Semantic web for scientific information

Streamlining how we write, find, link and reuse data and models

Ferdinando Villa
&
Integrated Modelling Partnership

bc³
BASQUE CENTRE
FOR CLIMATE CHANGE
Klima Aldaketa Ikergai

aquacross

An integrated solution for shared, distributed, collaborative modelling

SEMANTICS FOR DATA AND COMPUTATIONS

- Maintenance of the core conceptualization and language
- Maintenance and delivery of a **shared worldview** (ontologies) for cross-domain communication

APPLICATIONS

- Ecosystem services assessment (ARIES)
- Real-time monitoring using remotely sensed data
- Food and other environmental securities
- Integrating hydrology, primary production, nutrients with agent models of SES.

OPEN SOURCE SOFTWARE

- User-end (modelers and end users)
- Server technology (institutions)
- Developer team and user support

COLLABORATIVE MODELING

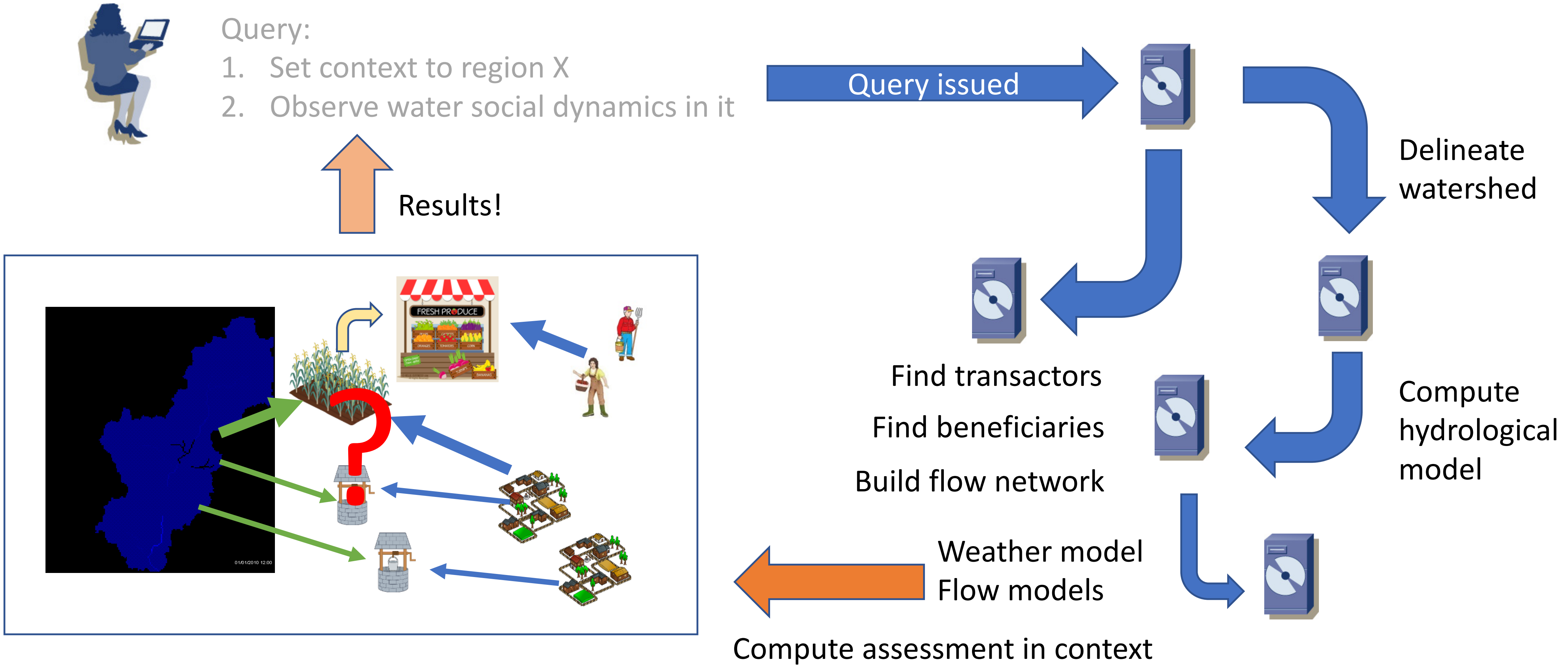
- Interoperable data and models
- Serving models on the Web
- Direct support of partner projects
- International Spring University since 2013

INTEGRATED MODELING INFRASTRUCTURE

- Assembly of models from networked data and model components
- Partners can manage their servers or use the partnership's

