The realist approach to building ontologies for science

Alan Ruttenberg

Science Commons

The perspective

- A realist ontology is concerned with enumerating those things that exist and how they relate to each other, in particular those things they existentially depend on.
- This enables a kind of audit trail from representations to elements of the world which can be checked.
- A realist ontology is one that has the best chance of being effectively communicated, it's creation objectively moderated, and as a result being widely accepted.

How to Build Consensus around Ontology

- Understand what terms mean by relating them (tracing them) to elements in reality. Or say when you are not.
- Know what the instances are
- Ask yourself how instances are related to each other
- Define classes as groups of instances that have shared properties
- Document and organize this knowledge in a way that can be managed in a distributed manner
- The product of this effort is your Ontology

Three Ways of Representing Scientific Knowledge

- *Record level*: Represent database records. Inconsistent if two sources disagree about contents of a field.
- *Statement level*: Represent what researchers say. Inconsistent if two people disagree about what a paper said
- *Domain level*: OBO Foundry approach. Represent your best understanding of consensus. Inconsistent if facts contradict.
- We need all three (but make clear which is which)

Questions to ask

- If you say there is a class, what are the particulars (syn: members, individuals, instances, entities)?
- What are the entities dependent on – without what else can't they exist?
- When do entities come into existence. When do they go out of existence?
- Is it with respect to a certain perspective (e.g. a certain granularity) that an entity can be referred to?

Upper Level Distinctions

- Continuants (Things)
  - Independent Continuants: fish, rivers, people
  - Dependent Continuants
    - Qualities: shape, mass, length
    - Realizables: trap function, regulator role, statistician role
  - Generically dependent continuants (Information)
- Occurrents (Processes)
  - swimming
  - measuring
  - a storm
  - part of a life (a stage of development)

Instances (1)

- "Objects" (*particulars, independent continuants*)
  - one coho salmon
  - Pine Creek on the Hoopa Valley Reservation
  - one fyke net
  - one passive integrated transponder tag
  - Bonneville dam
  - one fish ladder near the Bonneville dam
  - The National Oceanic and Atmospheric Administration
  - one person
  - one notebook

"fully present at every time when it exists"

Instances (2)

- "properties" (*dependent continuants*)
  - The ability of a passive integrated transponder tag to respond to a radio signal (a function)
  - NOAA Fisheries Service charge to review locally prepared salmon recovery plans (a role)
  - one salmon's fork length (changes over time)
  - the measured value of one salmon's fork length in millimeters

"fully present at every time when it exists"

Instances(3)

- Processes
  - a salmon swimming up a salmon ladder
  - Member of the NOAA Fisheries Service preparing advise to augment a salmon recover plan (realizing their role)
  - A PTT responding to a radio signal (executing its function)
  - A fisheries staff member weighing a salmon

"Takes place(unfolds) over a period of time"

Relations between instances

- trap17 **located_in** xxx river
- salmon23 **has_quality** {fork length of 12.2 mm} now
- attenna78 **has_function** PTT detection
- John Samuels **has_role** Biostatistian

Classes

- Those entities that are like in some way
- Mostly expressible as the relationships that their instances have to other instances
- **A PTT Antenna is a antenna that has function PTT detection function**
  - ***ALL instances of PTT Antenna are an instance of Antenna that has_function SOME instance of PTT detection function***

