# eScience and Geospatial Database

## GSB Workshop

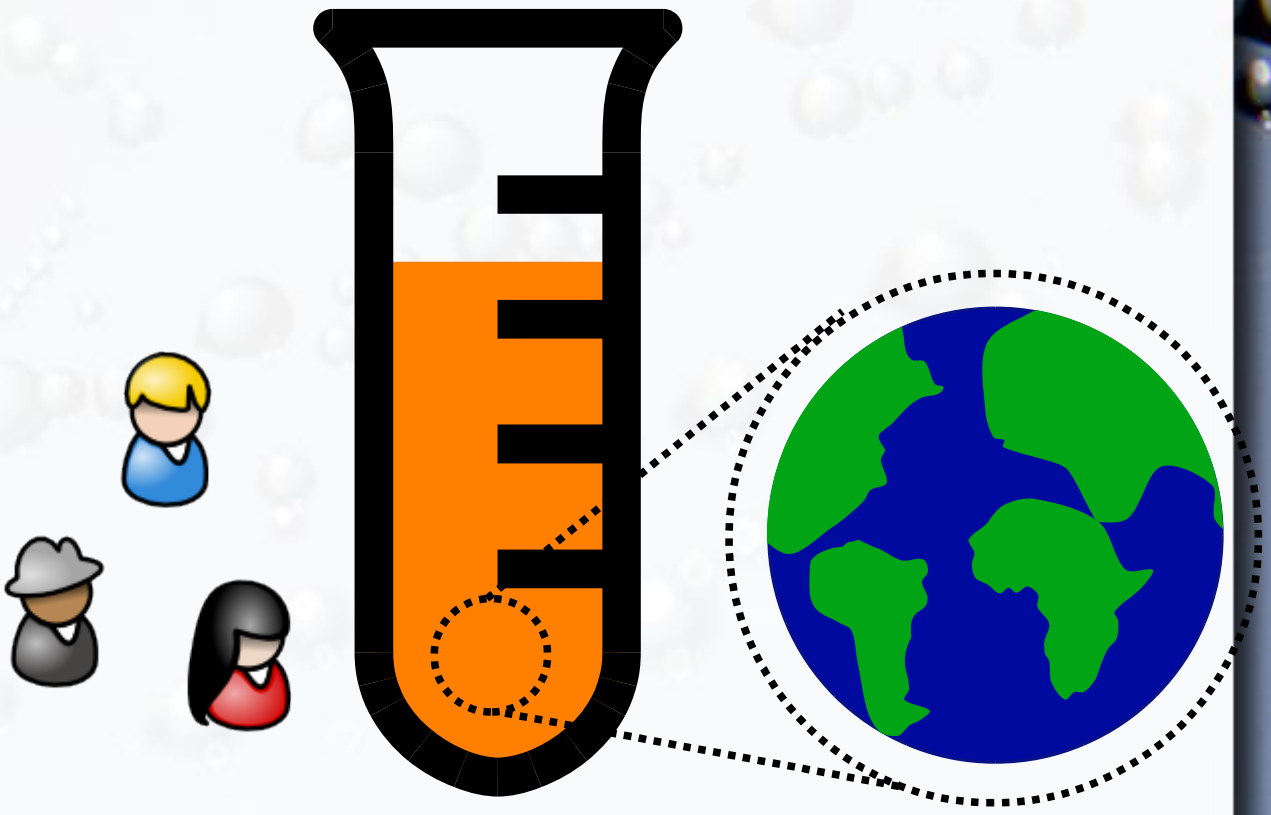
André Santanchè

LIS – Laboratoty of Information Systems
Institute of Computing – Unicamp

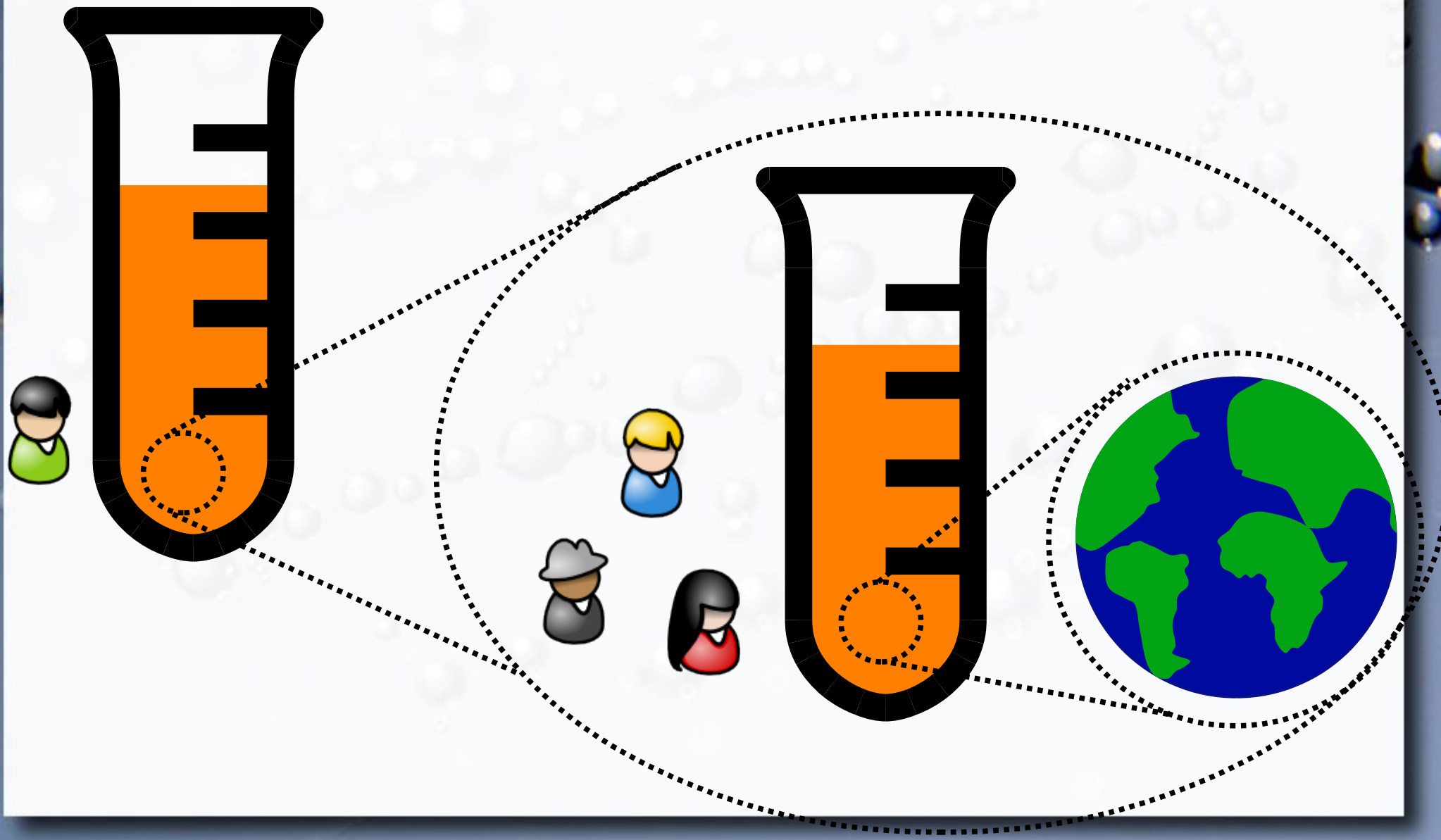
June 2013

