Empowering data sharing for genomics-based public health surveillance using ontologies

Emma Griffiths, PhD
Hsiao Public Health Bioinformatics Lab
Department of Pathology and Laboratory Medicine
University of British Columbia
Vancouver, Canada

IMMEM, September 21 2019
Sequence data

Contextual data

Sample metadata

Clinical/Epi

Lab

Methods

Methods toolbox
Data comes from different sources often needs to be shared for different purposes.

Food Safety & Public Health Activities Requiring Integrated Contextual Information:
- Surveillance
- Recalls
- Outbreak Investigation
- Source Attribution
- Risk Assessment
- More…
Challenges in public health data sharing

- Lack of a mechanism for data sharing between partners
- Security/privacy
- Control over data
- Semantic interoperability
- Manual curation
IRIDA

A comprehensive and distributed platform for public health genomic epidemiology

Easy, automated transfer of data from various sequencing platforms
(GitHub: SeqUDAS & IRIDA Uploader)

Reproducible and versioned workflows
(Assembly, annotation, SNP, MLST, AMR, serotyping, and more.)

Cloud enabled Version Coming!

-expanding Analytical Workflows with plug-in architecture

-Project and Sample Management

-Simple User Interface

Built-in Analytical Tools
Distributed Architecture:
Allow flexible Data sharing

• fine grained access controls: users specify fields of contextual data to share
## Shared Data Between Instances

<table>
<thead>
<tr>
<th>ID</th>
<th>SampleType</th>
<th>Geo</th>
<th>Commodity</th>
<th>Farm</th>
<th>VEpi</th>
<th>Epi Associations</th>
<th>Case#</th>
</tr>
</thead>
</table>

- **Both sites can access**
- **Site A access only**
- **Site B access only**
Harmonizing fields of data between partners/projects

*A field by any other name does NOT smell as sweet...*

SPECIMENSOURCE_1

Isolation

host_tissue_sampled

Source *(Isolation Source)*

Differences in labels,
Same meaning

*Computer doesn’t recognize these as the same thing*

Source *(Submitting Lab)*

Same label,
Different meaning

*Computer doesn’t recognize these as different*

...so, you can’t just combine fields of metadata.
Ontology, A Way of Structuring Information

Scenario: Comparing foodborne disease sources at different granularity

Avian Food Product

has_disposition Transmission Vehicle (Salmonella)

Poultry Food Product

is_a

Turkey Product

Turkey Roast
FOODON:03307670

is_input Surveillance Sample Plan

Chicken Product

Chicken Breast
FOODON:03311787

has_quality light meat

Chicken Breast (sliced, RTE)

Avian Egg Product

is_a

Duck Egg Product

Balut (Balot)
FOODON:03302184

has_disposition Transmission Vehicle (Salmonella)

is_input Surveillance Sample Plan

Hen Egg Product

Hen egg liquid (pasteurized)
FOODON:03311770

FOODON:03311787
Ontologies – an example

The Food Ontology (FoodOn)


- Standardizes food **products, feed, sources and processes** (27 918 classes)
- **Interoperable architecture** (OBO Foundry)
- Characterizes products by **facets**
  - e.g. source, processing, packaging, consumer group
- **Agency high level categories** enables **mapping between food schemes** e.g EFSA (FoodEx2), IFSAC

[Sources:
https://github.com/FoodOntology/foodon
www.foodon.org]
That’s great, but how do I use ontologies?
Lack of ontology-driven tools hinders widespread implementation:

LexMapr - A tool for transforming contextual data

Free text

Ontologized

Hamburger Patty (frozen)
FOODON:03309571

3rd Party Scheme

Cattle (NARMS)

Beef (IFSAC)

Data processing, mapping to ontologies

Map to 3rd party classification scheme
LexMapr Django is still in the development phase, and is currently catered towards food and environmental samples.

Input file*
Choose File: No file chosen
Submit

processing time: minutes

https://watson.bccdc.med.ubc.ca/lexmapr
GEEM: platform for building ontology-based data specifications

http://genepio.org/geem/form.html#GENEPIO:0002083
Summary: creating an ontology-based data sharing ecosystem

- Ontologies e.g. GenEpiO, FoodOn
- Ontology-driven tools e.g. LexMapr, GEEM
- Use cases drive app dev
- Ontology-based access controls
- Distributed system Platforms/systems
Future directions

1. One Health vocabulary

1. ontology integration into IRIDA, other systems (e.g. GenomeTrakr, Resistome Tracker)

1. building ontology-driven ecosystem (tools and partners)
Project Leaders
Fiona Brinkman – SFU
Will Hsiao – UBC/BCCDC
Gary Van Domselaar – NML

GenEpiO Team
Emma Griffiths – SFU
Damion Dooley – UBC
Ivan Gill – UBC
Gurinder Gosal – UBC
Dan Fornika – UBC

Simon Fraser University (SFU)
Geoff Winsor
Julie Shay
Claire Bertelli
Matthew Laird
Bhav Dhillon

McMaster University
Andrew McArthur
Brian Alcock

European Food Safety Agency
Leibana Criado Ernesto
Vernazza Francesco
Rizzi Valentina

National Microbiology Laboratory (NML)
Franklin Bristow
Aaron Petkau
Thomas Matthews
Josh Adam
Adam Olsen
Tara Lynch
Shaun Tyler
Philip Mabon
Philip Au
Celine Nadon
Matthew Stuart-Edwards
Morag Graham
Chrystal Berry
Lorelee Tschetter
Eduardo Toboada
Peter Kruczkiewicz
Chad Laing
Vic Gannon
Matthew Whiteside
Ross Duncan
Steven Mutschall

University of Lisbon
João Carriço

European Bioinformatics Institute
Melanie Courtot
Helen Parkinson

BC Centre for Disease Control and Public Health Lab
Judy Isaac-Renton
Patrick Tang
Natalie Prystajecky
Jennifer Gardy
Linda Hoang
Kim MacDonald
Yin Chang
Eleni Galanis
Marsha Taylor
Jennifer Law

University of Maryland
Lynn Schriml

Canadian Food Inspection Agency (CFIA)
Adam Koziol
Burton Blais
Catherine Carrillo

Dalhousie University
Rob Beiko
Alex Keddy

www.irida.ca
Acknowledgements

Will Hsiao
Damion Dooley
Gurinder Gosal
Ivan Gill
Dan Fornika
Hsiao Lab
Want to know more about IRIDA?
www.irida.ca

Want an IRIDA account?
irida-accounts@sfu.ca

Want to try LexMapr?
https://github.com/Public-Health-Bioinformatics/LexMapr
https://watson.bccdc.med.ubc.ca/lexmapr

Want to know more about our ontologies?
FoodOn – www.foodon.org
GenEpiO – www.genepio.org