



Sage Group B: Standards and Ontologies for Integration, Analysis, and Exchange of Global Coherent Datasets

23 April 2010

Acknowledgements



- Susanna-Assunta Sansone
- Cameron Neylon
- Philippe Rocca-Serra
- Jim Davies
- Steve Harris
- Paul Fisher
- Peter Li
- Carole Goble
- Andrew Kasarskis
- George A. Komatsoulis
- Alex Pico
- Kim Hartwell
- Jonathan Rees
- James Brenton
- Norman Paton
- Ilya Kuperschmidt
- Stephen Friend
- Eric Schadt
- Kaitlin Thaney

What do you think are the three top issues...?

- Consistent data format and metadata
- Reproducible data
- Ability to capture structured content
- Clear and useful data format standards for datasets and networks
- Agreement on types of data/how to standardize
- Data interoperability
- Standards, data sharing norms
- Knowledge representation and ontologies
- Data standardization
- ...

Ok, structured data standards

- Content
 - Enumeration of data elements
 - “Minimum information” lists
- Semantics
 - Actual meaning
 - Terminologies, ontologies, controlled vocabularies
- Syntax
 - Exchange format

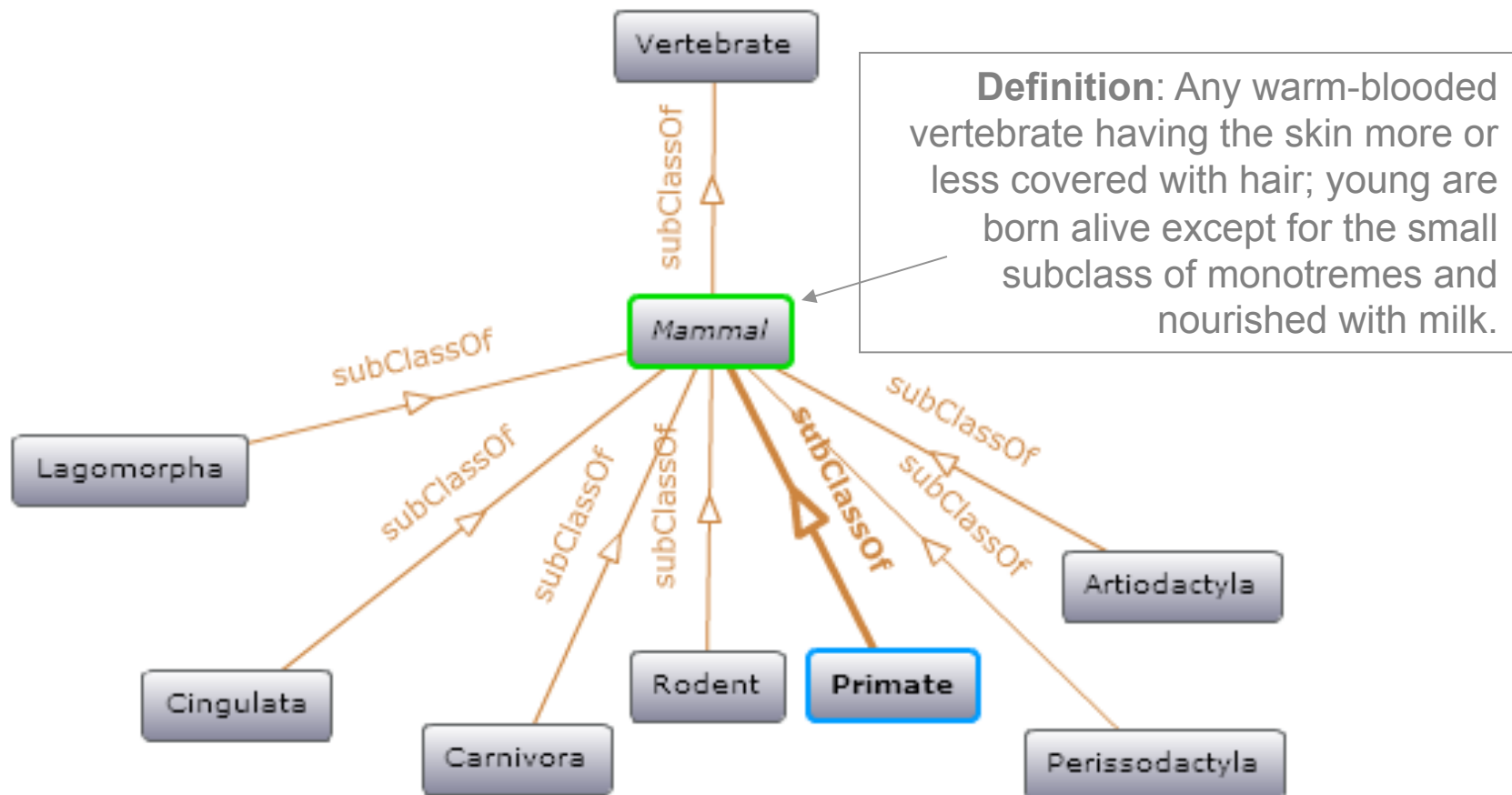
Content



- *Enumeration* of data elements
- “Minimum Information” lists
- E.g. MIAME- Minimum Information about a Microarray Experiment- 6 critical elements:
 - Raw data
 - Processed/normalized data
 - Sample annotation
 - Experimental design
 - Array annotation
 - Lab and data processing protocols

Semantics: the meaning

Ontology: a formal representation of a set of concepts within a domain and the relationships between those concepts.



Syntax: The Format

- Exchange format
- E.g. XML
 - `<gene> <name=tlr4>
<gbacc=NC_000070.5 >
<loc=9q345,987> </gene>`
- E.g. tabular format
 - Rows are subjects, columns are lab values
 - Rows are genes, columns are subjects

What we did

- Identified relevant standards
- Identified a set of open source, standards-based tools
- Performed dataset annotation as proof of concept for 2 Sage datasets:
 - BxH ApoE
 - TCGA
- Explored potential for synergies and integration between tools



What we did not do, by design

- Build a monolithic system that must be used in order to contribute to Sage
- Define a strict information model that must be followed in order to contribute to Sage
- Enumerate specific required ontologies

What is yet to be done

- Formalize some *recommended* minimum information models
- Extend, enhance and integrate tools
- Determine an optimum balance of structured annotation vs. ease and expressivity
- Determine optimum distribution of labor between curation experts and data contributors
- Integrate into pipeline with other working groups
- Architect and implement a queryable back-end storage system

Yet to be done, cont.



- Input from the broader community!

Group B: Afternoon session

- Background
 - ApoE -/- dataset as proof of concept
- Demos
 - Selection of annotation tools
- Discussion
 - Annotation: how, how much, who?
 - Next steps