# Semantics for interoperability of distributed data & models:

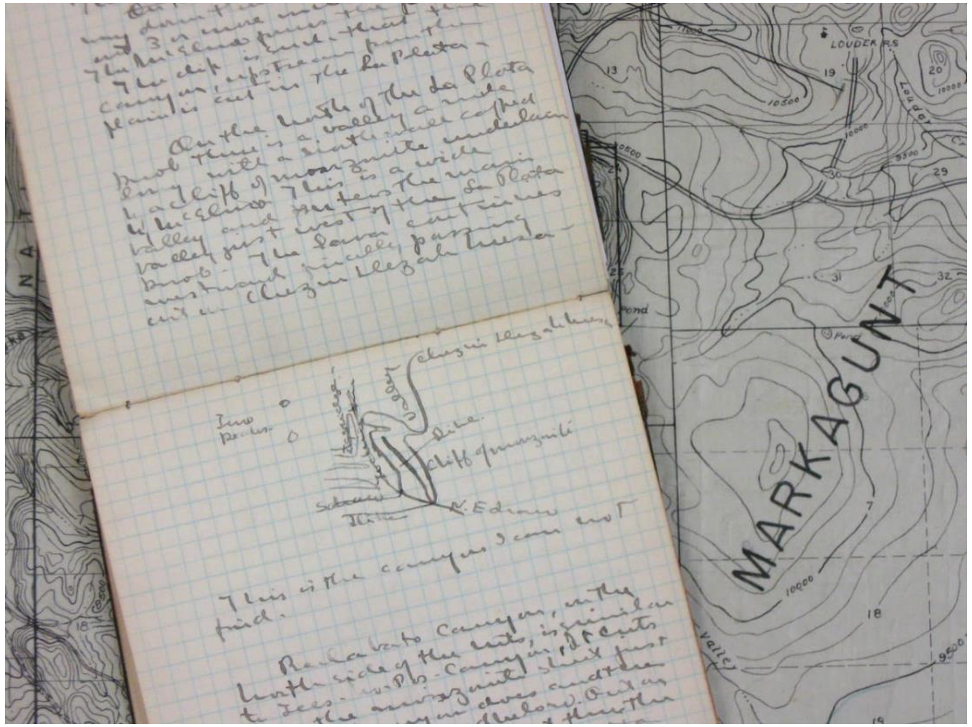
## Foundations for better connected information

bc3
BASQUE CENTRE FOR CLIMATE CHANGE
Klima Aldaketa Ikergai

International Spring University on Ecosystem Services Modeling

# Why hasn't this happened already?

- Movement to open data is well underway
- Semantics have worked for small disciplinary communities but so far have been very hard for interdisciplinary science
- General feeling that the semantic web has underperformed its promise
  - Need for a "killer app" that actually applies the semantic web to practical problems for science & society



# FAIR data stewardship principles (Wilkinson et al. 2016)

- Findable
- Accessible
- Interoperable
- Reusable
- FAIR+ (our interpretation): Information can be found, retrieved, linked, & operated upon in an *unsupervised way*, from multiple distributed repositories, with minimal risk of misalignment

# Types of ontologies

**CONTROLLED VOCABULARIES:**
Similar to domain ontologies; large number of terms

May use same vocabulary, even if logic is poorly thought out

**DOMAIN ONTOLOGIES:**
Define terms within a field

(e.g., SWEET, SPAN/SNAP, ENVO, Gene Ontology, PlantOntology)

**OBSERVATION ONTOLOGIES:**
How are scientific phenomena observed? (e.g., OBOE, O&M)

**FOUNDATIONAL ONTOLOGIES:**
Abstract, philosophical, high-level (e.g., DOLCE, SUMO, BFO)

How do we define a *scientific observable*,
and an *observation* of it?

- Three key dimensions make data *interoperable* & *reusable*:
  1. What is the observation about?
     Observable semantics (subject-quality-process-event)
  2. How is the observation carried out?
     Units, rankings, classifications: Properly annotated,
     a system could mediate between different units
  3. When and where is the observation carried out?
     Context and scale
- *Semantics first approach* (driving data collection, organization, processing, curation) vs. *annotation approach*

# Our approach

- Custom semantics & annotation language (k.IM)
  - Supported by open-source software (k.LAB)
  - Full support of FAIR+
  - Operates across domains of environmental & Earth systems modeling
- Move beyond "term matching" – textual metadata & controlled vocabularies
- Key requirements:

1. Fully compatible with accepted semantic web standards (OWL2)
2. Expressive, intuitively related to the scientific phenomena being described
3. Readable, as close as possible to English, to be easier to learn
4. Parsimonious, high descriptive power & flexibility – small core language to maintain logical consistency

# User types

| | |
|---|---|
| **SCIENTISTS/TECHNICIANS**: Annotate data & models using terms from domain ontologies with context-aware search tools | Science support staff |
| **DISCIPLINARY EXPERTS**: Build domain ontologies in collaboration with knowledge engineers | Research scientists |
| **KNOWLEDGE ENGINEERS**: Define semantic worldviews & guide development of logically consistent, parsimonious domain ontologies with disciplinary experts | Well-trained semantics experts |

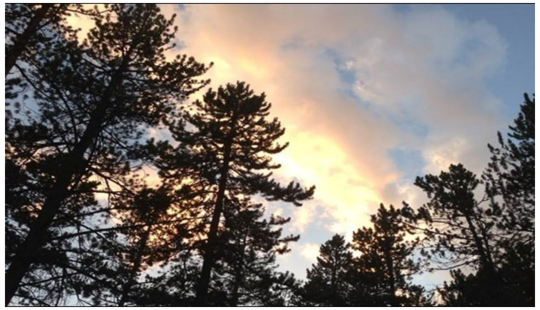
# Base observable & universal types

| Base category | Base concept | k.IM keyword | Explanation |
|---|---|---|---|
| | | **configuration** | Any combination of qualities or other arrangement that can be recognized by an observer without being identifiable as a single quality, subject or other observable. |
| continuant | subject | **thing** | Any inanimate physical body, as defined by an external observer. From the point of view of observation, cavities and observable "absences" are also things. |
| | | **agent** | Any physical body that is the context for autonomous processes that define its identity, including awareness of its own agency. |
| | quality | **priority** | A quality that can be ranked numerically, but no assumption is made about the scaling of the correspondent values beyond ordering. |
| | | **quantity** | A quality whose states are expressed numerically. |
| | | *physical properties* | k.IM provides keywords for the basic physical properties used in science, such as **temperature**, **energy**, **mass**, **volume**, **length**, **area**, among others. The keywords establish their physical nature (extensive or intensive) and enable validation of units of measurement. |
| | | **class** | A class is a special quality that **exposes** one or more traits of the context. Universals are not observable but a class allows to attribute "data" to describe combinations of traits. For example, land cover type can be seen as a combination of traits describing forests, urban texture etc. Using a class allows the semantics in complex classifications to be preserved and reasoned on. |
| | relationship | **structural relationship** | A relationship between two subjects whose existence does not depend on time, e.g., parent-child. |
| occurrent | process | **process** | A phenomenon that happens within a single subject and is observed as it evolves through time, creating change in the value of the subject's qualities and potentially creating or destroying subjects, events or relationships. |
| | event | **event** | Phenomena resulting from dynamic action that are seen as an atomic and countable at the temporal scale of observation. |
| | relationship | **functional relationship** | A relationship between two subjects whose observation implies a dynamic process, such as a flow of money between two commercial parties. |
| universal | trait | **attribute** | Any attribute that is not an ordering, realm, or identity. Attributes usually **describe** a quality that cannot be observed as a particular in the context of observation. |
| | | **ordering** | An attribute whose subclasses define an ordered list, e.g., high, medium, low |
| | | **realm** | An attribute that denotes |
| | | **identity** | An attribute used to identify a subject according to species, crop type, chemical element, etc. |
| | role | **role** | A function assumed by an observable when appearing in the context of another. A role adopted by an observable may **imply** other roles for observables related to it in that context. |

# Anything we can observe (with data) has a *subject*

- Countable, physical, recognizable object



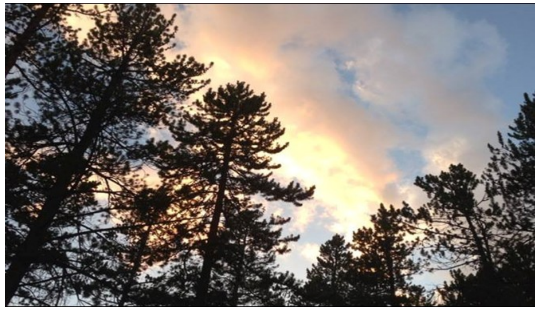



EXAMPLES

SUBJECTS: A mountain | A population of humans | A forest | A river

# Typical data describe a subject's specific *quality*

- Described by an *observer type* (measurement, count, percentage, proportion, etc.)