Picture by Ben Collins [http://www.flickr.com/photos/graylight/]

# Science



# eScience



# eScience

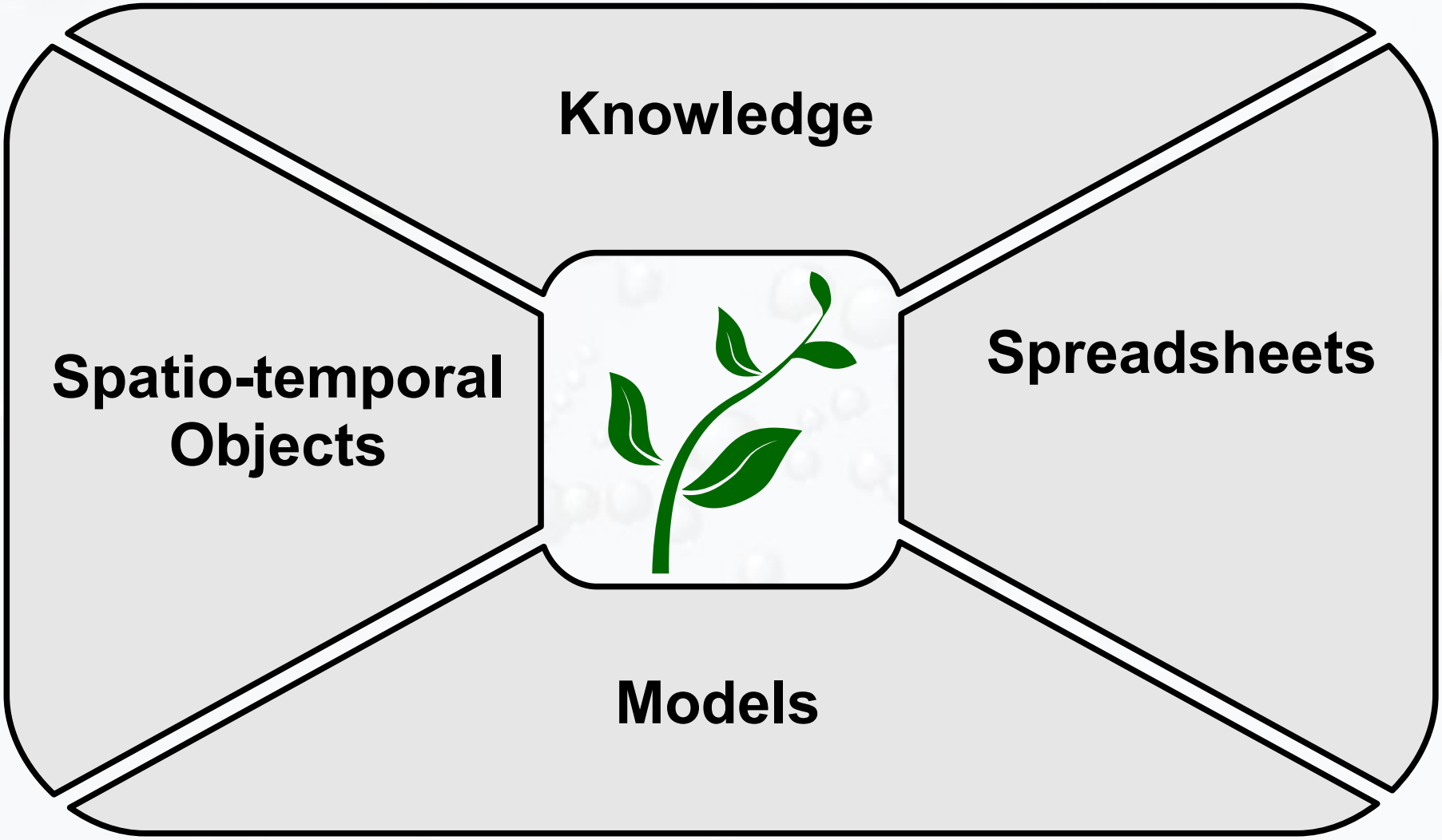
- The object of our science is your science when computers are involved
- “IT meets scientists” (Jim Gray on eScience, 2007)