Models and data live on a semantic web

An extensible network hosts data, models and model services available to users

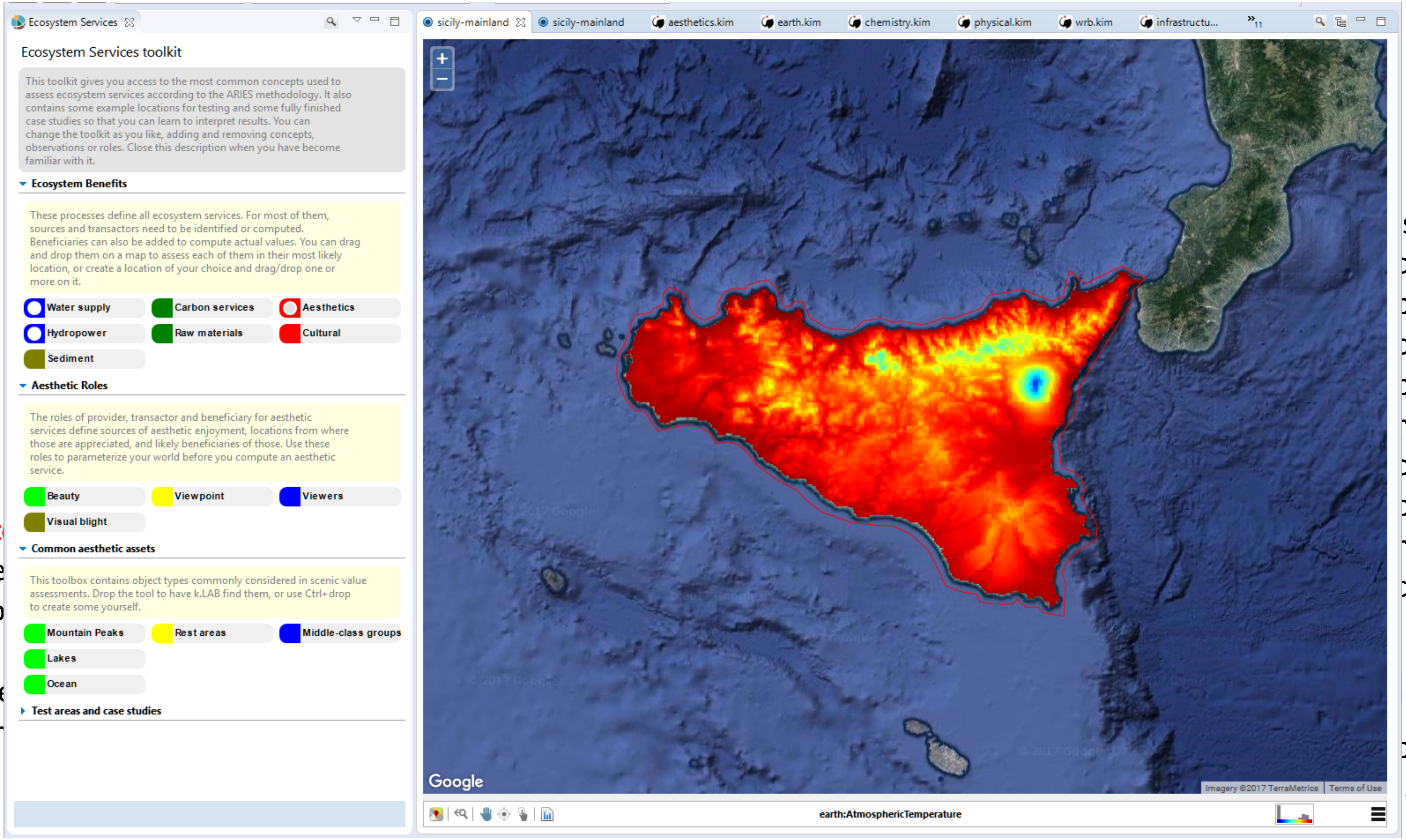


A user’s perspective: two-step assessment

Client software (desktop & soon web-based) allow modeling with minimal configuration and training. Provenance info is compiled into user documentation for each result set.

St
Se
co

(se
or



Drag-and-drop paradigm for end users

“Palette” of ES tools can store finished studies and scenario results

stem creates agents and processes from ontological specs builds best-case model out of components and data on the semantic network computes it…

Models are built and computed when user drops the concept

Full reports are built to document the computation

ogged into network secure certificate

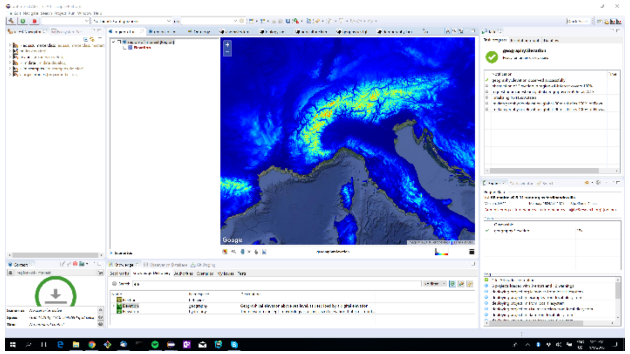
A modeler's/data scientist's perspective

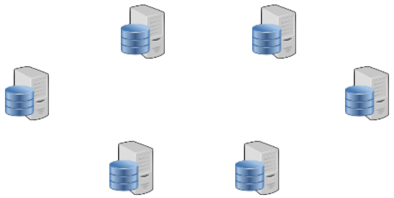
streamlining and supporting the annotation workflow for both data and computations

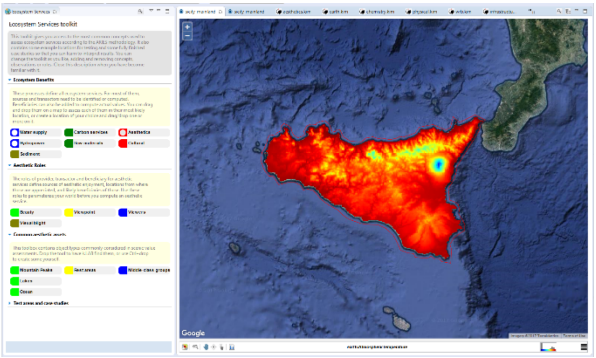
Raster DEM

Synchronized domain ontologies (worldview)

```
model data/dem.tiff as geography:Elevation in m;
```







One language for knowledge, data and computations

Uses English-like syntax to express the observation ontology; linguistic approach keeps ontologies small and learnable

```
model hydrology:SurfaceWaterFlow,
    hydrology:RunoffVelocity,
    hydrology:RunoffWaterVolume named runoff
  observing
    earth:PrecipitationVolume,
    geography:Elevation,
    earth:AtmosphericTemperature,
    earth:Stream
  using
      hydrology.swat.distributed()
  over time (step = "1 day")
      do [
          ... actions
      ],
      change runoff to [
          ...
      ]
;
```

Hydrological model

```
learn value of ecology:Biodiversity 0 to 100
   observing
      conservation:Protected ecology:Vegetation earth:Site as im:Archetype,
      not conservation:Protected ecology:Vegetation earth:Site as im:Archetype,
      geography:Aspect in degree_angle as im:ExplanatoryQuality,
      geography:Slope in degree_angle as im:ExplanatoryQuality,
      im:Annual im:Mean earth:AtmosphericTemperature in Celsius as im:ExplanatoryQuality,
      im:Annual  earth:PrecipitationVolume in mm as im:ExplanatoryQuality,
      distance from landcover:UrbanFabric earth:Region in m as im:ExplanatoryQuality,
      distance from earth:Coast in m as im:ExplanatoryQuality,
      distance from infrastructure:Road in m as im:ExplanatoryQuality,
      proportion of soil:Silt in soil:TopSoil im:Volume as im:ExplanatoryQuality
   using weka.bayesnet();
```

Machine learning model

Ontological statements read as English and are validated while editing. Inconsistent concepts are flagged as errors and discarded.

Color coding, assisted editing, and informative error messages help user

Often dozen of OWL axioms in 1-2 lines

```
thing Coast
   "A portion of land adjacent to a major marine or lacustrine water body."
   is earth:Terrestrial earth:Region
       adjacent to (earth:Marine or earth:Lacustrine) earth:Region;

thing Coastline
   "The boundary between land and an adjacent @Coast."
   is im:Boundary
       of (earth:Region adjacent to (earth:Marine or earth:Lacustrine) earth:Region)
       adjacent to (earth:Marine or earth:Lacustrine) earth:Region;
```

from IM worldview (general users receive it from the network and search it)

Worldviews merge domains and vocabularies reliably and intuitively

```
namespace chemistry
  using im, physical
  in domain im:Chemistry;
```

```
abstract identity Compound
  "Concrete subclasses of Compound must be identified with
   an InChl code validated by the IUPAC authority."
  is ChemicalSpecies
  requires authority IUPAC
  has disjoint children
      (Water identified as "1S/H2O/h1H2" by IUPAC),
      (CO2   identified as "1S/CO2/c2-1-3" by IUPAC),
      (NH3   identified as "1S/H3N/h1H3" by IUPAC),
      (H3O   identified as "1S/H2O/h1H2/p+1" by IUPAC),
      (SO4   identified as "1S/H2O4S/c1-5(2,3)4/h(H2,1,2,3,4)" by IUPAC),
      (NaCl  identified as "1S/ClH.Na/h1H;/q;+1/p-1" by IUPAC);

quality MassConcentration
  is ratio of ${inherent extends ChemicalSpecies} im:Mass
           to ${context  extends ChemicalSpecies} im:Volume;

@origin("SWEET")
quality Acidity
  "Capability of a molecule to donate a hydron (proton or hydrogen ion H+)."
  is MassConcentration of H3O within Water
;

quantity Ph
  "A measure of acidity or alkalinity of an aqueous solution used universally, with
   values of 7 indicating neutrality. Computed as the negative logarithm of the activity
   of H ions."
 decreases with Acidity
;
```