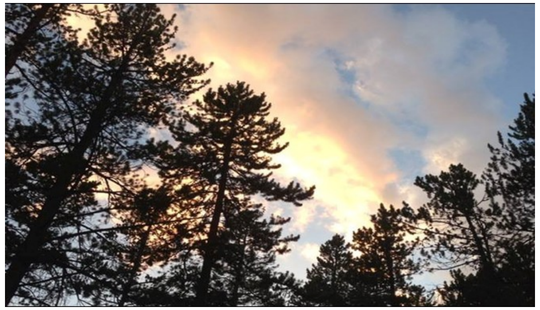
EXAMPLES

| | | | | |
|---|---|---|---|---|
| SUBJECTS: | A mountain | A population of humans | A forest | A river |
| QUALITIES: | Elevation (measurement, m) | Per capita income (value, $) | Percent tree canopy cover (%) | Stream order (ranking – $2^{nd}$) |

# Over time, subjects may experience *processes*

- Described by an *observer type* (e.g., measurement, count, percentage, proportion, etc.)



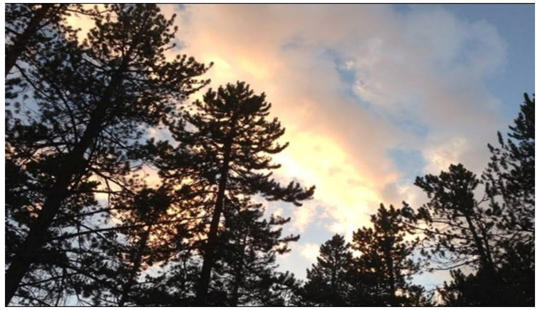



EXAMPLES

| | | | | |
|---|---|---|---|---|
| SUBJECTS: | A mountain | A population of humans | A forest | A river |
| QUALITIES: | Elevation (measurement, m) | Per capita income (value, $) | Percent tree canopy cover (%) | Stream order (ranking – $2^{nd}$) |
| PROCESSES: | Erosion (measurement, T/ha*yr) | Migration (people/yr) | Tree growth (T/yr) | Streamflow ($m^3$/sec) |

# A single, time-limited process is an *event*



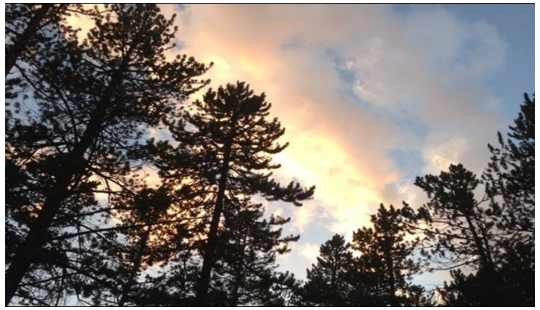



EXAMPLES

| | | | | |
|---|---|---|---|---|
| SUBJECTS: | A mountain | A population of humans | A forest | A river |
| QUALITIES: | Elevation (measurement, m) | Per capita income (value, $) | Percent tree canopy cover (%) | Stream order (ranking – 2nd) |
| PROCESSES: | Erosion (measurement, T/ha*yr) | Migration (people/yr) | Tree growth (T/yr) | Streamflow ($m^3$/sec) |
| EVENTS: | Snowfall | A birth | Death of a tree | A flood event |

# *Relationships* connect two subjects

- Structural & functional components
  (<u>Parenthood</u> connects <u>parents</u> to <u>children</u>; <u>Ecosystems</u> provide benefits to <u>human beneficiaries</u>
- Very important for agent-based models



EXAMPLES

| | | | | |
|---|---|---|---|---|
| SUBJECTS: | A mountain | A population of humans | A forest | A river |
| QUALITIES: | Elevation (measurement, m) | Per capita income (value, $) | Percent tree canopy cover (%) | Stream order (ranking – 2nd) |
| PROCESSES: | Erosion (measurement, T/ha*yr) | Migration (people/yr) | Tree growth (T/yr) | Streamflow ($m^3$/sec) |
| EVENTS: | Snowfall | A birth | Death of a tree | A flood event |

RELATIONSHIPS: ↖ Skiers using a mountain for recreation ↗ ↖ A city using a river for water supply ↗

