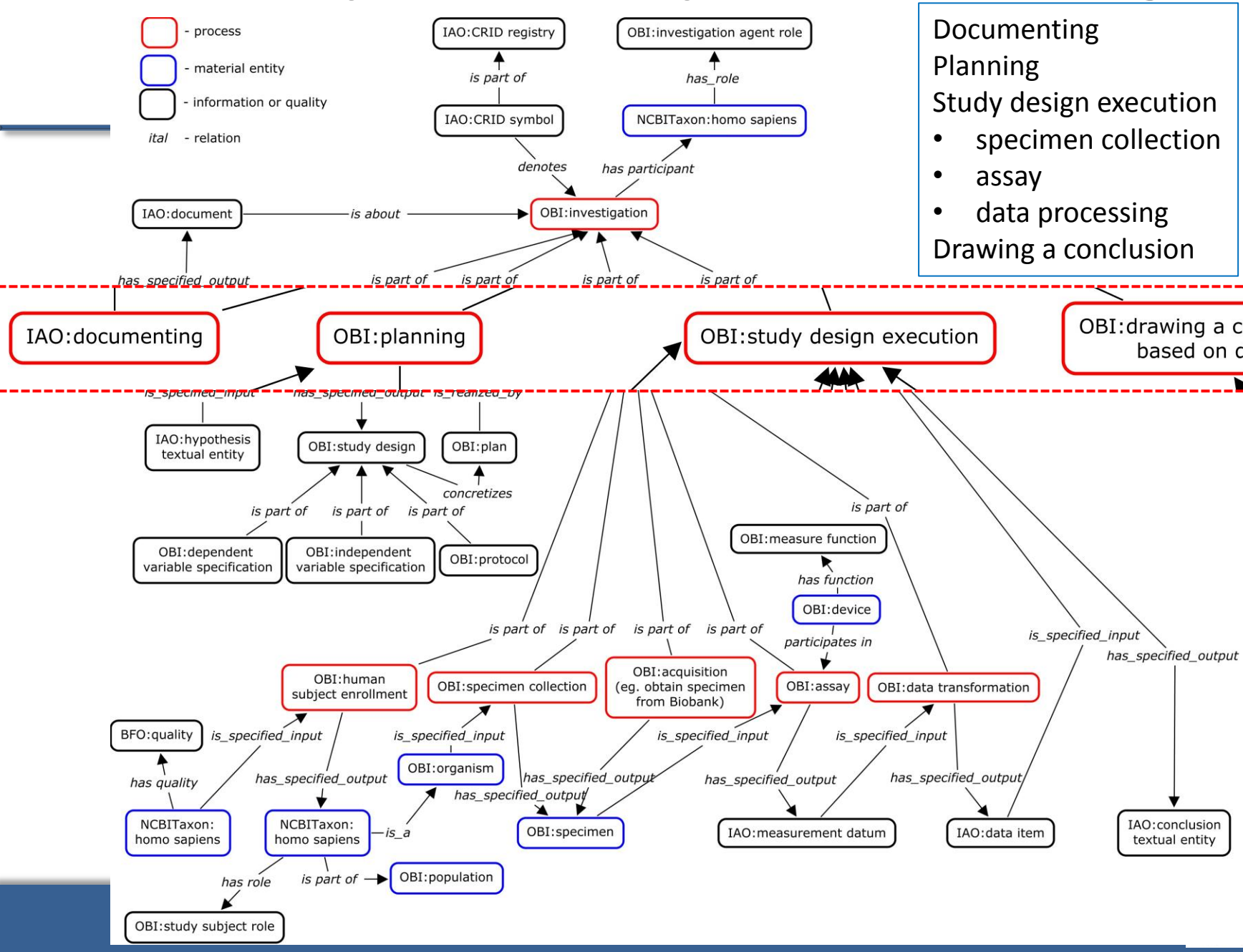
Measurement of Glucose concentration in blood



OBI represents an investigation focusing on processes



OBI can represent all aspects of an investigation



Documenting
Planning
Study design execution
- specimen collection
- assay
- data processing

Drawing a conclusion

Penn
University of Pennsylvania

OBI-related Resources

- The release version of OBI is available on:
  - NCBO Bioportal website:

    http://bioportal.bioontology.org/
  - Ontobee website:

    http://www.ontobee.org/
- The link of latest release version of OBI is:
  - http://purl.obolibrary.org/obo/obi.owl

Many projects are using OBI

Projects Using This Ontology Create new project

PROJECT	DESCRIPTION	PEOPLE	INSTITUTION
Electrophysiology Ontology	The Electrophysiology (EP) Ontology is part of Project 2 of the...	Dr. Raimond L. Winslow, Stephen J. Granite	The Johns Hopkins University
Influenza Ontology	The Influenza Ontology is an application ontology covering the ...	Burke Squires, Lynn Schriml, Joanne Luciano	UT Southwestern, Univ of Maryland, MITRE
Neuroscience Information Framework	The Neuroscience Information Framework (NIF; http://nif.nih.gov...	Maryann Marto...	
An Ontology for Drug Discovery Investigations	The goal of DDI project is to develop an ontology for the descr...	Da Qi, Larisa...	Aberystwyth University
Adverse Event Reporting Ontology (AERO)	The Adverse Event Reporting Ontology (AERO) is an ontology aime...	Melanie Courtot	
ISA software suite	An open source ISA software suite and an extensible hierarchica...	International collaborative effort	Multiple institutions; leads at University of Oxford, UK
Immune Epitope Database	The IEDB contains data related to antibody and T cell epitopes ...	http://www.immuneepitope.org/acknowledgements.php	La Jolla Institute for Allergy & Immunology
Integrative Tools for Protozoan Parasite Research (ITPPR)	The Integrative Tools for Protozoan Parasite Research project i...	Christian Stoeckert	University of Pennsylvania
NCBO Annotator	A Web service that tags free text with ontology concepts. NCBO ...	NCBO	Stanford University
FGED-MGED Ontology	The Functional Genomics Data (FGED) Society has incorporated th...		