Explicit domains

Assisted editor with as-you-type syntax validation

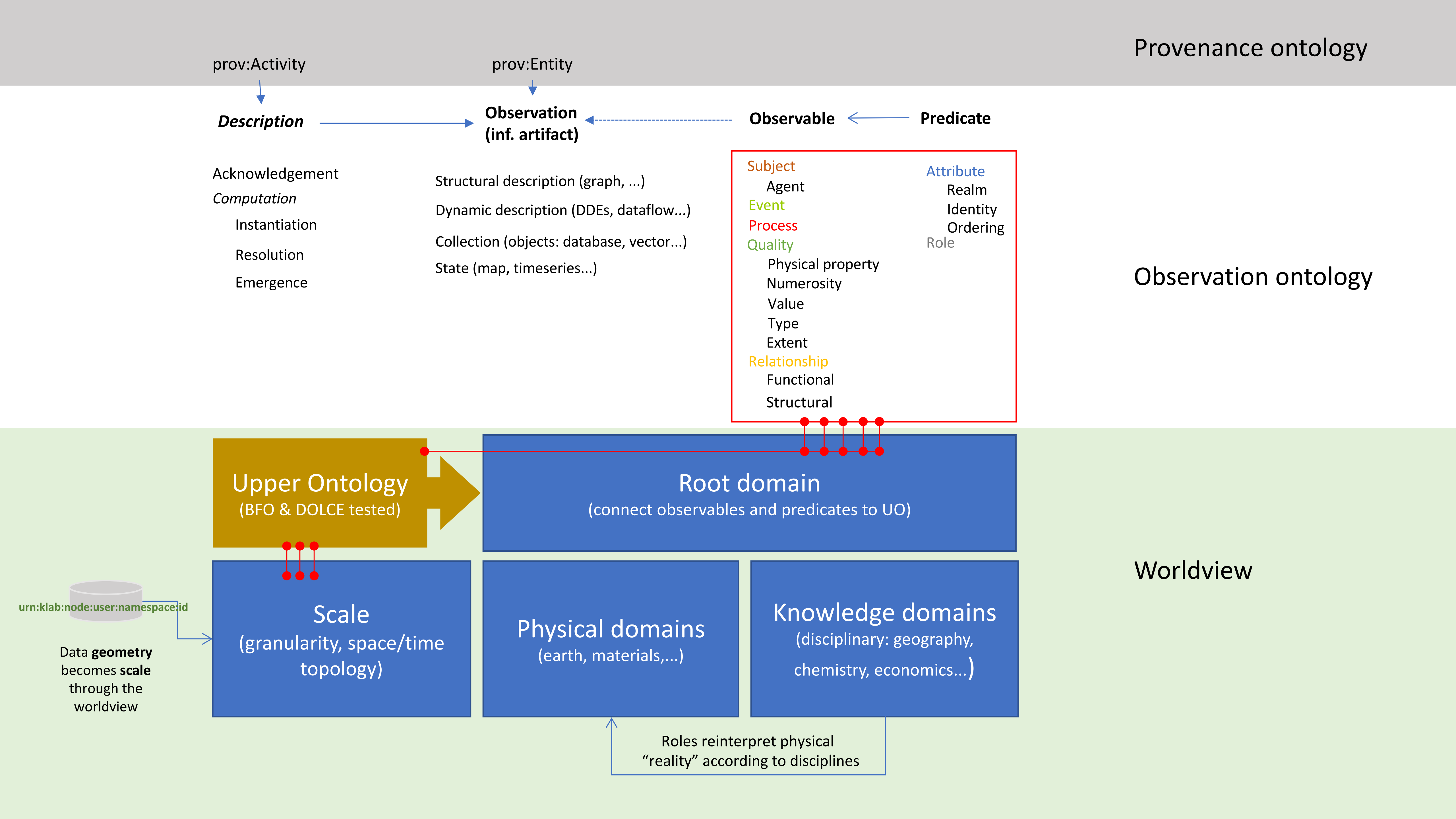
- syntax embodies observation semantic
- predicate and observable composition
- usage of attributes

Logical validation at each save

- Uses reasoner of choice
- Consistency of usages with upper ontology
- (potentially nested) inherency...

Bridging to vocabularies through endorsed **authorities**

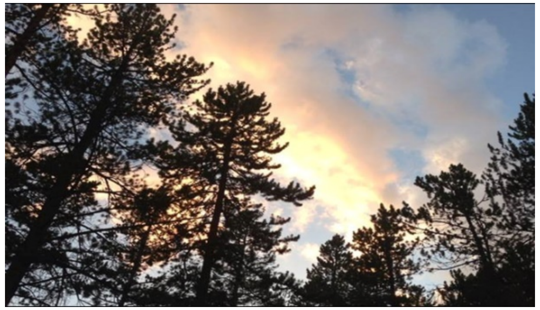
- GBIF (taxonomic identities)
- IUPAC (chemical identities)
- WRB (soil identities)
- AGROVOC (agri processes and practices)
- ... (plug-in)


Provenance ontology
prov:Activity
prov:Entity
Description
Observation
(inf. artifact)
Observable
Predicate
Acknowledgement
Computation
Instantiation
Resolution
Emergence
Structural description (graph, ...)
Dynamic description (DDEs, dataflow...)
Collection (objects: database, vector...)
State (map, timeseries...)
Subject
Agent
Event
Process
Quality
Physical property
Numerosity
Value
Type
Extent
Relationship
Functional
Structural
Attribute
Realm
Identity
Ordering
Role
Observation ontology
Upper Ontology
(BFO & DOLCE tested)
Root domain
(connect observables and predicates to UO)
Worldview
urn:klab:node:user:namespace:id
Data geometry becomes scale through the worldview
Scale
(granularity, space/time topology)
Physical domains
(earth, materials,...)
Knowledge domains
(disciplinary: geography, chemistry, economics...)
Roles reinterpret physical "reality" according to disciplines

A semantics-first approach - for wide user groups

Address all the "W's of information – what, where, when, why, and how – without becoming too large or complex to learn and use.