# Observables can also have one or more *traits*

- "Adjectives" that add descriptive power to further modify a concept
- Add flexibility without adding more complexity to the ontologies

| 1. ATTRIBUTES | 2. IDENTITIES | 3. REALMS | 4. ORDERINGS |
| --- | --- | --- | --- |
| (Temporal, frequency, min/max/mean, etc.) | (Authoritative species or chemical names) | (strata - Soil, atmosphere, ocean, forest) | (High-Moderate-Low) |

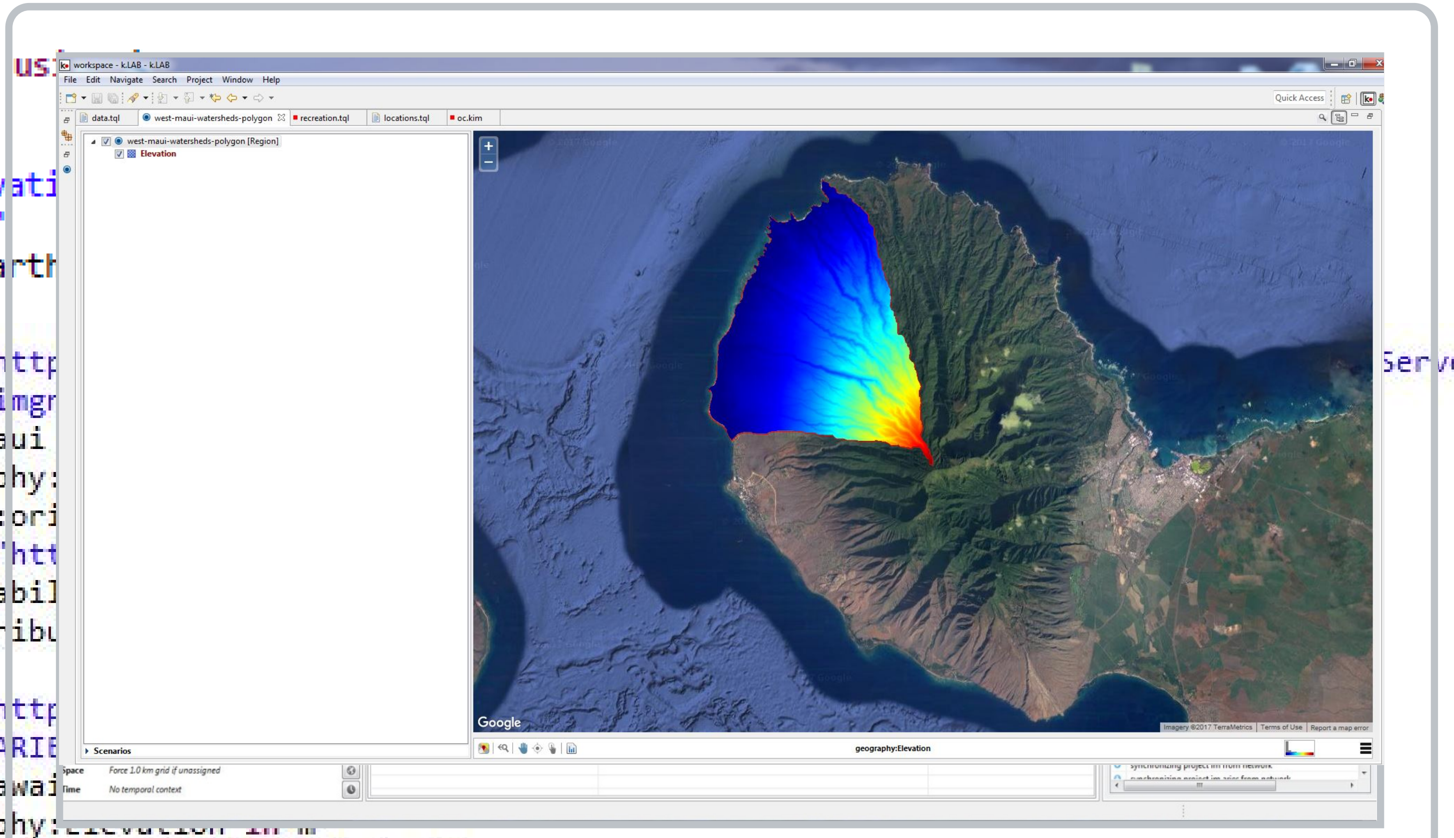
# Defining, annotating, & observing concepts

```
namespace geography us

length Elevation
  "Geographical elevati
   elevation model."
  is im:Height of earth

model wcs(service = "http                                                Server",
          id      = "imgr
    named elevation-maui
    as measure geography:
    with metadata {dc:ori
              dc:url "htt
              im:reliabil
              im:distribu

model wcs(service = "http
          id      = "ARIE
    named elevation-hawai
    as measure geography:
    with metadata {dc:comment "NOAA IfSAR data"
        dc:url "http://coast.noaa.gov/digitalcoast/data/coastalifsar"
        im:reliability 95
        im:distribution "restricted"};
observe earth:Region named west-maui-watersheds-polygon
   over space (shape="EPSG:4326 MULTIPOLYGON (((-156.6405400269599 21.022965540999312, -156.63881777157485 21.0223997822568
```