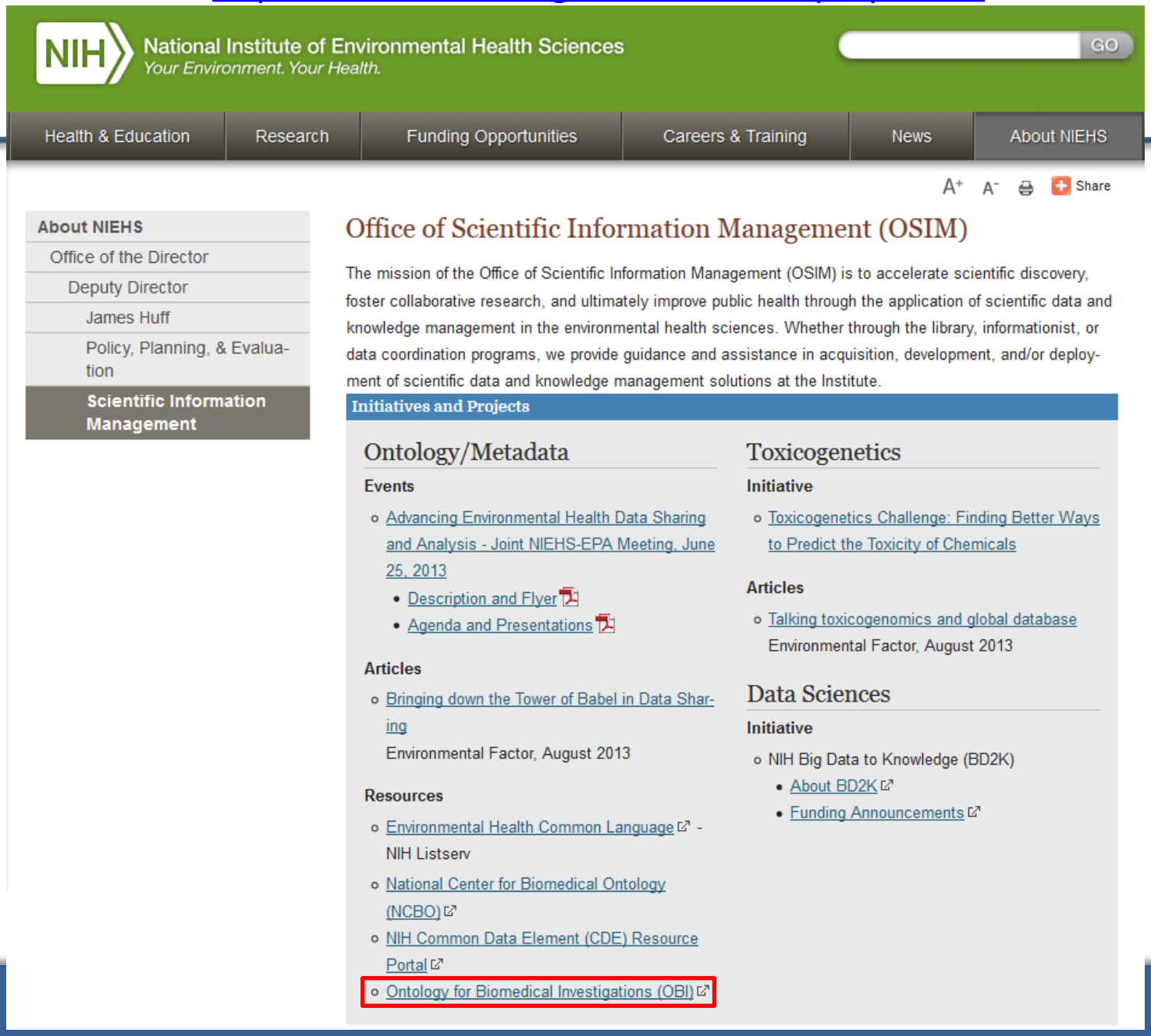
Over 20 projects are using OBI

- Annotations: Database (*e.g.* IEDB, EupathDB, ArrayExpress) and tools (*e.g.* ISA tool)
- Build application ontologies (e.g. Influenza Ontology, BCO, BCGO)
- Semantic framework: metadata standardization (*e.g.* NIAID Core Metadata)

Penn
University of Pennsylvania

OBI is listed as an Ontology/Metadata resource at NIEHS

http://www.niehs.nih.gov/about/od/deputy/osim/



NIH National Institute of Environmental Health Sciences
Your Environment. Your Health.

GO

Health & Education | Research | Funding Opportunities | Careers & Training | News | About NIEHS

A+ A- Share

About NIEHS
- Office of the Director
  - Deputy Director
    - James Huff
    - Policy, Planning, & Evaluation
    - **Scientific Information Management**

Office of Scientific Information Management (OSIM)

The mission of the Office of Scientific Information Management (OSIM) is to accelerate scientific discovery, foster collaborative research, and ultimately improve public health through the application of scientific data and knowledge management in the environmental health sciences. Whether through the library, informationist, or data coordination programs, we provide guidance and assistance in acquisition, development, and/or deployment of scientific data and knowledge management solutions at the Institute.

Initiatives and Projects

Ontology/Metadata

Events

- Advancing Environmental Health Data Sharing and Analysis - Joint NIEHS-EPA Meeting, June 25, 2013
  - Description and Flyer
  - Agenda and Presentations

Articles

- Bringing down the Tower of Babel in Data Sharing
  Environmental Factor, August 2013

Resources

- Environmental Health Common Language - NIH Listserv
- National Center for Biomedical Ontology (NCBO)
- NIH Common Data Element (CDE) Resource Portal
- Ontology for Biomedical Investigations (OBI)

Toxicogenetics

Initiative

- Toxicogenetics Challenge: Finding Better Ways to Predict the Toxicity of Chemicals

Articles

- Talking toxicogenomics and global database
  Environmental Factor, August 2013

Data Sciences

Initiative

- NIH Big Data to Knowledge (BD2K)
  - About BD2K
  - Funding Announcements

Penn University of Pennsylvania

Applying OBI to Metadata Standardization

Metadata for human pathogen/vector genomic sequences

- Genome Sequencing Centers for Infectious Diseases (GSCIDs), the Bioinformatics Resource Centers (BRCs), and the U.S. National Institute of Allergy and Infectious Diseases (NIAID)
  - Project
  - Specimen
  - Sequencing
- http://www.niaid.nih.gov/labsandresources/resources/dmid/metadata/pages/default.aspx
- Dugan, Vivien G., et al. "Standardized Metadata for Human Pathogen/Vector Genomic Sequences." *PloS one* 9.6 (2014): e99979.