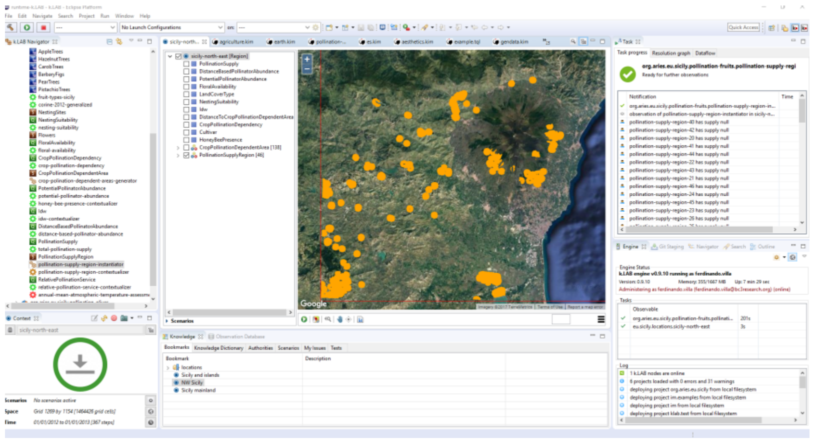

SUBJECTS:	A mountain	A group of humans	A forest	A river
QUALITIES:	Elevation (measurement)	Per capita income (value)	Percent tree canopy cover (%)	Stream order (ranking)
PROCESSES:	Erosion	Migration	Tree growth	Streamflow
EVENTS:	Snowfall	A birth	Death of a tree	A flood event
RELATIONSHIPS:	↖ Skiers using a mountain for recreation ↗		↖ A city using a river for water supply ↗	

Semantics for **predicates** allow to compose attributes, realms and identities without inheritance; interface to vocabularies
Roles account for **usages** of general observables in disciplinary contexts without giving up consistency and FAIR goals

Tooling (1): k.IM language and support software

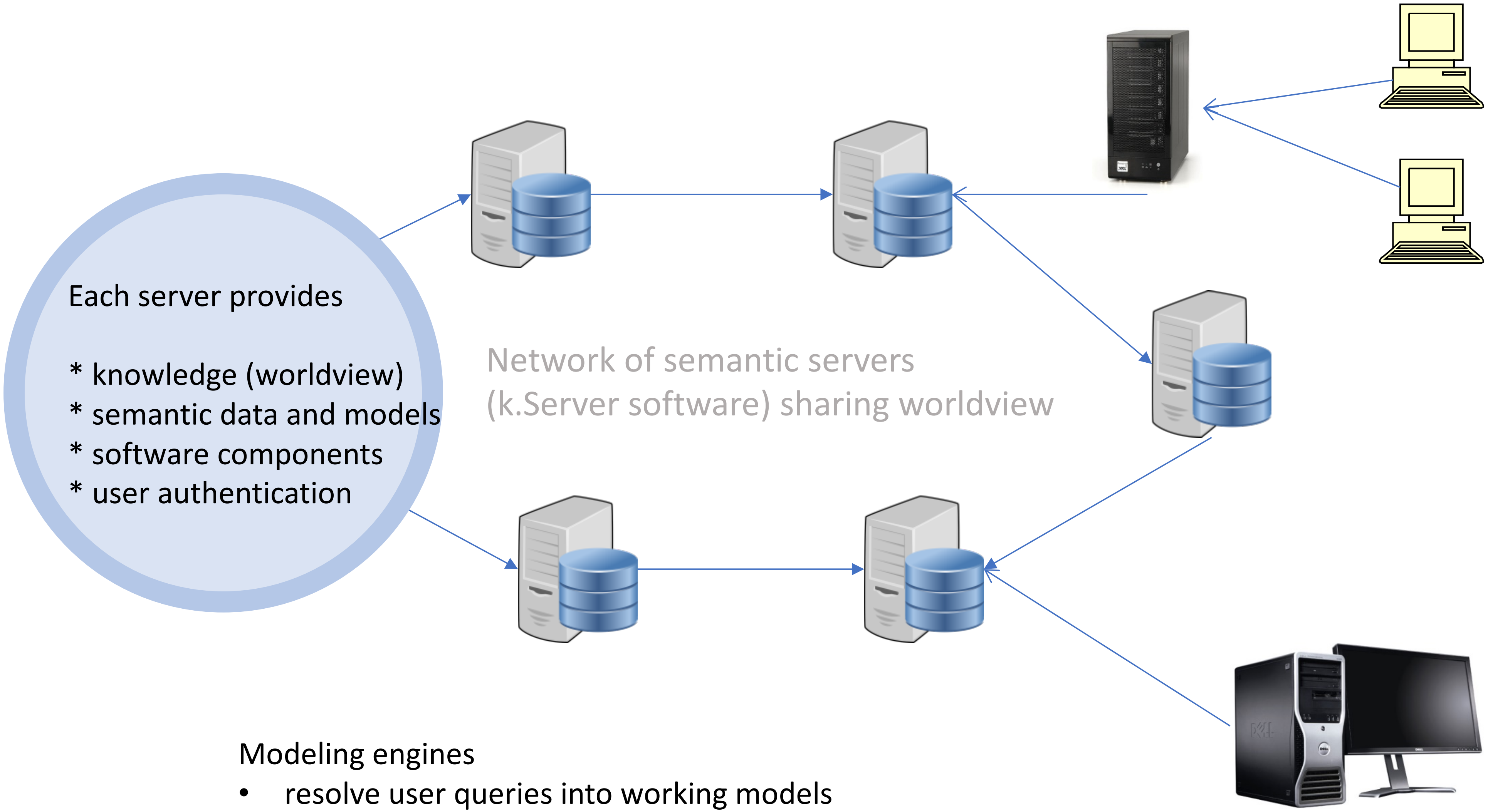
```
role PollinatorSupplier
    is ses:Provider
    applies to earth:Region
    implies PollinatorAbundance as ses:Supply;

role AgriculturalProductionDependent
    is ses:Beneficiary
    implies PollinatedYield as ses:Demand
    applies to observation:Subject;

/**
 * Roles that define the P->T and B->T relationships.
 */
role PollinationSupplyConnection
    is ses:ProvisionFlow
    applies to im:MatterTransferConnection between PollinatorSupplier and PollinationDependent;

role AgriculturalUseConnection
    is ses:UseFlow
    applies to im:MatterTransferConnection between AgriculturalProductionDependent and PollinationDependent;

/**
 * Role for the ES, tying everything together.
 */
role PollinationEcosystemBenefit
    "The benefit obtained by any user of the yield made possible by pollination. This is
     easier to monetize than most ES when defined this way."
    is ses:ProvisioningEcosystemBenefit
    implies at least 1 PollinationSupplyConnection, at least 1 AgriculturalUseConnection
;

/*
```

The k.IM language is used to express both the worldview and the data/models that use it



- Tools and interfaces enable end users, modelers, and network administrators
- Simplify the tasks of semantically describing, coding, and publishing data and models.
- Provide and maintain documentation, community resources for discussion, user support and bug reporting
- Create tools for participatory, graphical model building that can be directly translated into templates for working models.