# Fourth Paradigm

- **Thousand years ago:**
  - science was empirical; describing natural phenomena
- **Last few hundred years:**
  - theoretical branch; using models, generalizations
- **Last few decades:**
  - a computational branch; simulating complex phenomena
- **Today: data exploration (eScience)**
  - unify theory, experiment, and simulation

(Jim Gray on eScience, 2007)

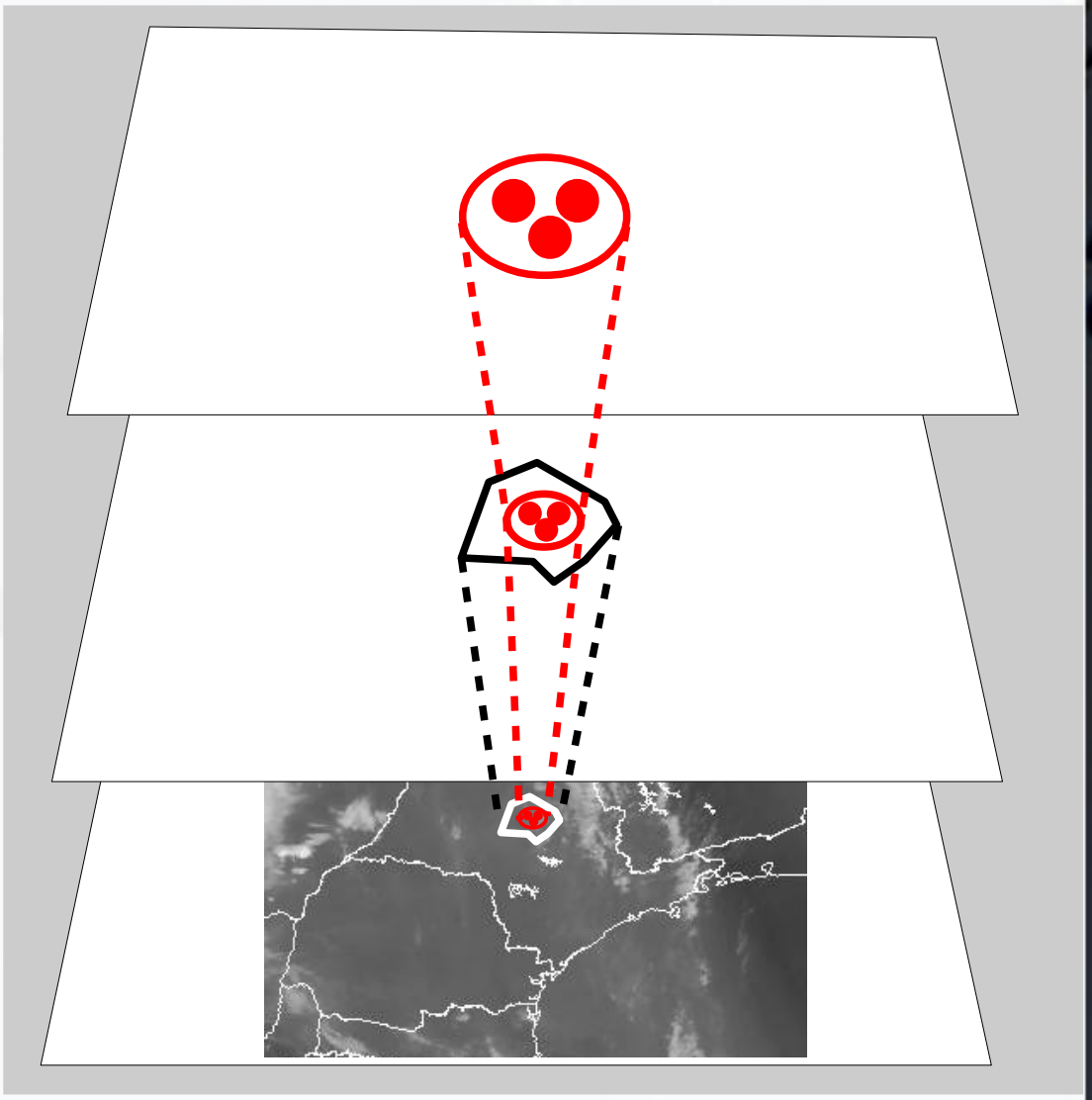