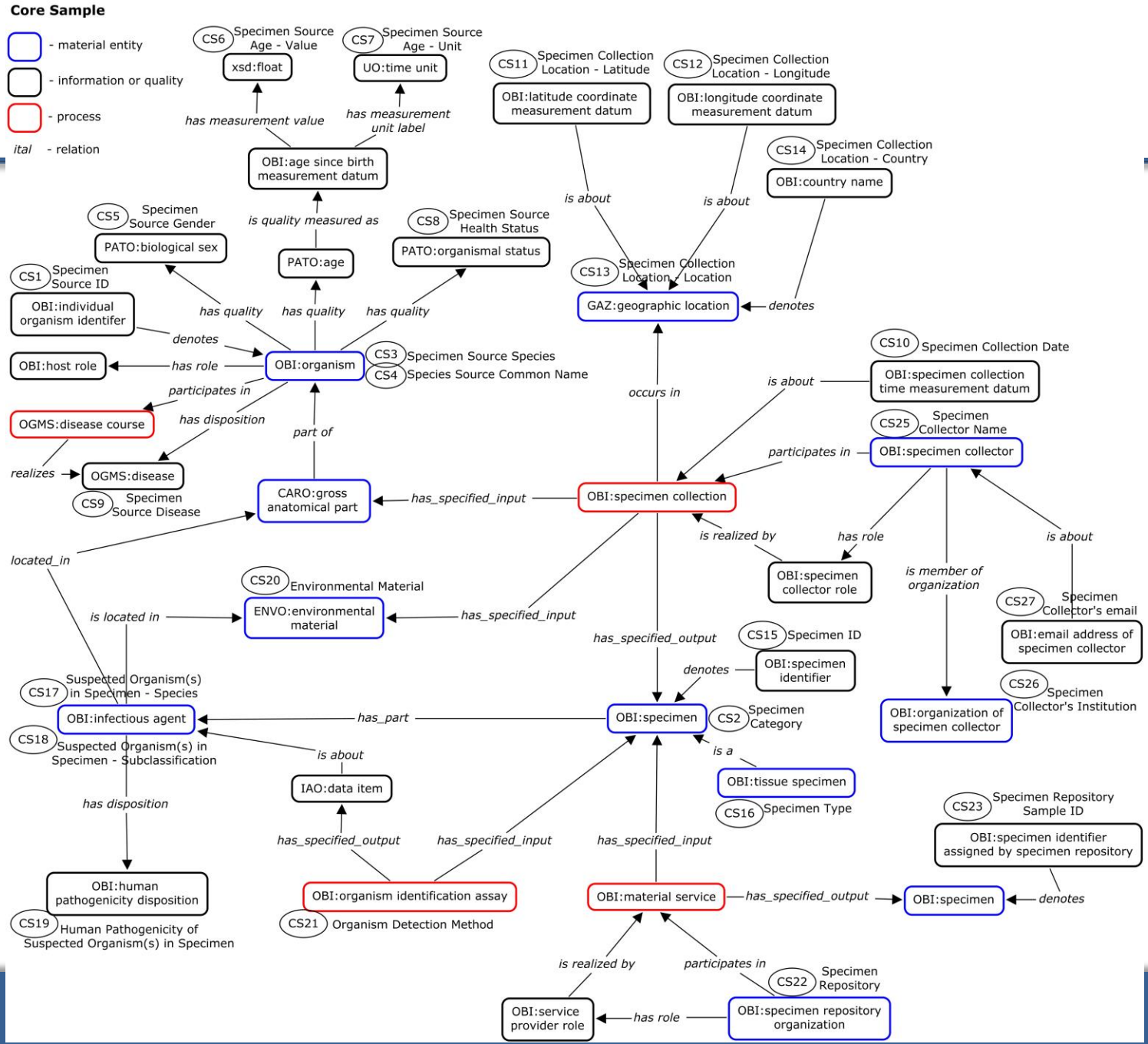
Standardized metadata for human pathogen/vector genomic sequences

- Generate checklist/data dictionary
- Semantic representation of metadata using OBI
- Map fields to other data standards, including the Genomic Standards Consortium's minimal information (MIxS) and NCBI's BioSample/BioProjects checklists via OBI/OBO ontology terms

Core Sample Attributes And Mappings

Project Field ID	Field Name	Data Categories	OBO Foundry Purl	BioProject Synonyms	MIxS Synonym
CP1	Project Title	Investigation	http://purl.obolibrary.org/obo/OBI_0001622	Title*	project name
CP2	Project ID	Investigation	http://purl.obolibrary.org/obo/OBI_0001628		
CP3	Project Description	Investigation	http://purl.obolibrary.org/obo/OBI_0001615	Description*	
CP4	Project Relevance	Investigation	http://purl.obolibrary.org/obo/OBI_0500000	Relevance*	
CP5	Sample Scope	Investigation	http://purl.obolibrary.org/obo/OBI_0001884	Sample Scope*	
CP6	Target Material	Investigation	http://purl.obolibrary.org/obo/OBI_0001882	Material*	
CP7	Target Capture	Investigation	http://purl.obolibrary.org/obo/OBI_0001899	Capture*	
CP8	Project Method	Investigation	http://purl.obolibrary.org/obo/OBI_0001896	Methodology*	
CP9	Project Objectives	Investigation	http://purl.obolibrary.org/obo/OBI_0001892	Objective*	
CP10	Grant Agency	Investigation			
CP11	Supporting Grants/Contract ID	Investigation	http://purl.obolibrary.org/obo/OBI_0001629	Grant ID	
CP12	Publication Citation	Investigation	http://purl.obolibrary.org/obo/OBI_0001617	PubMed ID; DOI	ref_ biomaterial
CP13	Sample Provider Principal	Investigation	http://purl.obolibrary.org/obo/OBI		