# Attributes & their types

- Enable a construction of a large, flexible, yet parsimonious & logically consistent system

```
namespace ecology using chemistry, earth;

abstract attribute Salinity
  describes chemistry:Salinity within earth:Aquatic earth:Region
  has children
    Saline,
    Brackish,
    Freshwater;


namespace chemistry using im, physical;

quantity Salinity
  is proportion of (NaCl im:Mass) to (Water im:Mass) within physical:DelimitedBody;
```

# *Semantic observers* produce observations of concepts

| Prototype | Description |
|---|---|
| **measure** <O> **in** <unit> | Specifies the unit for a concrete physical property and ensures that it is compatible with the physical nature and the spatiotemporal context of use. Ensures that units are converted when dependencies are matched to data. |
| **rank** <O> [min **to** max] | Used with priorities, can specify a scale for bounded ranks and ensure that scales are properly converted when dependencies are matched to data. |
| **classify** <O><br>[**into** <O1 [**if** <condition>], O2, …>]<br>[**according to** <metadata field>]<br>[**as identified by** <authority>] | Used with **class** concepts, enables many useful ways of specifying the semantic content of categorical classifications; in addition to the direct specification of the concepts that each possible value or range of values should map to, it allows specifying metadata for conversion (e.g. the standard encodings of categories in common land cover datasets) and to match values to concepts by converting identifiers through a specified authority. |
| **value** <O> [**over** <O2>]<br>[**in** <currency>]<br>[min **to** max] | Values can be direct or relative (an example of the latter is the pairwise comparisons used in multiple criteria analysis) and refer to a currency (monetary or conceptual) or have a scale like in the case of rankings. When the currency is monetary, a year must also be specified; k.LAB contains functionalities that bridge to conversion services so that values can be adjusted for inflation and converted to different currencies in many cases. |
| **distance to** <O> **in** <length unit> | A distance observer will observe all the objects of the type mentioned in the context of observation and compute the distance to them. In k.IM, this observer can also be used with reference to the URI of a specific observation, which can be located anywhere. |
| **count** <O> [**per** <extent unit>] | Count observers observe all the objects of the type mentioned and produce their numerosity, if necessary distributed over space and/or time. A count concept is produces unless O is already a count. |
| **ratio [of]** <O1> **[to** <O2>] | Ratio observers describe ratios between qualities. A ratio concept is produces unless O is already a ratio. |
| **proportion [of]** <O1> **[in** <O2>]<br>**percentage [of]** <O1> **[in** <O2>] | Proportion and percentage are differently scaled ways to observe a proportion concept, which is created according to rules in Table 2 unless O1 is already a proportion. |
| **uncertainty [of]** <O> | The numeric scaling and computation of uncertainty is not mandated in k.IM. In k.LAB, currently, numeric uncertainties are computed as standard deviations of probability distributions, and the Shannon index of diversity is used for categorical information. |
| **probability [of]** <O> | Probability observers validate their data in the [0-1] interval. A probability concept is produces unless O is already a probability. |
| **occurrence [of]** <O> | This observer is a “fluent” shorthand to specify the probability of a presence. |
| **presence [of]** <O> | Validates data as boolean (true/false). A presence concept is produced unless O is already a presence. |