# Summary



# Spatio-temporal Objects

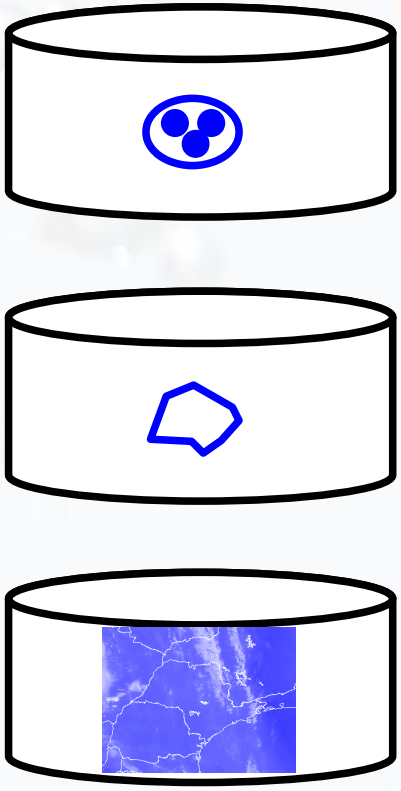
# Spatio-temporal Objects

- Multiscale
- Multi-representation
- Multi-focus

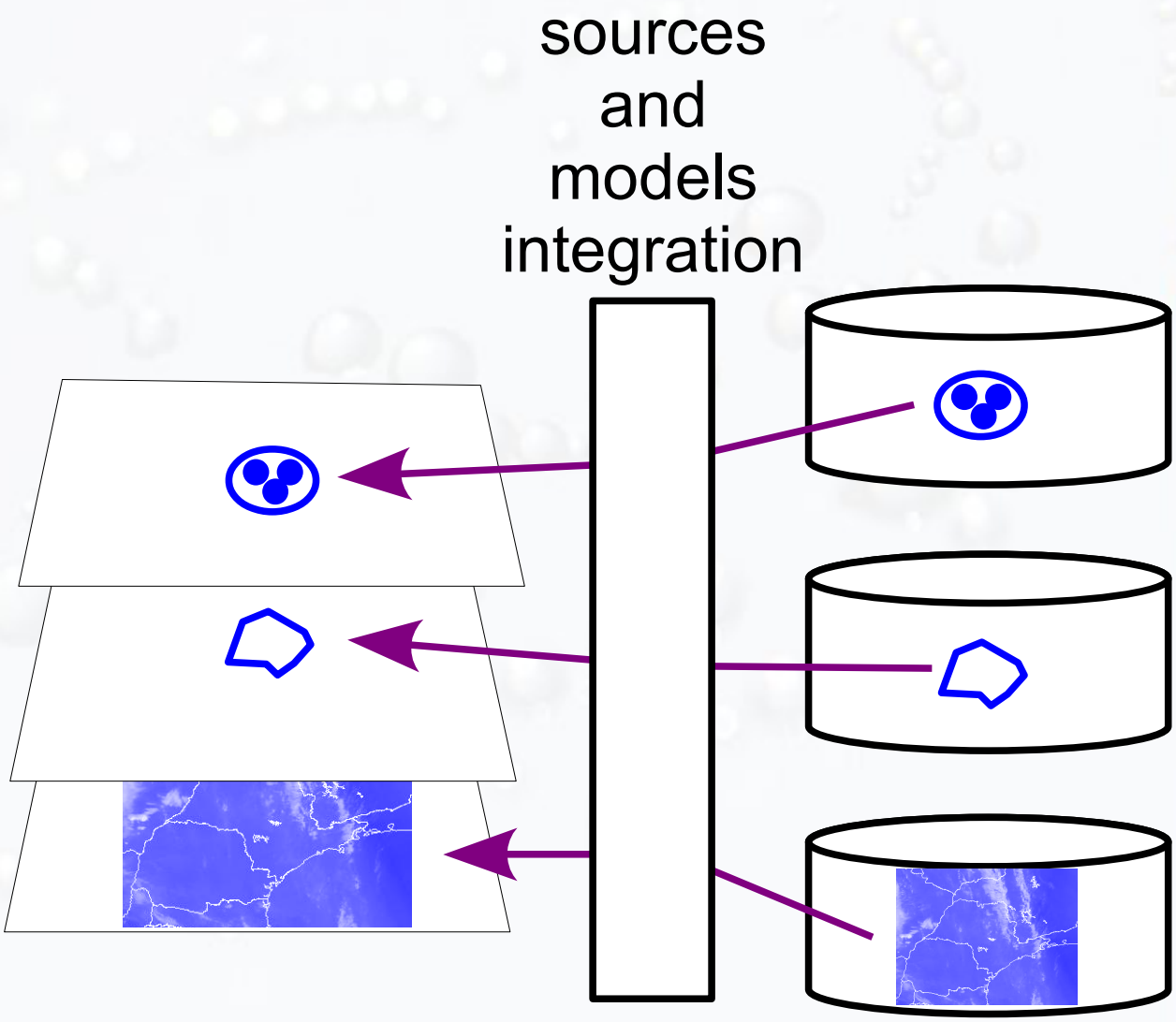