OBI reveals relations between sample attributes



Applying OBI to ICEMR Protein Array Data Data Collection and Integration

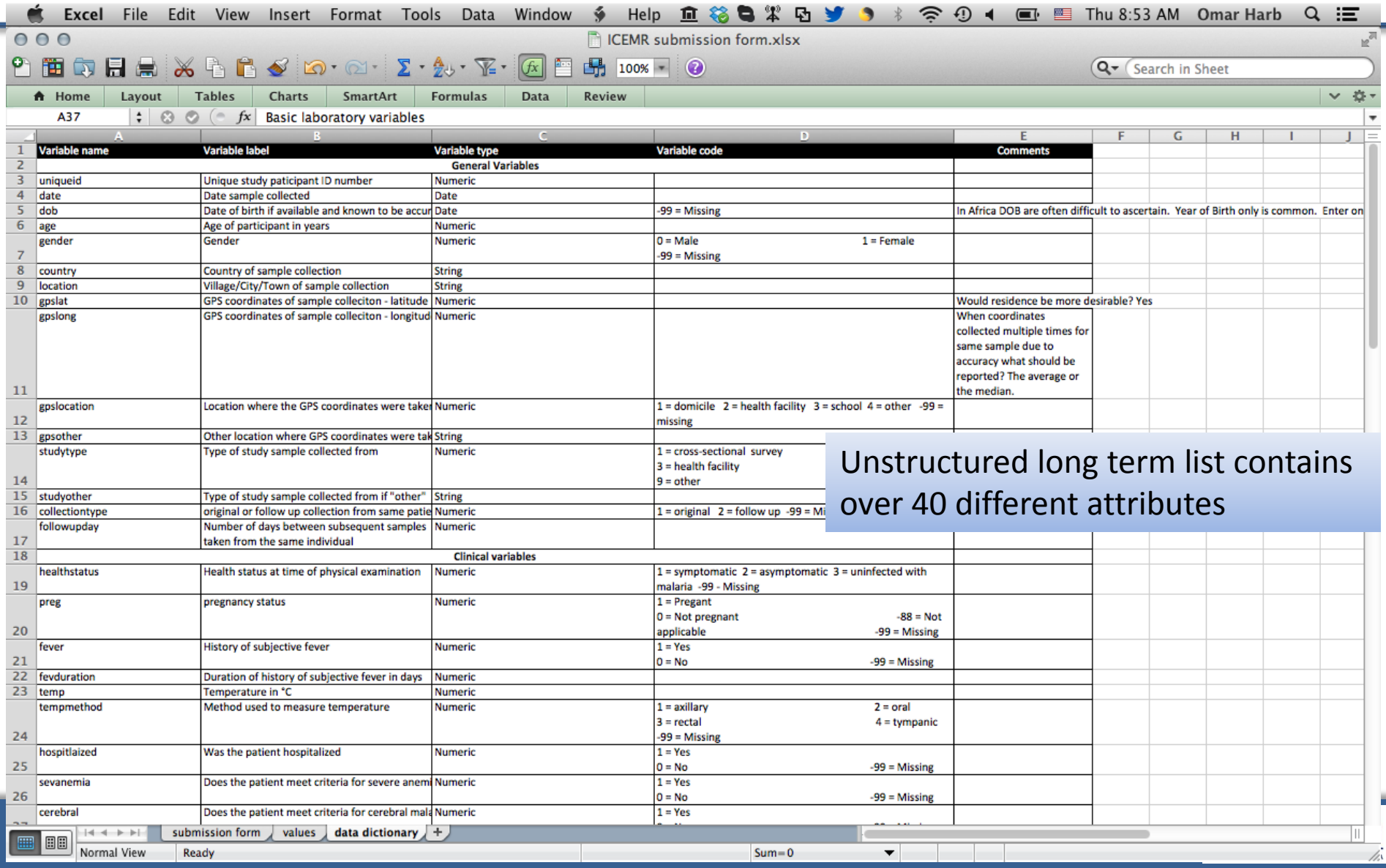
ICEMR Projects

- Global collaborative projects aim to understand the epidemiology and transmission patterns of malaria in different geographic regions in malaria research
  - 10 regional ICEMR groups (*e.g.* Amazonia, Malawi, South Asia, West Africa, East Africa, India, etc.)
- Data produced as a result of the research activities undertaken by the International Centers of Excellence in Malaria Research (ICEMRs)
  - heterogeneous with respect to origin, type of data, and format of data
- Well-structured data and consistent representation of metadata are needed for accurate data integration and cross-study analysis

Challenges

- Minimum information should be captured for malaria studies
  - what host, parasite or vector samples were collected
  - where and when the samples were collected
  - what clinical phenotypes of host had
  - what kinds of assay were performed …
- Values and format of data should be used for consistent representation

ICEMR data dictionary was provided by representatives of ICEMR groups



Excel File Edit View Insert Format Tools Data Window Help Thu 8:53 AM Omar Harb

ICEMR submission form.xlsx

Home | Layout | Tables | Charts | SmartArt | Formulas | Data | Review

A37 — Basic laboratory variables

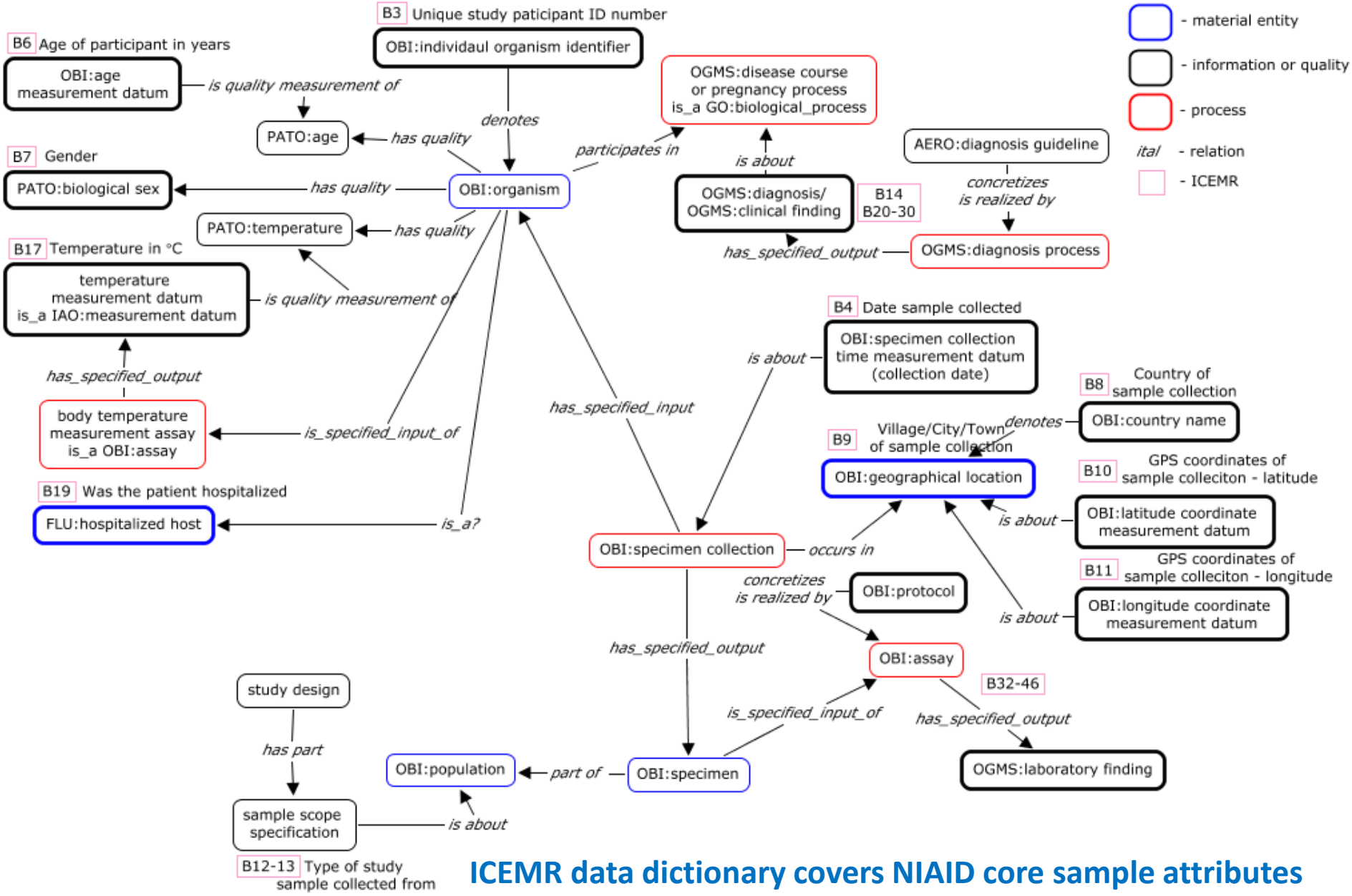
	Variable name	Variable label	Variable type	Variable code	Comments
2			General Variables		
3	uniqueid	Unique study paticipant ID number	Numeric		
4	date	Date sample collected	Date		
5	dob	Date of birth if available and known to be accur	Date	-99 = Missing	In Africa DOB are often difficult to ascertain. Year of Birth only is common. Enter on
6	age	Age of participant in years	Numeric		
7	gender	Gender	Numeric	0 = Male 1 = Female -99 = Missing	
8	country	Country of sample collection	String		
9	location	Village/City/Town of sample collection	String		
10	gpslat	GPS coordinates of sample colleciton - latitude	Numeric		Would residence be more desirable? Yes
11	gpslong	GPS coordinates of sample colleciton - longitud	Numeric		When coordinates collected multiple times for same sample due to accuracy what should be reported? The average or the median.
12	gpslocation	Location where the GPS coordinates were take	Numeric	1 = domicile 2 = health facility 3 = school 4 = other -99 = missing	
13	gpsother	Other location where GPS coordinates were tak	String		
14	studytype	Type of study sample collected from	Numeric	1 = cross-sectional survey 3 = health facility 9 = other	
15	studyother	Type of study sample collected from if "other"	String		
16	collectiontype	original or follow up collection from same patie	Numeric	1 = original 2 = follow up -99 = Mi	
17	followupday	Number of days between subsequent samples taken from the same individual	Numeric		
18			Clinical variables		
19	healthstatus	Health status at time of physical examination	Numeric	1 = symptomatic 2 = asymptomatic 3 = uninfected with malaria -99 - Missing	
20	preg	pregnancy status	Numeric	1 = Pregant 0 = Not pregnant -88 = Not applicable -99 = Missing	
21	fever	History of subjective fever	Numeric	1 = Yes 0 = No -99 = Missing	
22	fevduration	Duration of history of subjective fever in days	Numeric		
23	temp	Temperature in °C	Numeric		
24	tempmethod	Method used to measure temperature	Numeric	1 = axillary 2 = oral 3 = rectal 4 = tympanic -99 = Missing	
25	hospitlaized	Was the patient hospitalized	Numeric	1 = Yes 0 = No -99 = Missing	
26	sevanemia	Does the patient meet criteria for severe anemi	Numeric	1 = Yes 0 = No -99 = Missing	
	cerebral	Does the patient meet criteria for cerebral mal	Numeric	1 = Yes	