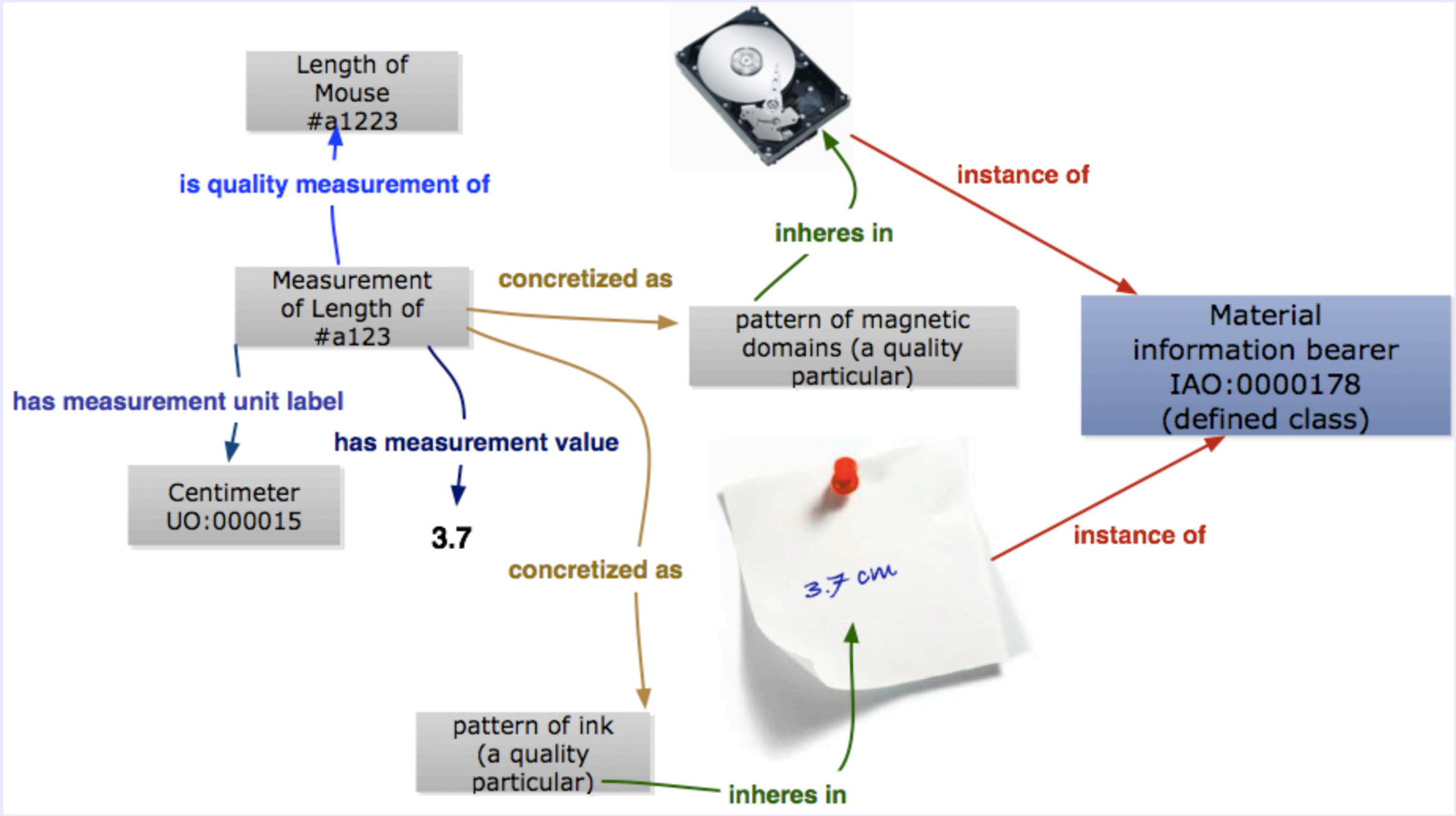
Information artifact ontology

- Aiming to be a mid-level ontology
- Information content entities
- Processes that consume or produce information content entities
- Material bearers of information
- Relations in which one of the relata are information content entities

Stance of IAO

1. Information entities are created by a sentient, intentionally, or by a machine made for that purpose.
2. Information is defined by what it is at creation.
3. Information content entities are related to other things by being "about" them – they are dependents, not independent entities