# Application-centered Integration

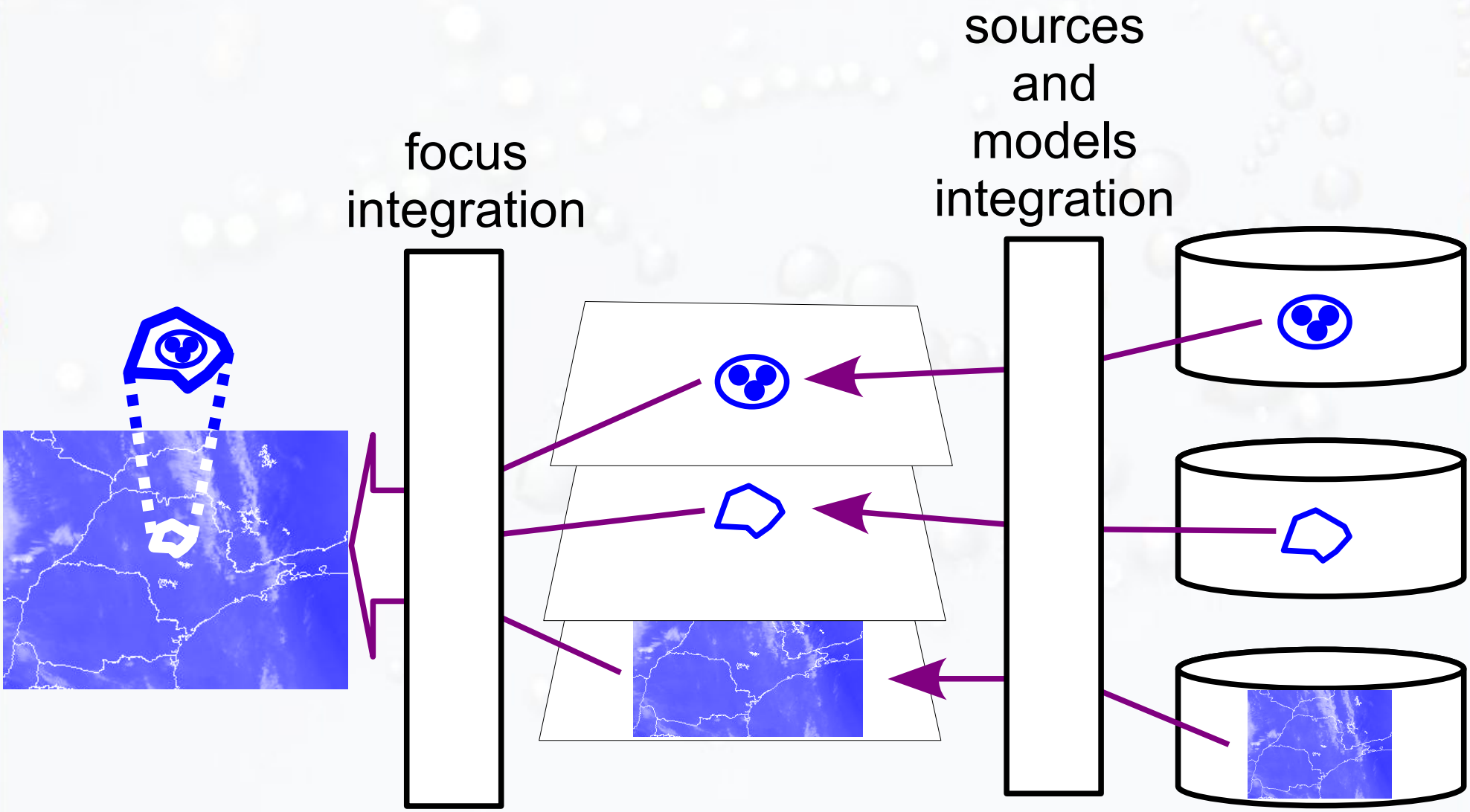
many sources
many models



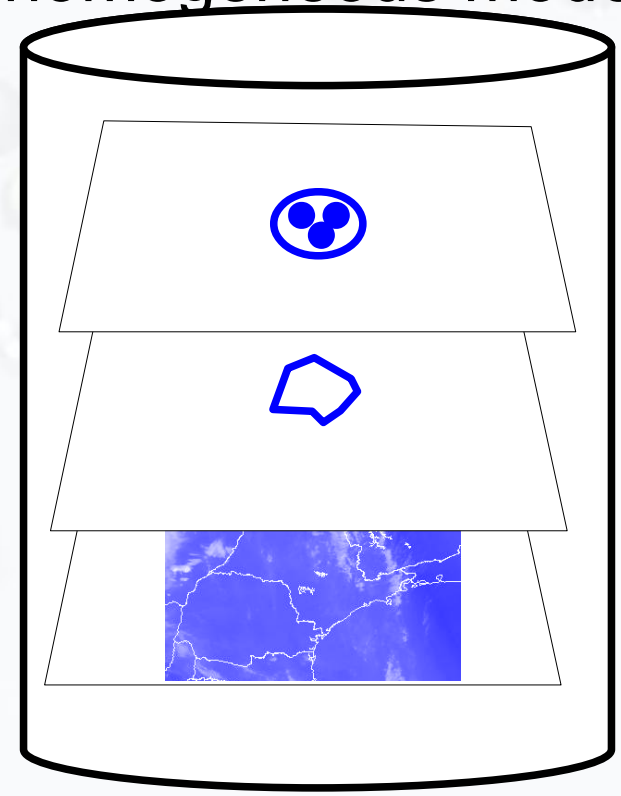
# Application-centered Integration



