Some Thoughts on Data and eScience

Mike Conlon
University of Florida
mconlon@ufl.edu
What Does Data Look Like?
Get a world map showing temperature sensors.
What are the Data Processes?
Data Reasoners
Agencies, universities, corporations
Strong semantics
Irregular persistence
Variable processes

Data Consumption
Access control, semantic alignment, format alignment

Data Archives
Data trusts, not for profits, government agencies
Some semantics
Strong persistence
Variable means of production

Data Sharing
Disclosure, usage agreements, some linkage, some formatting, some semantics

Data Producers
Labs, universities, corporations, government agencies, hospitals
Limited semantics
Irregular persistence
Highly variable means of production
Data Sharing

Photograph by J. G. Park. Flickr.com

Photograph by Ell Brown. Flickr.com
Donate to Freedom and Sharing: CC’s Annual Campaign

JOI ITO

"Help make sharing easier, safer and more effective. CC Licenses enable sharing. Support sharing. Please donate to Creative Commons."

— Joi Ito, CC Board Chair & Director of MIT Media Lab

2011 Creative Commons Annual Campaign
DONATE TODAY

Mission
WHAT IS CREATIVE COMMONS?
Creative Commons helps you share your knowledge and creativity with the world.

License
HOW CAN I LICENSE MY WORK?
There is no registration to use the Creative Commons licenses. Licensing a work is as simple as selecting which of the six licenses best meets your goals, and then marking your work in some way so that others

Explore
LOOKING FOR CREATIVE WORKS?
Looking for music, video, writing, code, or other creative works?

Creative Commons has got you covered.
The Role of the Archive

• Collate data, finalize semantics, ready for consumption
A Consumption Scenario

*Find all faculty members whose genetic work is implicated in breast cancer*

VIVO will store information about faculty and associate to genes. Diseaseome associates genes to diseases. Query resolves across VIVO and data sources it links to.
Data Reasoning

Data integration continues to be a serious bottleneck for the expectations of increased productivity in the pharmaceutical and biotechnology domain.

“Linked Life Data” integrates common public datasets that describe the relationships between gene, protein, interaction, pathway, target, drug, disease and patient and currently consist of more than 5 billion RDF statements.

The dataset interconnects more than 20 complete data sources and helps to understand the “bigger picture” of a research problem by linking previously unrelated data from heterogeneous knowledge.

From the LarKC (Large Knowledge Collider) [http://www.larkc.eu/overview/](http://www.larkc.eu/overview/)
Public, structured linked data about investigators interests, activities and accomplishments, and tools to use that data to advance science.
A Web of Data – The Semantic Web

Information is stored using the Resource Description Framework (RDF) as subject-predicate-object “triples”

Jane Smith

professor in

has affiliation with

author of

Dept. of Genetics

College of Medicine

Genetics Institute

Journal article

Book

Book chapter

Subject

Predicate

Object

A Web of Data – The Semantic Web

Information is stored using the Resource Description Framework (RDF) as subject-predicate-object “triples”

Jane Smith

professor in

has affiliation with

author of

Dept. of Genetics

College of Medicine

Genetics Institute

Journal article

Book

Book chapter

Subject

Predicate

Object
VIVO produces human and machine readable formats

Software reads RDF from VIVO and displays...
**Malaria**

**Abstract**

We review previous studies on the genetic diversity of malaria vectors to highlight the major trends in population structure and demographic history. In doing so, we outline key information about molecular markers, sampling strategies and approaches to investigate the causes of genetic structure in *Anopheles* mosquitoes. Restricted gene flow due to isolation by distance and physical barriers to dispersal may explain the spatial patterns of current genetic diversity in some *Anopheles* species. Nonetheless, there is noteworthy disagreement among studies, perhaps due to variation in sampling methodologies, choice of molecular markers, and/or analytical approaches. More refined genealogical methods of population analysis allowing for the inclusion of the temporal component of genetic diversity facilitate the evaluation of the contribution of historical demographic processes to genetic structure. A common pattern of past unstable demography (i.e., historical fluctuation in the effective population size) by several *Anopheles* species, regardless of methodology (DNA markers), mosquito ecology (anthropophilic vs zoophilic), vector status (primary vs secondary) and geographical distribution, suggests that Pleistocene environmental changes were major drivers of divergence at population and species levels worldwide.
Some Questions Regarding Data Processes
Who pays?
Conlon, Michael  |  Associate Director and Chief Operating Officer

**Positions**

- **Biomedical Informatics**, Director 2010 -
- **Clinical and Translational Science Institute**, Faculty 2010 -
- **Clinical and Translational Science Institute**, Associate Director and Chief Operating Officer 2010 -
- **Health Outcomes and Policy**, Faculty 2008 -
- **Health Outcomes and Policy**, Research Associate Professor 1992 -
- **Health Outcomes and Policy**, Interim Division Chief, 2010 -

[http://vivo.ufl.edu/display/n25562](http://vivo.ufl.edu/display/n25562)