Tooling (2): distributed semantic web infrastructure



Web users connect to modeling engine and use web interface for modeling

UI is drag-and-drop only, optimized for non-technical user

Modelers have a private engine and connect to the network using the modeling IDE

UI is optimized for developers and modelers

Modeling engines

- resolve user queries into working models
- find content on the semantic network

The Integrated Modelling Partnership

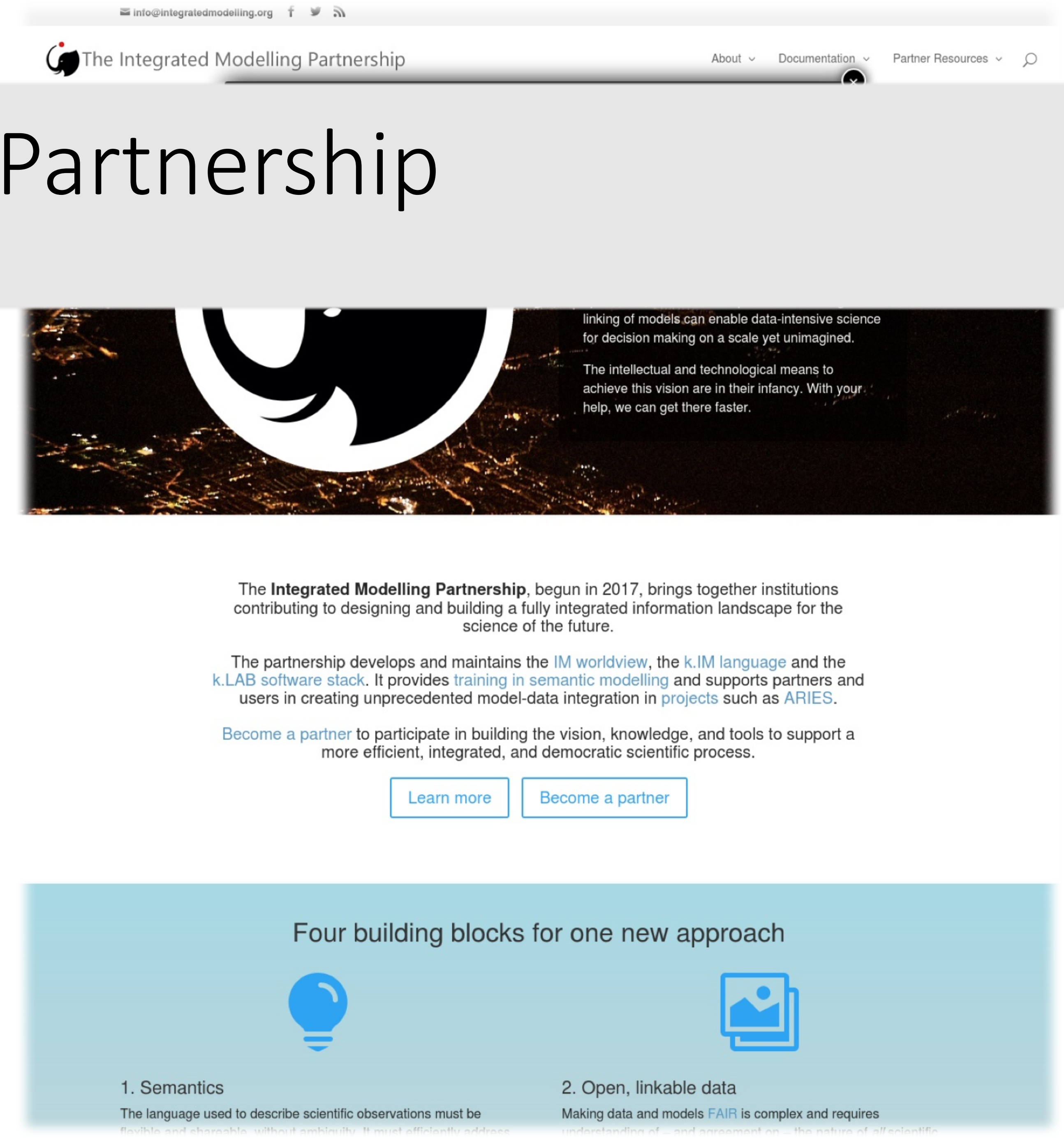
http://www.integratedmodelling.org

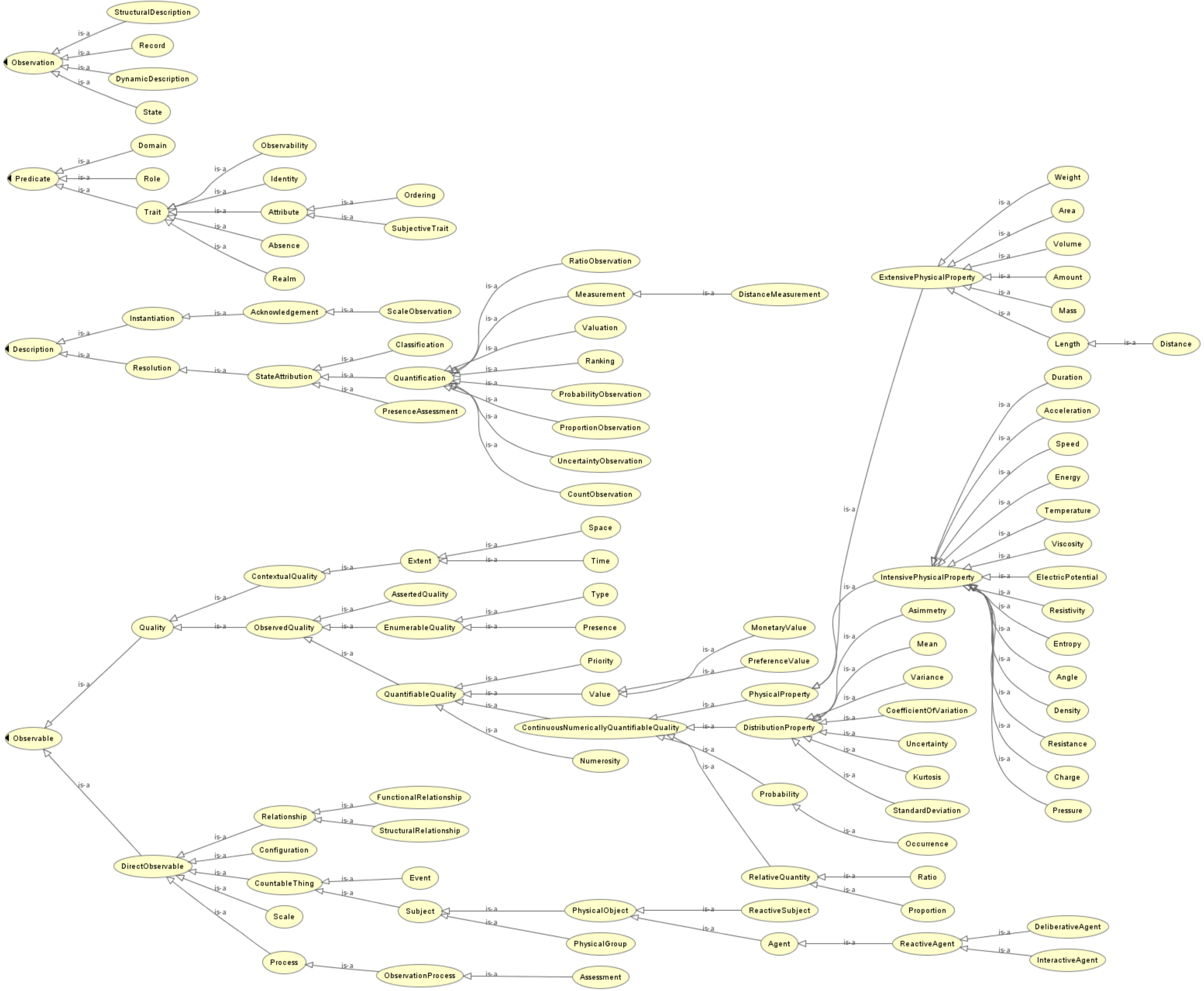
Partners share:

- **Participation**
  - design their applications in collaboration with the core staff
  - Work Packages drive the development of larger initiatives
- **Ownership**
  - all the products (worldview, software, online courses) are open source/open access
  - Products bear the copyright of the partnership with its member institutions.
- **Control**:
  - partners enter the steering committee that defines activities, governance and directions of development

Maintains:

- Downloadable software and documentation for users and developers.
- Certification for users and institutions
- The shared *worldview for all domains, prioritizing needs of partners*.
- Online data and models annotated with worldview semantics and identified by unique
- Supports partners in deploying their own servers and modeling engines.
- Online courses and training material on integrated modelling.
- Public issue tracking, knowledge base and developer online support channels.





StructuralDescription
Record
Observation
DynamicDescription
State
Domain
Predicate
Role
Trait
Observability
Identity
Attribute
Ordering
SubjectiveTrait
Absence
Realm
Instantiation
Acknowledgement
ScaleObservation
Description
Resolution
StateAttribution
Classification
Quantification
PresenceAssessment
RatioObservation
Measurement
DistanceMeasurement
Valuation
Ranking
ProbabilityObservation
ProportionObservation
UncertaintyObservation
CountObservation
ExtensivePhysicalProperty
Weight
Area
Volume
Amount
Mass
Length
Distance
IntensivePhysicalProperty
Duration
Acceleration
Speed
Energy
Temperature
Viscosity
ElectricPotential
Resistivity
Entropy
Angle
Density
Resistance
Charge
Pressure
Space
Time
Extent
ContextualQuality
AssertedQuality
Type
Quality
ObservedQuality
EnumerableQuality
Presence
Observable
Priority
QuantifiableQuality
Value
MonetaryValue
PreferenceValue
PhysicalProperty
ContinuousNumericallyQuantifiableQuality
DistributionProperty
Asimmetry
Mean
Variance
CoefficientOfVariation
Uncertainty
Kurtosis
StandardDeviation
Numerosity
Probability
Occurrence
RelativeQuantity
Ratio
Proportion
Relationship
FunctionalRelationship
StructuralRelationship
DirectObservable
Configuration
CountableThing
Event
Subject
Scale
PhysicalObject
ReactiveSubject
PhysicalGroup
Agent
ReactiveAgent
DeliberativeAgent
InteractiveAgent
Process
ObservationProcess
Assessment
is-a

Promoting **semantics-first** solutions for open, linkable data and models

- In today's dialogue on interoperability, *information* equals *data*. Both data and models can be seen as ways to make scientific observations – *definitions* of observations. Doing so enables a consistent discussion on how to semantically connect data to models and how to build complex models by assembling simpler ones.
- Semantics-first data/models are *first-class research objects* that can be found online, read and understood by computers and humans alike. They can reuse existing vocabularies and thesaura while ensuring consistent semantics throughout the information landscape.
- Powered by semantics, artificial intelligence can transparently match data and models to a chosen time, place, problem, and (multiple!) scales.
- **Much of the complexity of building and running models can be handled by machines, with substantial advantages for science and decision making**.

The challenge of data/model integration and reuse

Scientists in the past collected data in notebooks. In the digital age, we want scientific data and models to be FAIR - Findable, Accessible, Interoperable, and Reusable, to ensure their maximum value.

A fully connected information landscape using open, safe, accurate, "Wikipedia-like" sharing and linking of models can enable data-intensive science for decision making on a scale yet unimagined:

1. **reuse** the abundance of data and specialized knowledge available and needed to analyse social and natural processes (and their interactions)
2. **avoid** the risk of **fragmentation** hidden in the use of ad-hoc (or no) semantics to describe data
3. enable **simple user workflows** in modelling, supporting **direct** questions like: What is the social dynamics of water in basin X? How does switching to crop Y affect rural food security in region Z?

Where are we along this path in 2017?

Using and reusing data: The state of the art

1. Distributed access to datasets over the web (OGC, OpenDAP, …)
2. Linked Open Data paradigm: open standards, each artifact is coupled with a URI pointing to its “meaning”.
3. Problem: the meaning *differs for each observer* - unless semantics is coherent across domains, uses and goals.
4. If it’s not consistent, it’s not FAIR