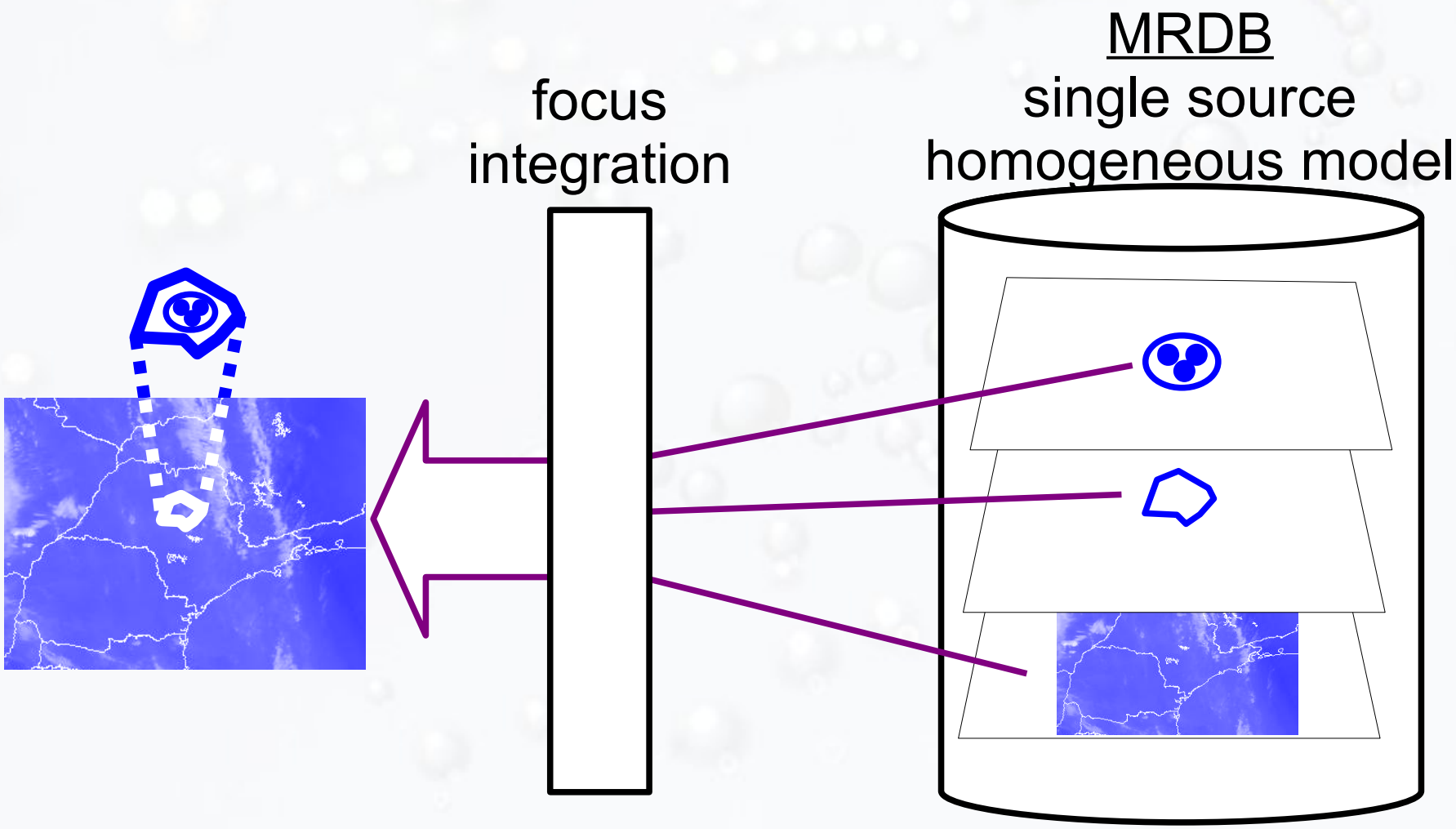
# Application-centered Integration



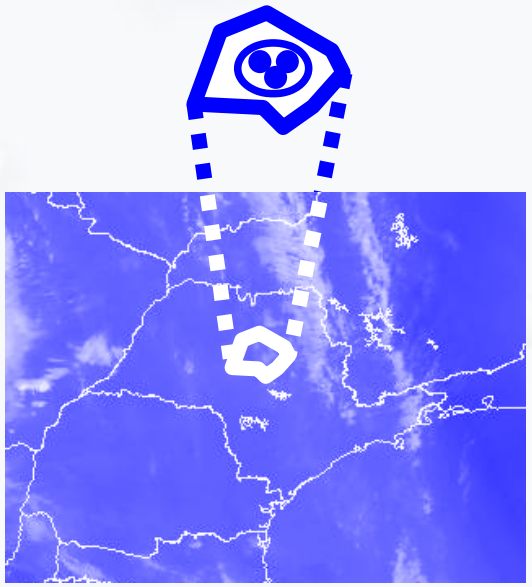
# Multi-representation Database

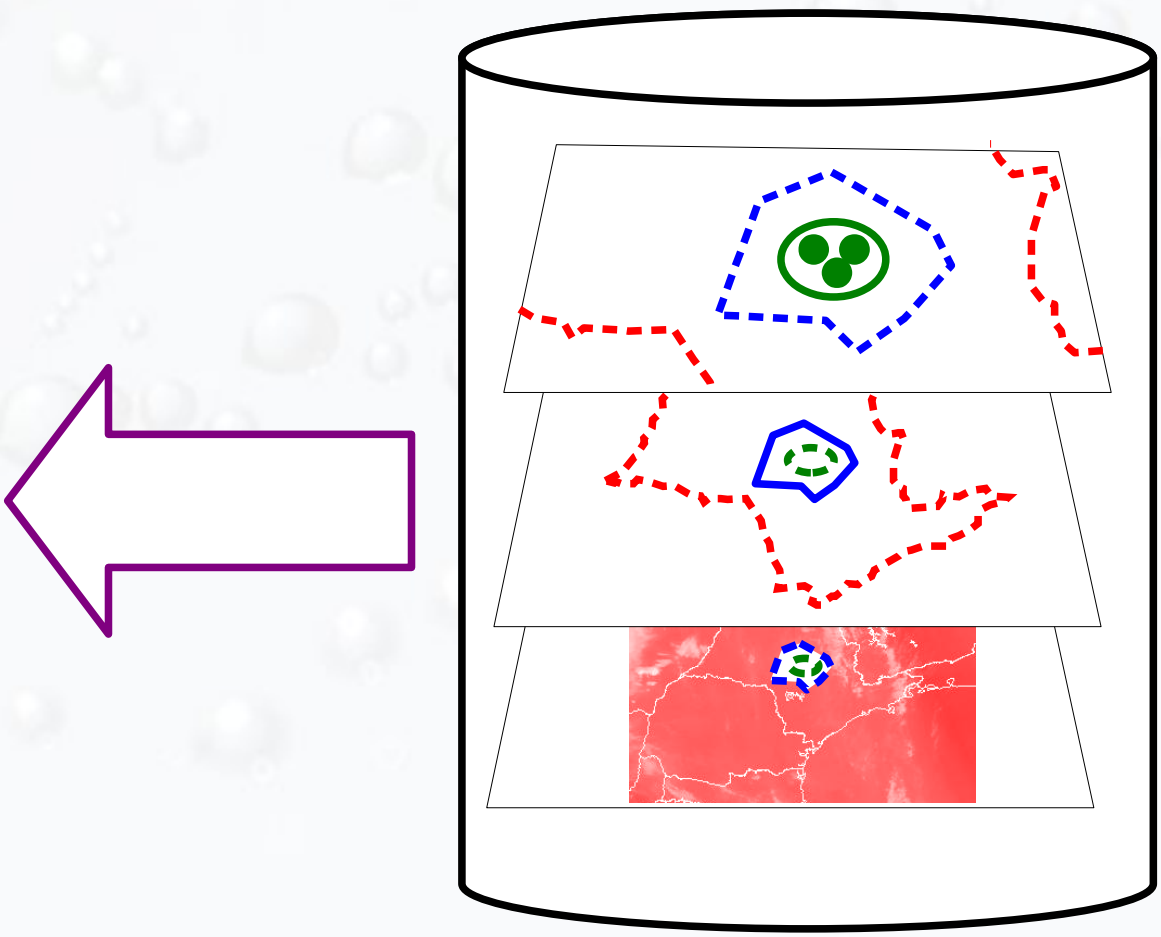