Information entity, concretized



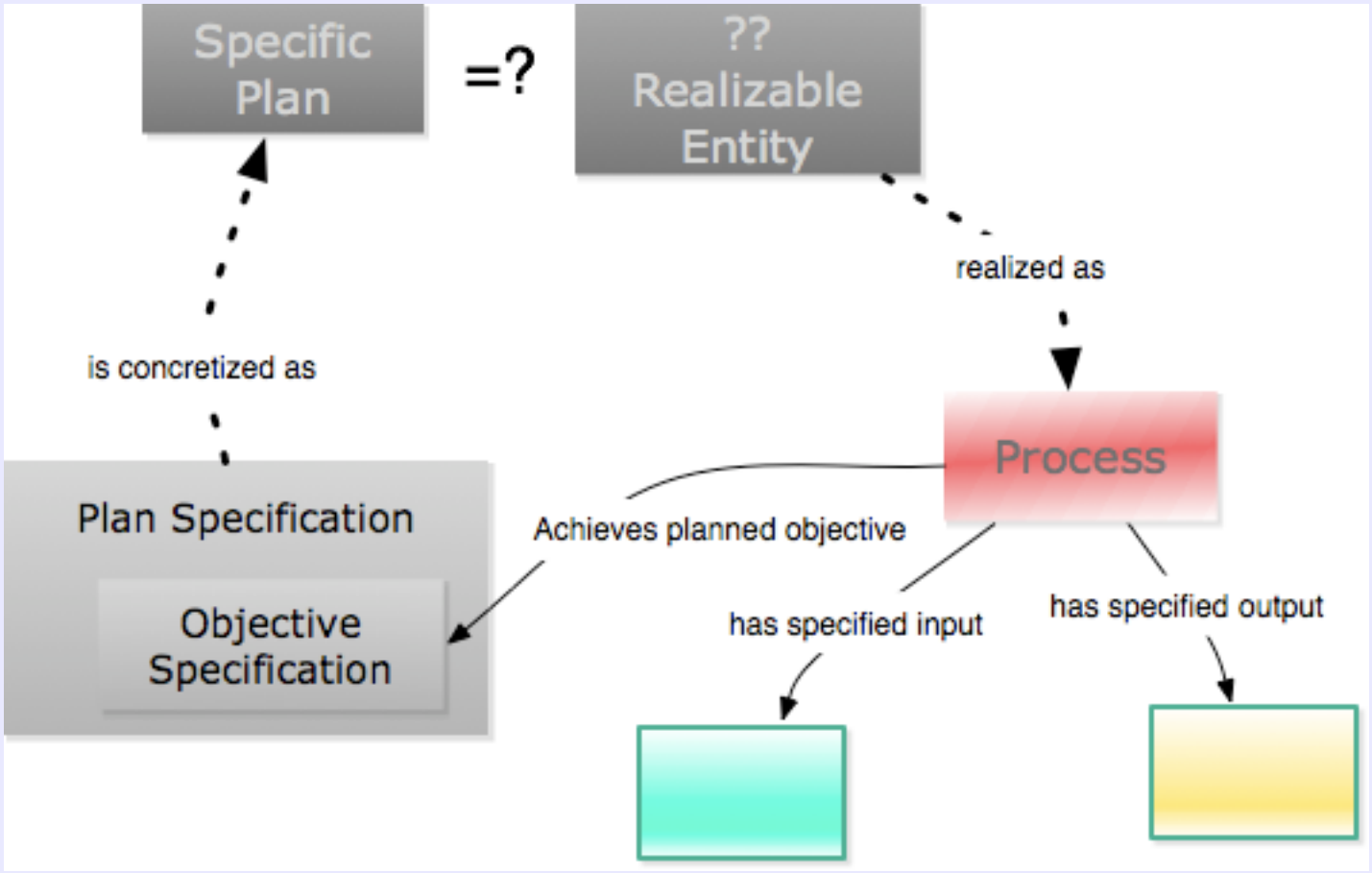
Examples of information content entities

- Novels
- Legal documents
- Charts
- Symbol
- Traffic directions
- Recipes
- Computer programs
- XML files
- File formats
- "Referring particulars"
- Ontologies
- Class descriptions
- Sentences
- URIs
- Simulation models
- Ideas
- Questions
- Hypotheses
- Databases
- Licenses
- Poems
- Journal article
- Rejection letters
- Things that can be true or false
- Propositions
- Advertisements
- Specifications
- Serial numbers
- Model numbers
- Formulas