Unstructured long term list contains over 40 different attributes

submission form | values | data dictionary

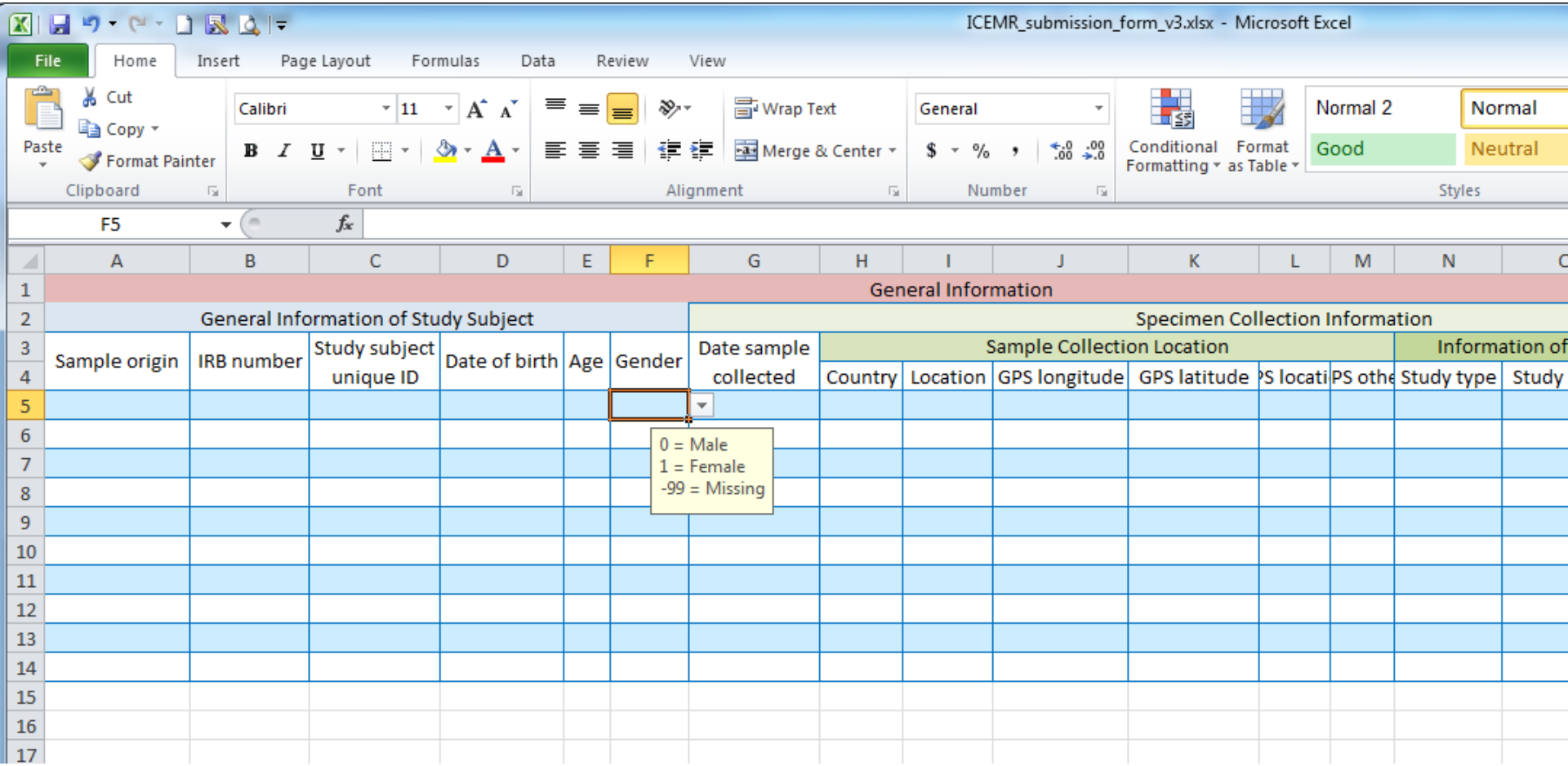
Normal View | Ready | Sum=0

Ontology mapping and representation allows comparison between the ICEMR data dictionary and the NIAID standard and to identify missing attributes