# Authorities

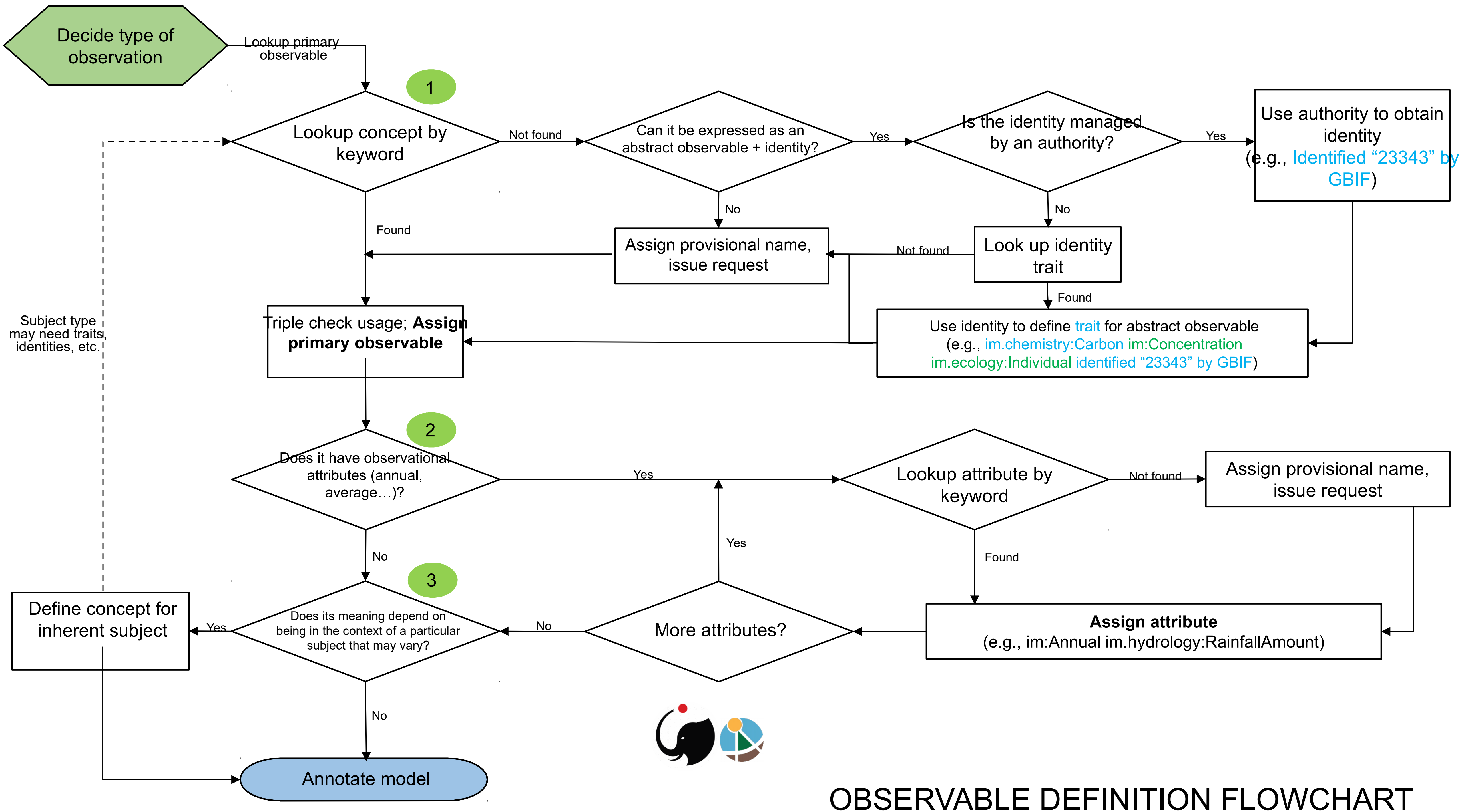
- Reuse well-accepted domain ontologies & controlled vocabularies: GBIF (biological taxonomy), IUPAC (chemical elements & compounds), Soil WRB (soil), AGROVOC (agriculture)
  - For honeybees (*Apis mellifera):*

```
model raster("data/bees.tif")
  as count biology:Individual identified as "1341976" by GBIF.SPECIES per km²;

agent HoneybeeIndividual
  is biology:Individual identified as "1341976" by GBIF.SPECIES;

model raster("data/bees.tif")
  as count HoneybeeIndividual per km²;
```

- *Bridging authorities* could mediate between domain ontologies/controlled vocabularies from the same field (not yet attempted).




OBSERVABLE DEFINITION FLOWCHART

# Benefits & challenges

- Benefits:
  1. Clear focus on how foundational, observation, and domain ontologies fit together to clearly define scientific observables
  2. Simple phenomenology to describe observables
  3. Distributed, web-based language and software enforces consistency but allows uncoordinated use & expansion to appropriate domain ontologies/controlled vocabularies, all in support of FAIR+
- Challenges: Use across larger, more diverse communities