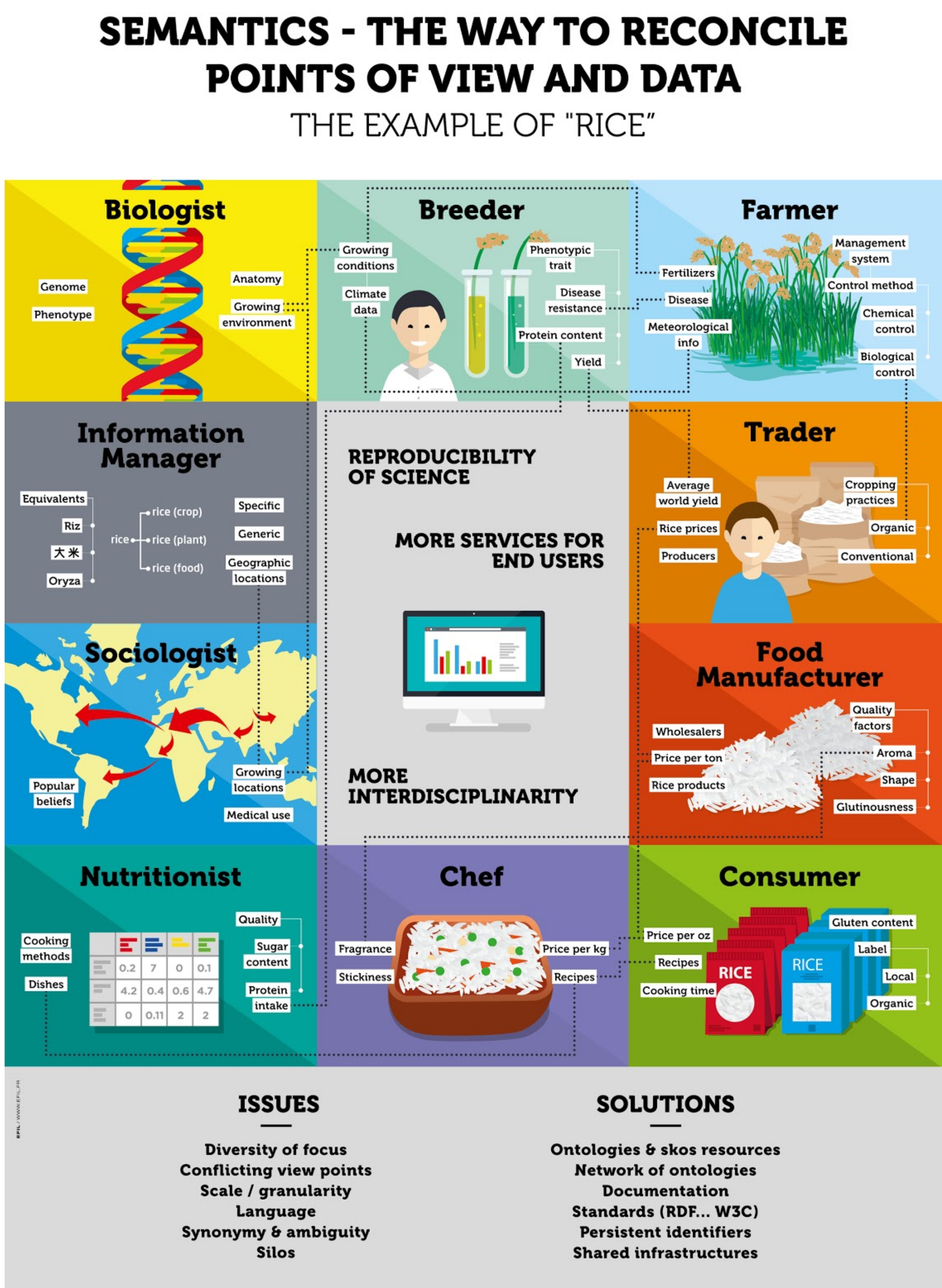


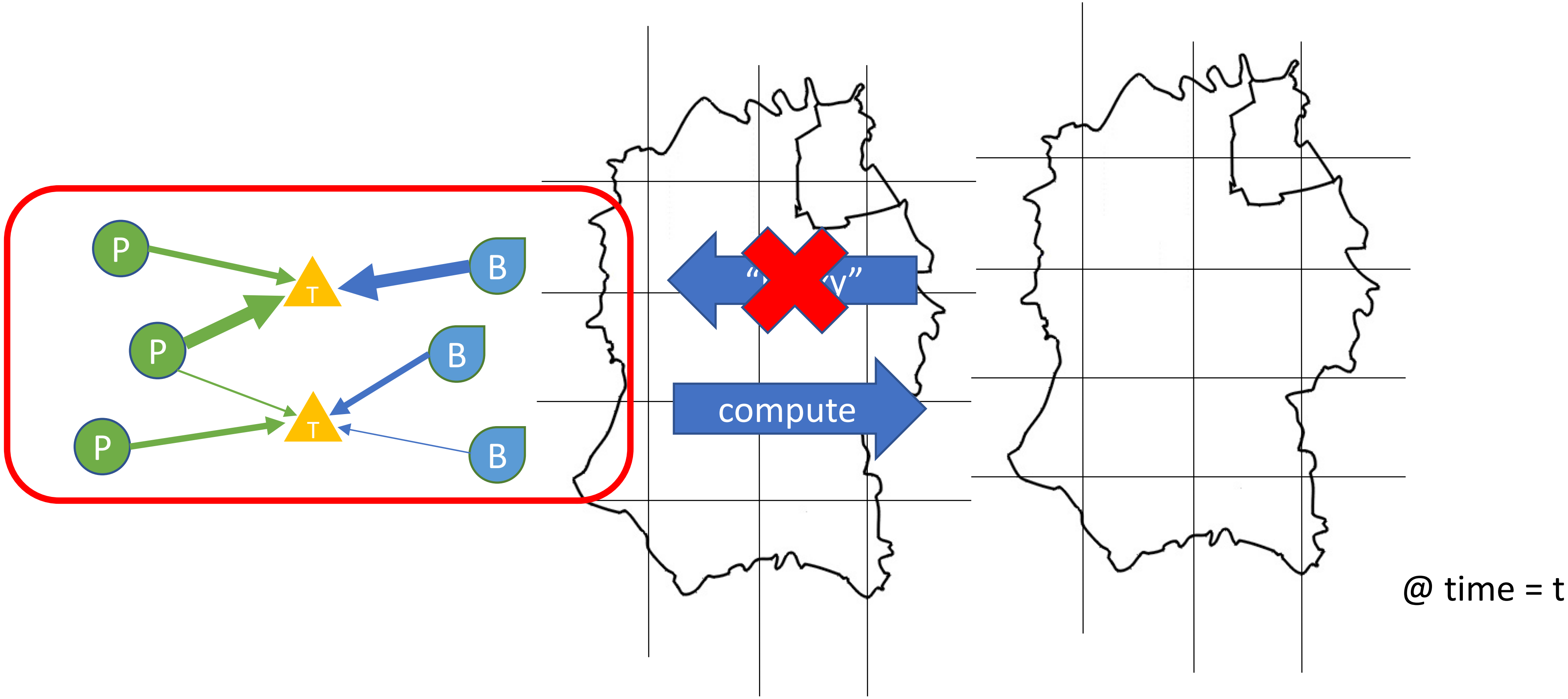
Image credits: INRA, AgriSemantics RDA working group

Reusing models

- Modeling paradigms represent different “metaphors” adopted during model design:
  - process-based vs. agent-based
  - stochastic/probabilistic vs.deterministic models
  - spatial vs. non-spatial, raster/vector, continuous vs. discrete time, etc.
- It remains **difficult to mix and match models incarnating different paradigms** across the lifecycle of an application.
- Often, complex problems are handled with one paradigm that fits some components but must be “tricked” to handle the rest.
- As a result models are still brittle **monoliths**, hard to disassemble and reassemble.
- Integrating architectures (OpenMI &C.) only handle the technical aspects of integration, addressing only a subset of the problem.