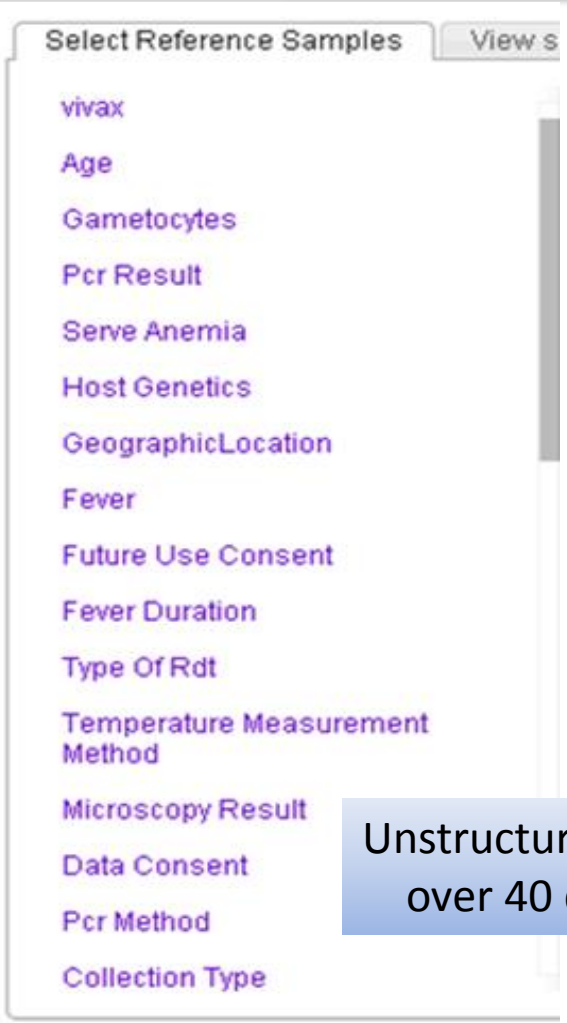


ICEMR data dictionary covers NIAID core sample attributes

OBI was used to build the ICEMR data submission form by helping to group data dictionary entries in a structured and logical manner

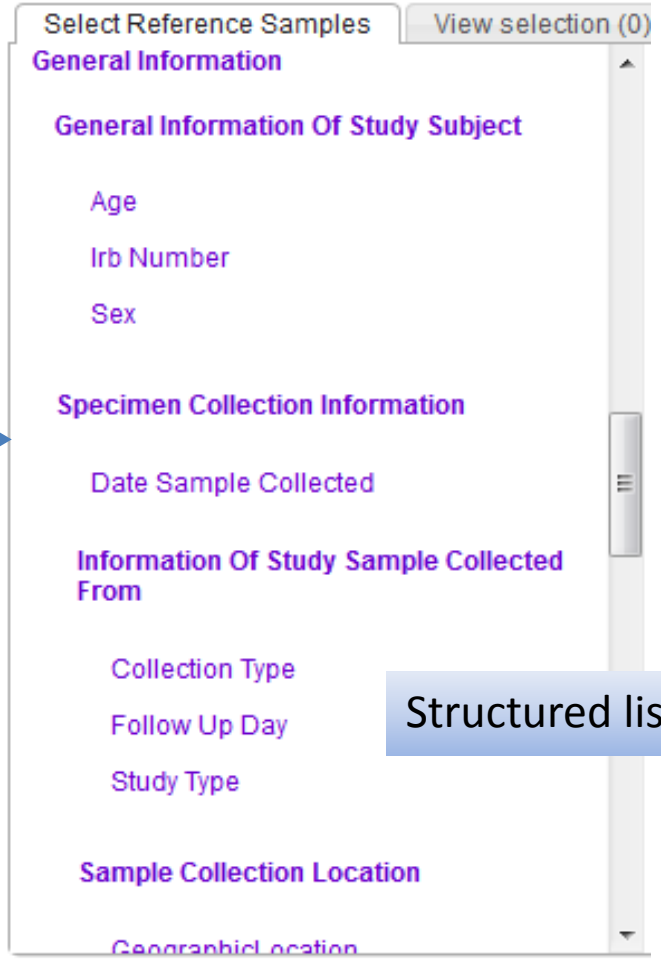


OBI enables a more user-friendly interface for queries based on sample attributes