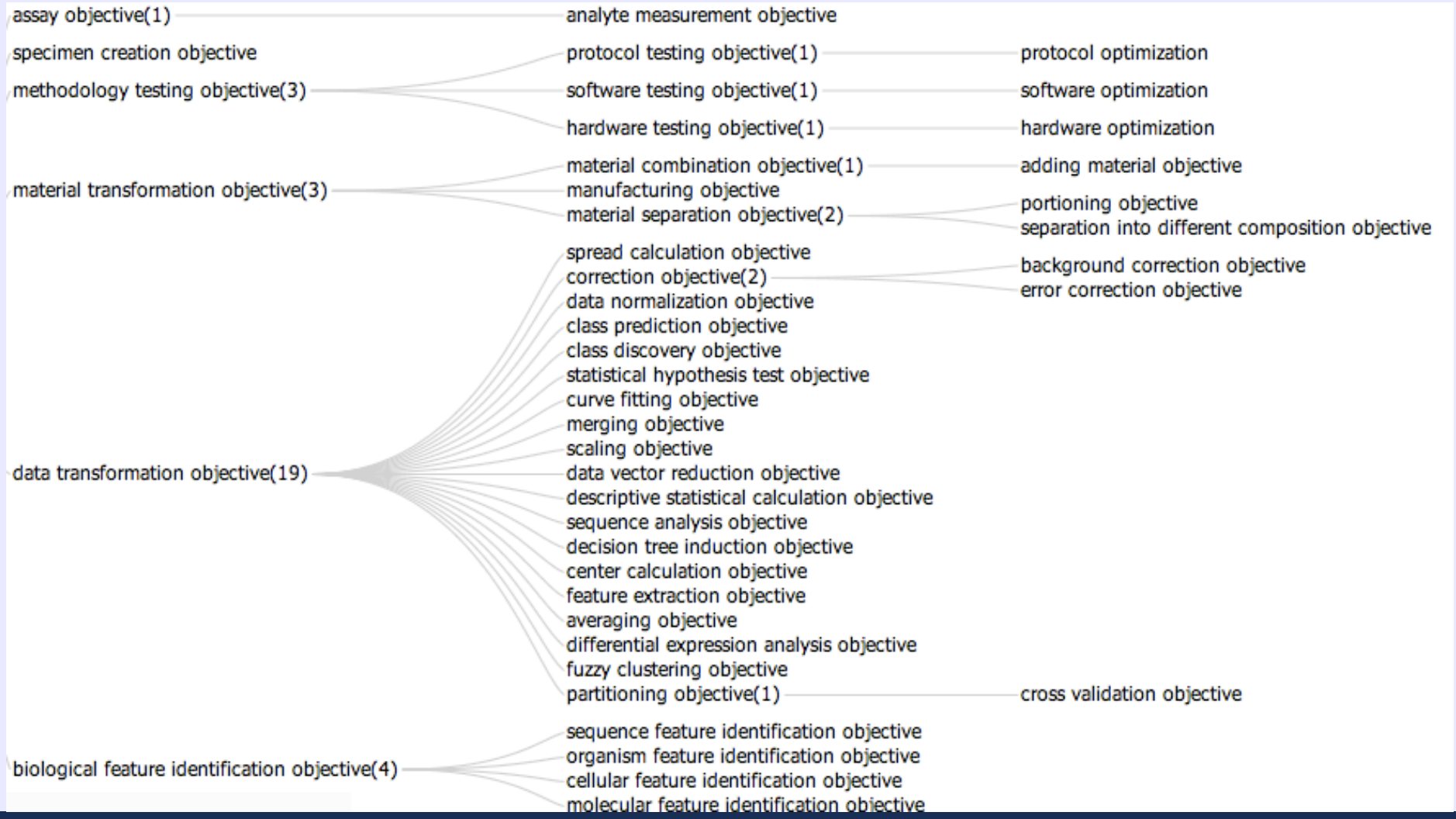
Examples of relations

- is about
- denotes
- mentions
- is measurement of
- encodes
- is rendering of
- is topic of
- makes claim
- cites
- is translation of

From the Ontology of Biomedical Investigations(OBI) Connecting plans, objectives, processes



Hierarchy below objective specification



Guidelines: The open world assumption

- Assertions we make are not (and need not be) complete
- What we don't state, or can not infer from what is stated, we have no knowledge of
- When we have all possibilities enumerated, we explicitly say so – "closure".
- Something to watch out for – ontology may expose detail and complication – this does not imply our representations must record all of that

Guidelines: Fewer relations are better

- Most queries use relations as connectives and bind classes or instances to variables
- The more relations we have, the more possible queries we have (from a topological point of view)
- Reasoning systems have more expressivity to reason about classes than about relations
- If we can constrain the number of possibilities, we have a higher chance that queries return results
- I start with the OBO Relations ontology and add sparingly as necessary

Guidelines: Reusing terms is hard, but essential

- Some existing OBO ontologies: (http://obofoundry.org)
  - Biological process (http://www.geneontology.org/)
  - Chemical entities of biological interest (http://www.ebi.ac.uk/chebi/)
  - Sequence ontology (http://www.sequenceontology.org/)
  - PATO – Phenotypes/Qualities (http://obofoundry.org/wiki/index.php/PATO:Main_Page)
- Consider creating or joining a community organized to share ontologies
- If we don't reuse we lose in at least two ways:
  - We have to redefine the terms we need
  - We have to do work to do data integration
- It's hard because we don't have adequate tools for making this easy – in may cases we haven't even *conceived* of what the form of such tools will be.

Links

Places to find more information

- http://ontology.buffalo.edu/smith/
- http://www.bioontology.org/wiki/index.php/Introduction_to_Biomedical_Ontologies
- http://www.bioontology.org/wiki/index.php/Dissemination_Wiki
- http://obofoundry.org/
- http://www.co-ode.org/resources/tutorials/
- http://sw.neurocommons.org/presentations/
- http://neurocommons.org/page/Main_Page
- http://www.ebi.ac.uk/ontology-lookup/
- http://www.geneontology.org/GO.contents.doc.shtml
- http://amigo.geneontology.org/cgi-bin/amigo/go.cgi
- http://wiki.reactome.org/index.php/Reactome_Curator_Guide
- http://tesuque.stanford.edu/pubsearch.org/AGP_help.html
- http://www.bioontology.org/wiki/index.php/PATO:Main_Page
- http://tools.immuneepitope.org/wiki/index.php/Curation_Manual
- http://www.yeastgenome.org/help/gotutorial.html
- http://obi.sourceforge.net/
- http://senselab.med.yale.edu/

Editing Tools

- http://www.phenote.org/
- http://oboedit.org/
- http://protege.stanford.edu/

The Neurocommons

- http://neurocommons.org/