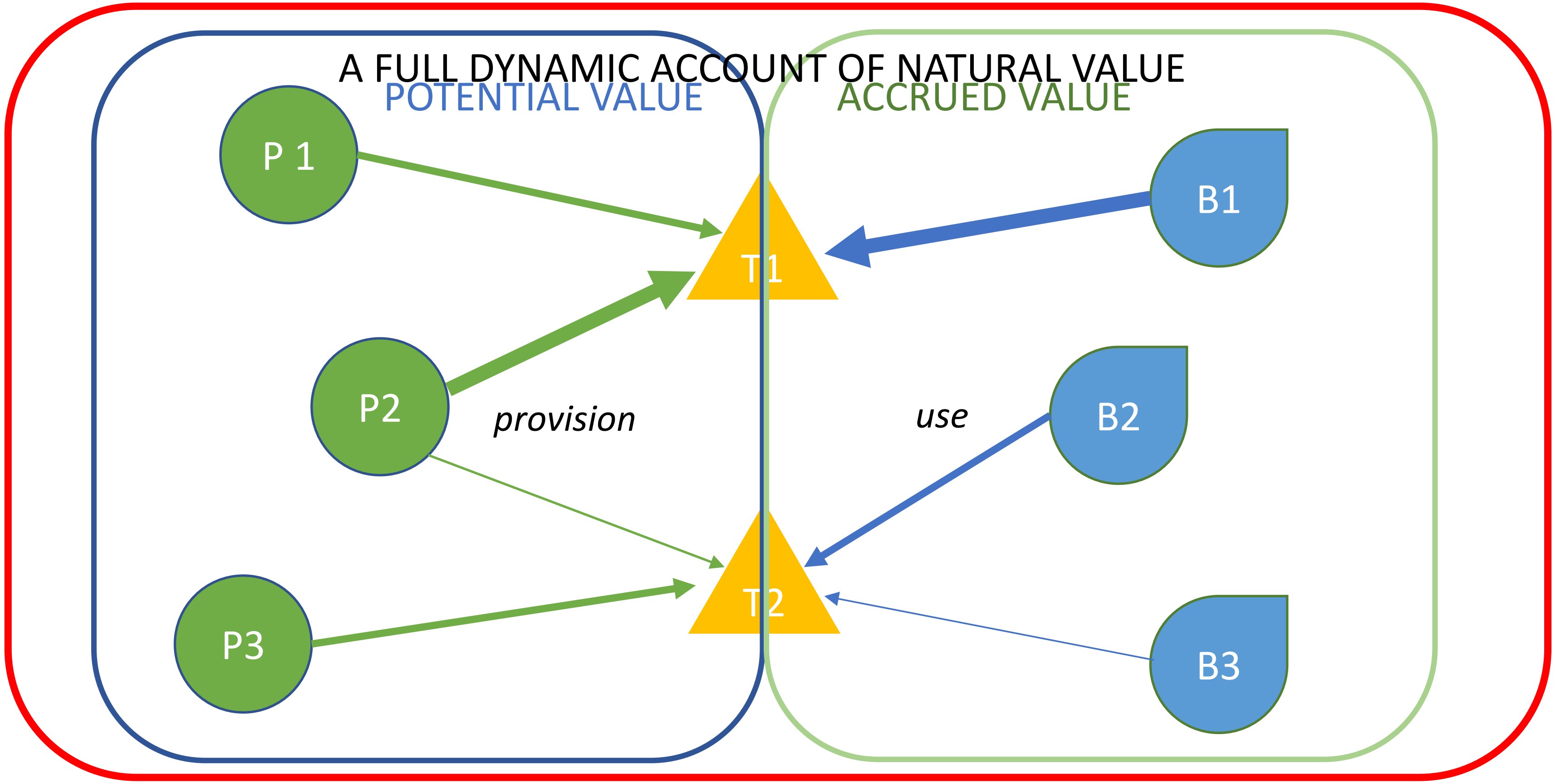
A case in point: accounting for human-natural interactions

- We know the limitations of “proxy” models – and it’s not because of decision makers.
- Still, building models of the *true* system models is hard - impossible in rapid assessments



Adaptive, assisted system characterization

Driven by semantics and by *roles*, supporting a specific view of physical phenomena without introducing ambiguities



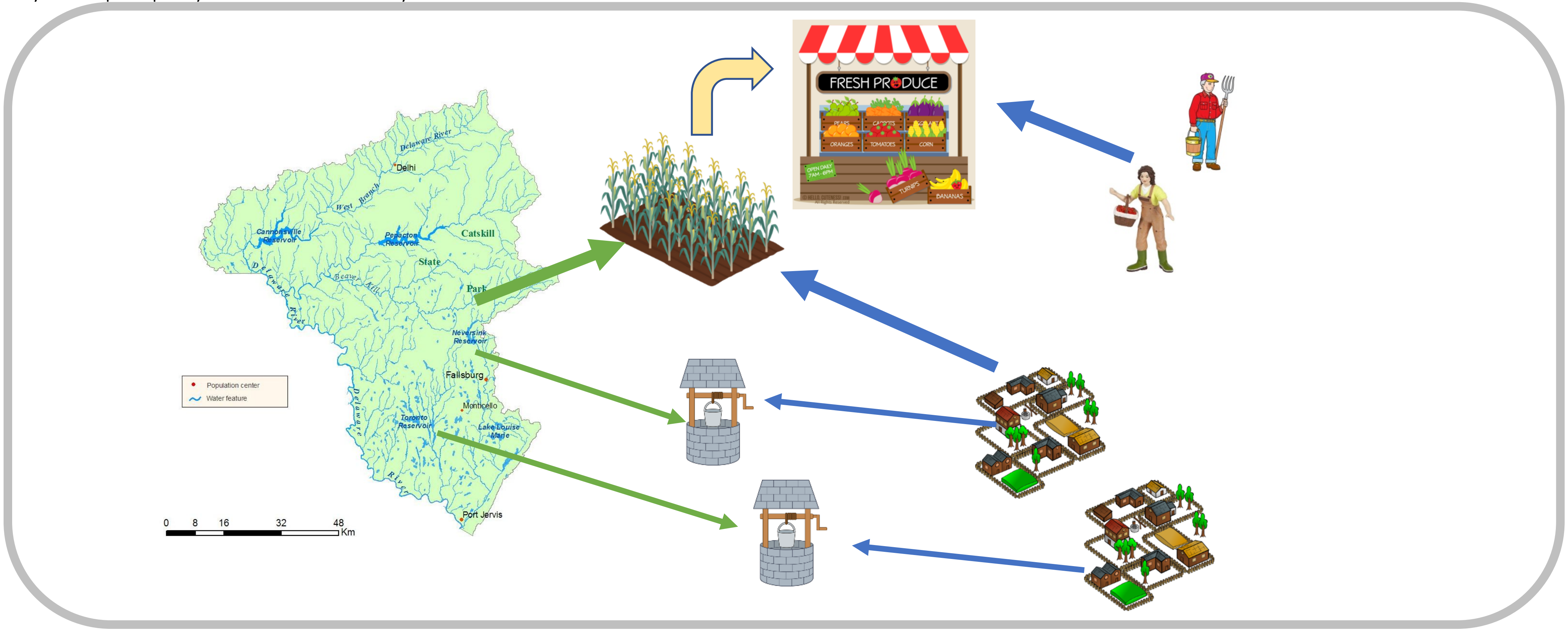
Providers (e.g. forests, watersheds): where valuable ecosystem function happens

Transactors (e.g. wells, crops, atmosphere): where natural value is generated

Beneficiaries (e.g. farmers, coastal dwellers): demand agents for natural value

Example: building an eco-social flow network

Triggered by a simple query: “observe social dynamics of water in watershed X”



The model for the *system* then creates and classifies all relationships, starting with *provision* (provider->transactor)…

…and following with use (beneficiary <- transactor), building a (potentially) differently scaled model for each flow.

Intermediate transactors (e.g. markets) are brought in according to the ontologies. They can be local or remote.

Providers (e.g. forests, watersheds) are first identified and built by the AI engine.

The ontologies define types of Transactors (e.g. wells, crops, atmosphere), identified last.

Beneficiaries (e.g. farmers, coastal dwellers) are identified next.

Resolution of models based on semantics

Model statements are stored in a distributed database. Each dependency is stated conceptually and resolved contextually.