Unstructured long list
over 40 different attributes

organized sample attributes based on OBI framework

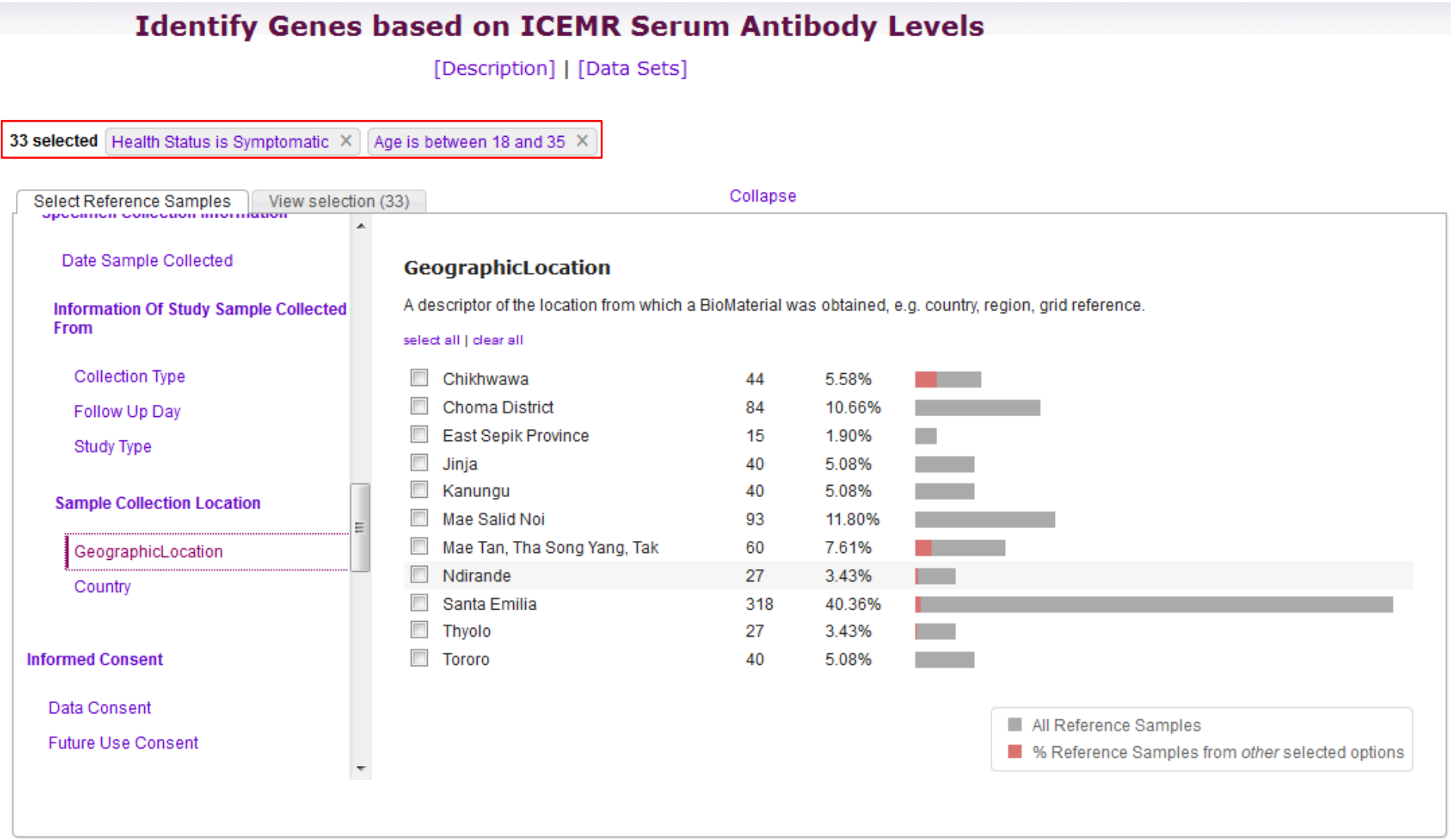


Structured list

Consistent representation of ICEMR data using ontology

- Initial loading data: antibody array data
- For example, ICEMR ‘health status’
  - Asymptomatic, Symptomatic, Uninfected with malaria, unknown
  - Data from Amazonia group
    - sint7d: 1 -> 'health status': symptomatic
    - sint7d: 0 -> 'health status': asymptomatic
  - Data from Southwest Pacific group
    - parasite detection positive + fever or other symptom (e.g. anemia) -> 'health status': symptomatic
    - parasite detection positive + no other symptom -> 'health status': asymptomatic malaria

Data can be better explored after ontology-based harmonization and integration: Identify study subjects between ages 18-35 with symptomatic malaria and their geographic location



screenshot from ICEMR Plasmodb.org web site

Penn
University of Pennsylvania

Applying OBI to PRISM Studies Metadata Representation and Organization

PRISM Studies

- Longitudinal cohort study following participants from 300 households in three regions of Uganda with diverse demographics and transmission intensity:
  - Jinja (low incidence of malaria)
  - Kanunga (mild incidence of malaria)
  - Tororo (high incidence of malaria)
- Quarterly routine visits, plus additional sick visits
- Monthly mosquito collection in each dwelling

Extensive metadata: hard to understand what they represent and how they are related to each other

Household data (over 80 fields)

VISDATE	STARTIME	HHNUM	DISTRICT	INTNUM	AGREE	NUMPEOF	SWATER	OTHERSCS	TFACLTY	OTHERFCY	ELECTIRC	RADIO	CASSETTE	TV
10-Sep-11	12:18:46	201033121	JINJA	1	YES	6	PUBLIC T	[Skipped]	UNCOVER	[Skipped]	NO	NO	NO	N
5-Sep-11	11:30:01	206002105	JINJA	1	YES	10	PUBLIC T	[Skipped]	UNCOVER	[Skipped]	NO	YES	NO	N
8-Sep-11	9:46:11	205016305	JINJA	2	YES	5	PUBLIC T	[Skipped]	UNCOVER	[Skipped]	YES	NO	NO	N
7-Sep-11	10:31:50	216001607	JINJA	1	YES	8	PUBLIC T	[Skipped]	COVERED	[Skipped]	YES	YES	YES	YE

Household member data (about 20 fields)

hhid	uniqueid	cohort	LINE	id	G6PD	alphathal	hbs	RLTSHP	SEX	LIVHER	STYHR	AGE	ANSW	RLTSHPcat
HH205011301	1205011301	0	1					Head of household	MALE	YES	YES	36		Head of househo
HH201030403	10201030403	0	10					GRANDCHILD	FEMALE	YES	YES	10		Grandchild
HH205005903	8205005903	1	8	1091		0	0	SON OR DAUGHTER	MALE	YES	YES	0		Son or daughter
HH201033121	4201033121	0	4					SON OR DAUGHTER	FEMALE	YES	YES	15		Son or daughter

Clinical visits data (about 170 fields)

id	DATE	startdate	siteid	gender		itnlastnight
1001	5-Aug-11	5-Aug-11	Jinja	Female	1 0:00	
1001	1-Nov-11	5-Aug-11	Jinja	Female	1 0:00	1
1001	30-Nov-11	5-Aug-11	Jinja	Female	1 0:00	1

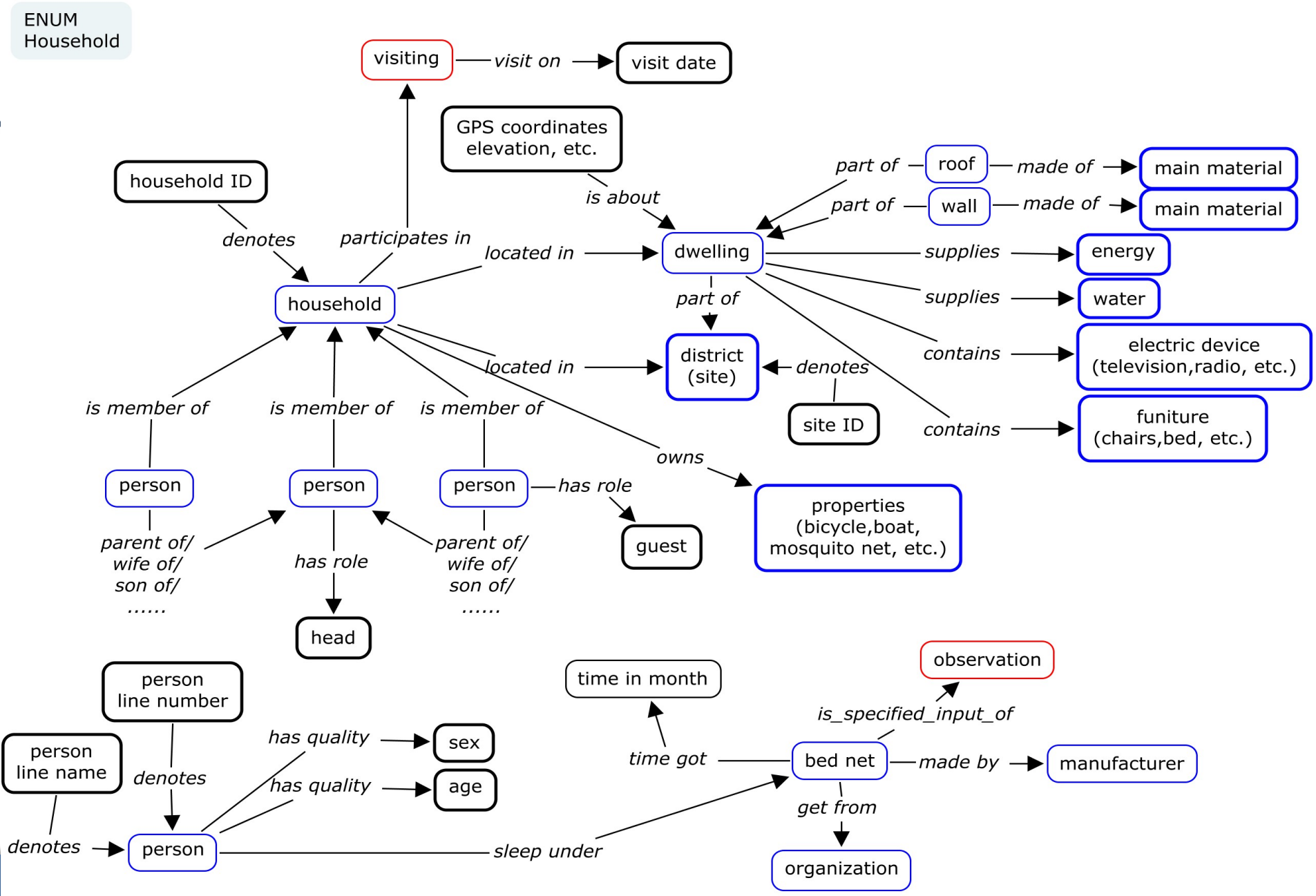
Challenges:
How to load data?
How to query data and retrieve useful information?

Mosquito trapping data (about 10 fields)

hhid	barcode1	fedgamb1	unfedgamb1	gravidgambie1	fedfunes1	unfedfunes1	totalgravid1	unableassess1	femotherano1	gambiaeunableassess1
101009801	T4-M8XL	0	14	1	0	3	0	0	1	0
102014401	T4-CAU9	0	37	4	0	0	0	0	0	0
102019101	T4-6FVH	2	7	0	0	0	0	0	3	0

Total over 280 different kinds of metadata

OBI helped to understand metadata and relations between them

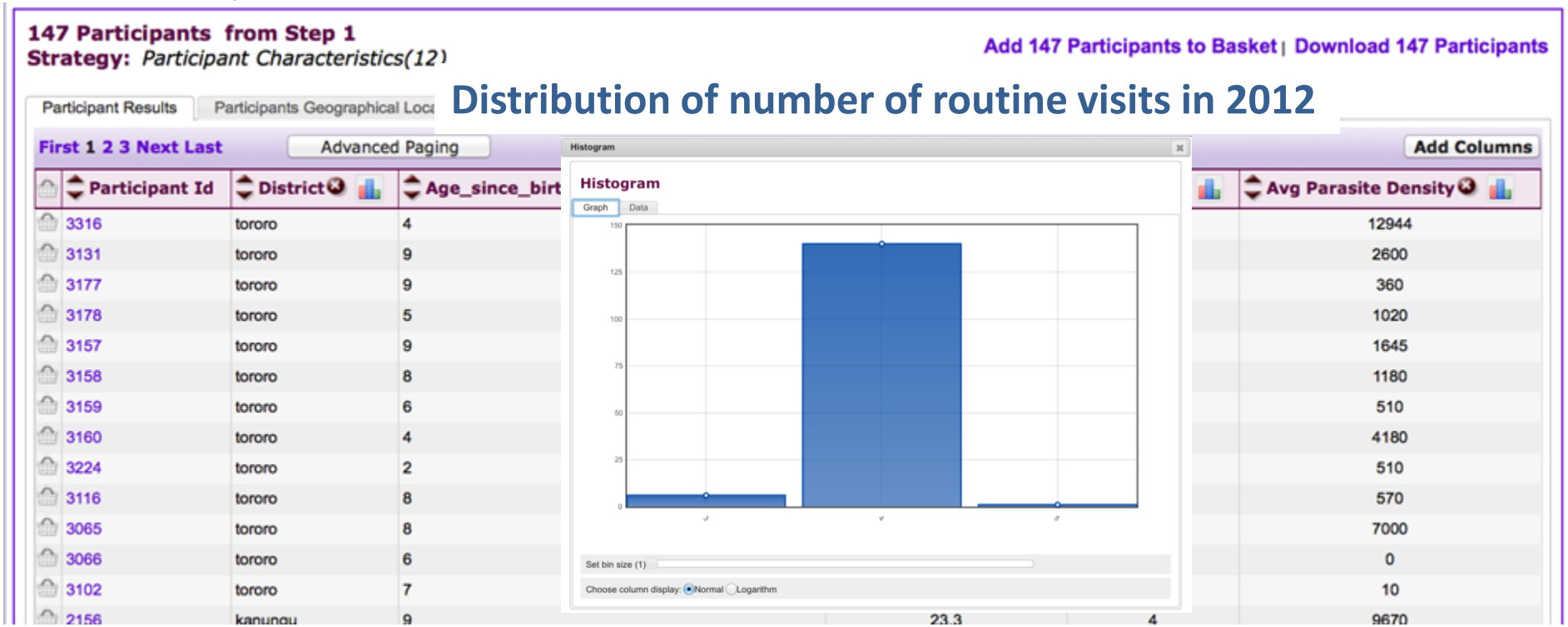


Ontologies support consistent representation of data

- Other OBO ontologies are need for PRISM data annotation, such as
  - Gene ontology (GO)
  - Human Disease Ontology (DOID)
  - Human Phenotype Ontology (HPO)
  - Ontology for General Medical Science (OGMS)
  - Protein Ontology (PRO)
- OBI is a starting point to pull in other important ontologies in a semantically consistent manner

Query PRISM Data

- Data annotated using OBI and OBO ontologies
- Data loaded into database
- Samples can be retrieved based on clinical metadata



Credit: Omar Harb, EuPathDB Outreach Coordinator

Penn
University of Pennsylvania

Relations between data revealed based on ontology enable complexity queries

- Asymptomatic infection? Identify children with high exposure but no clinical malaria symptoms.
  - what is the impact of age?
  - what is the impact of prior exposure?
  - geographic correlates?
- Hyper-susceptibles? Children with low exposure but multiple bouts of malaria.
  - human genotypes?
  - parasite genotypes?
- Families with both? What clinical / behavioral correlates?

Credit: David Roos, EuPathDB Principle Investigator

Penn
University of Pennsylvania

Summary

- Ontologies like OBI help in metadata standardization and category organization by:
  - supporting consistent data representation
  - providing a semantic framework to understand massive data and reveal inter-connections between them
  - helping in information retrieval and enables complex queries

Acknowledgements

OBI Consortium

Bioinformatics Resource Centers (BRCs)

- Richard Scheuermann

ICEMR

PRISM

- Grant Dorsey
- San James

EuPathDB (PlasmoDB)

- Omar Harb
- David Roos
- Brian Brunk
- Shon Cade
- John Brestelli