# Multi-representation Database



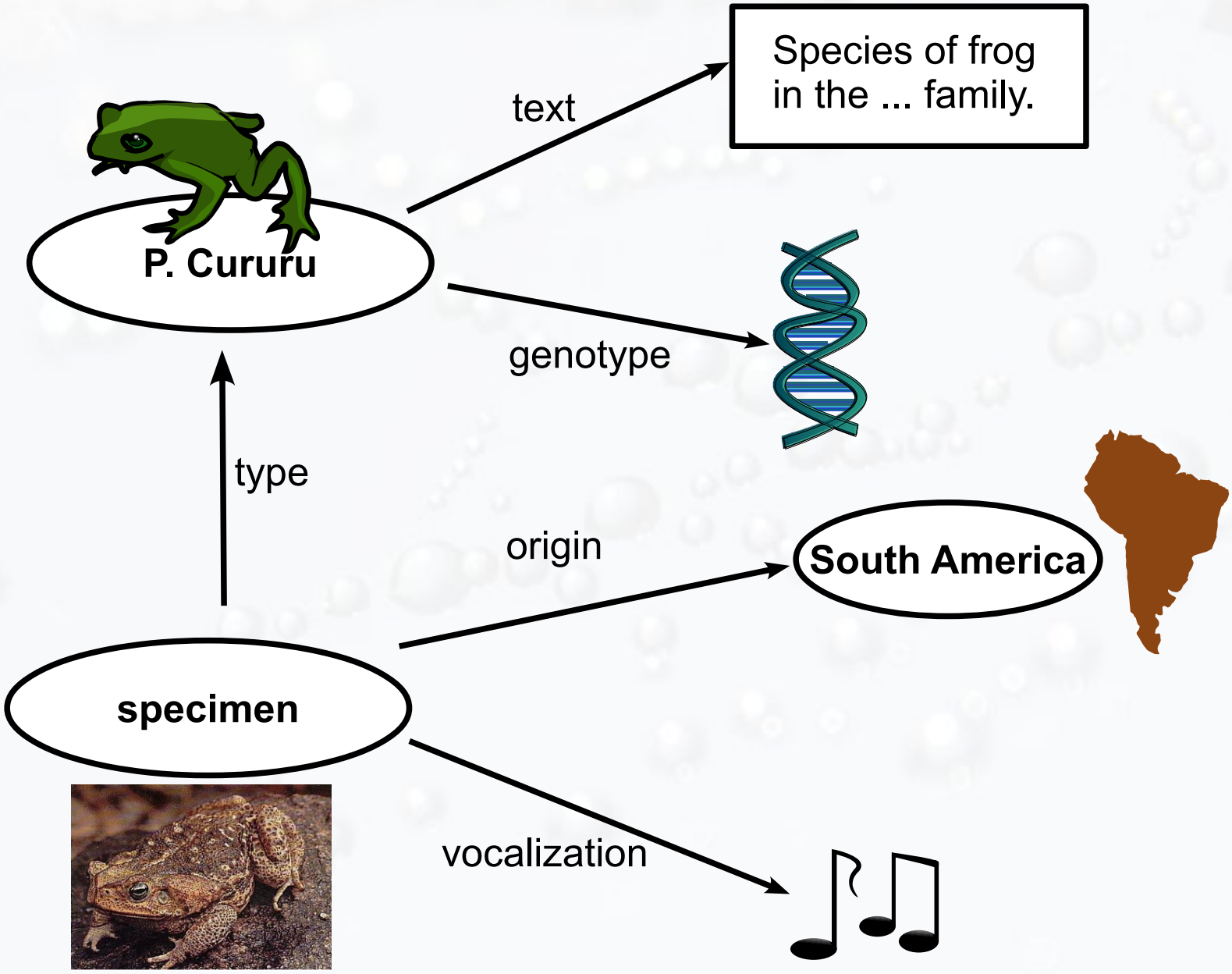
# Our Approach



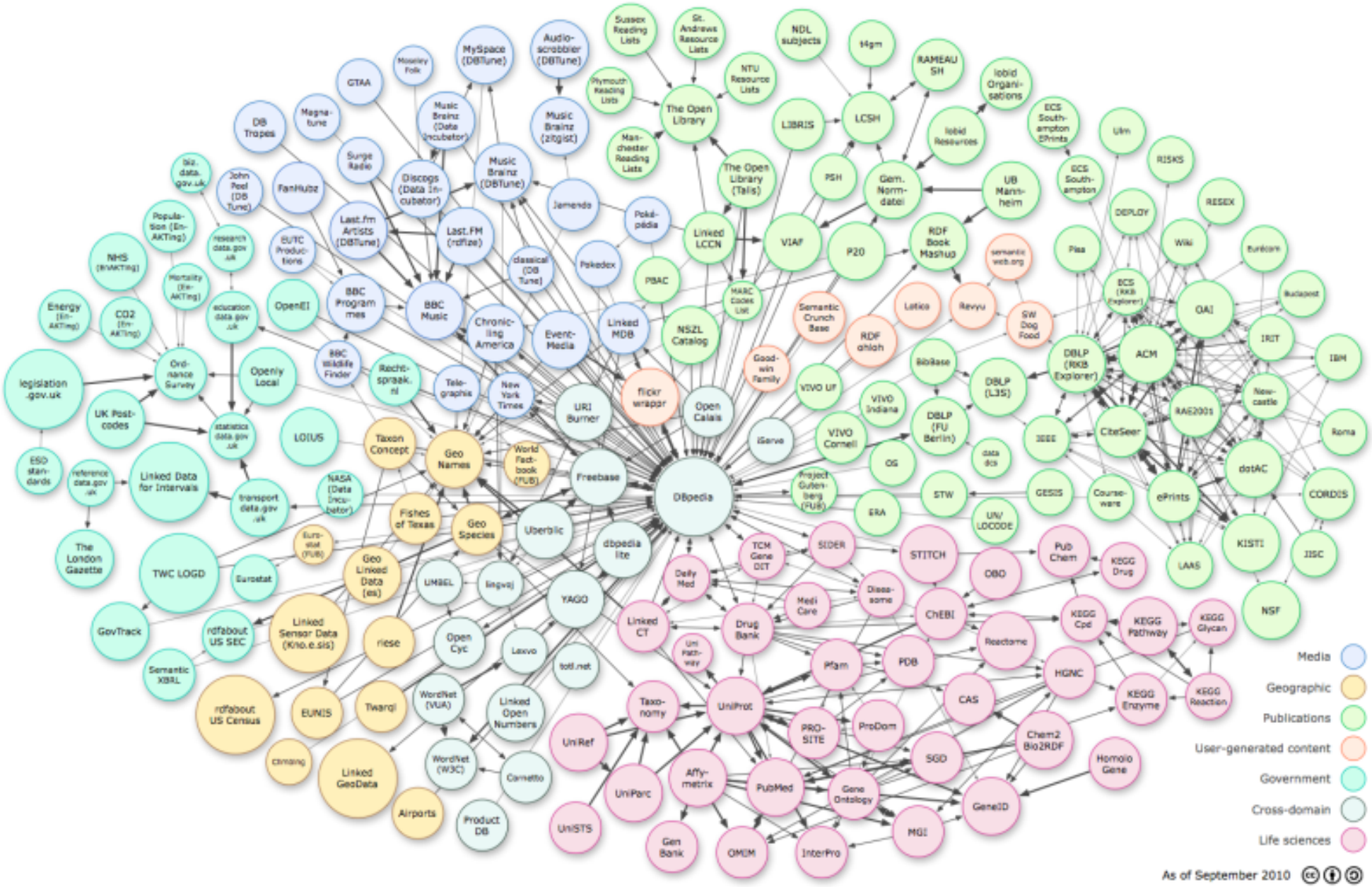


# Knowledge

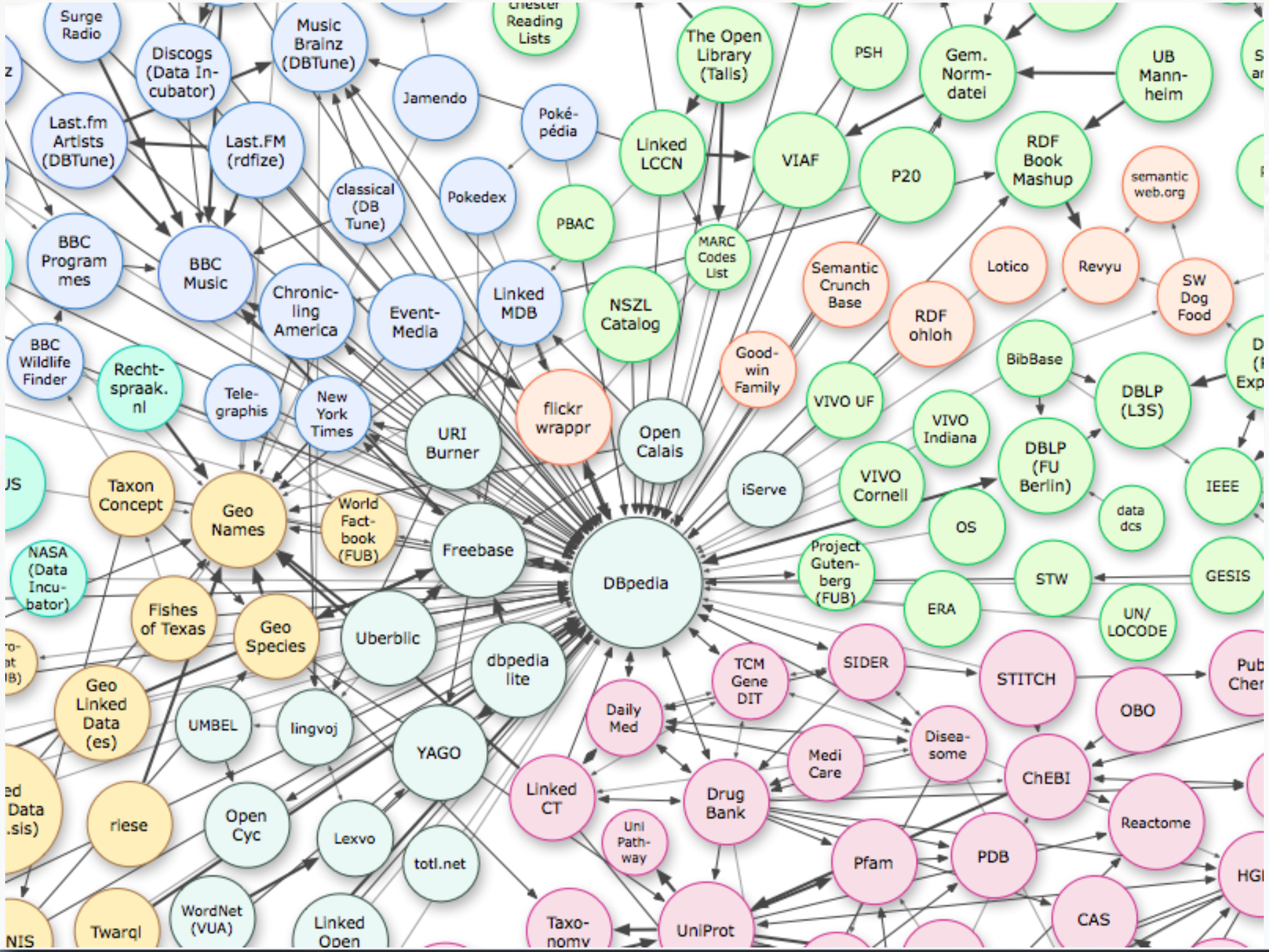
# Description as a Graph



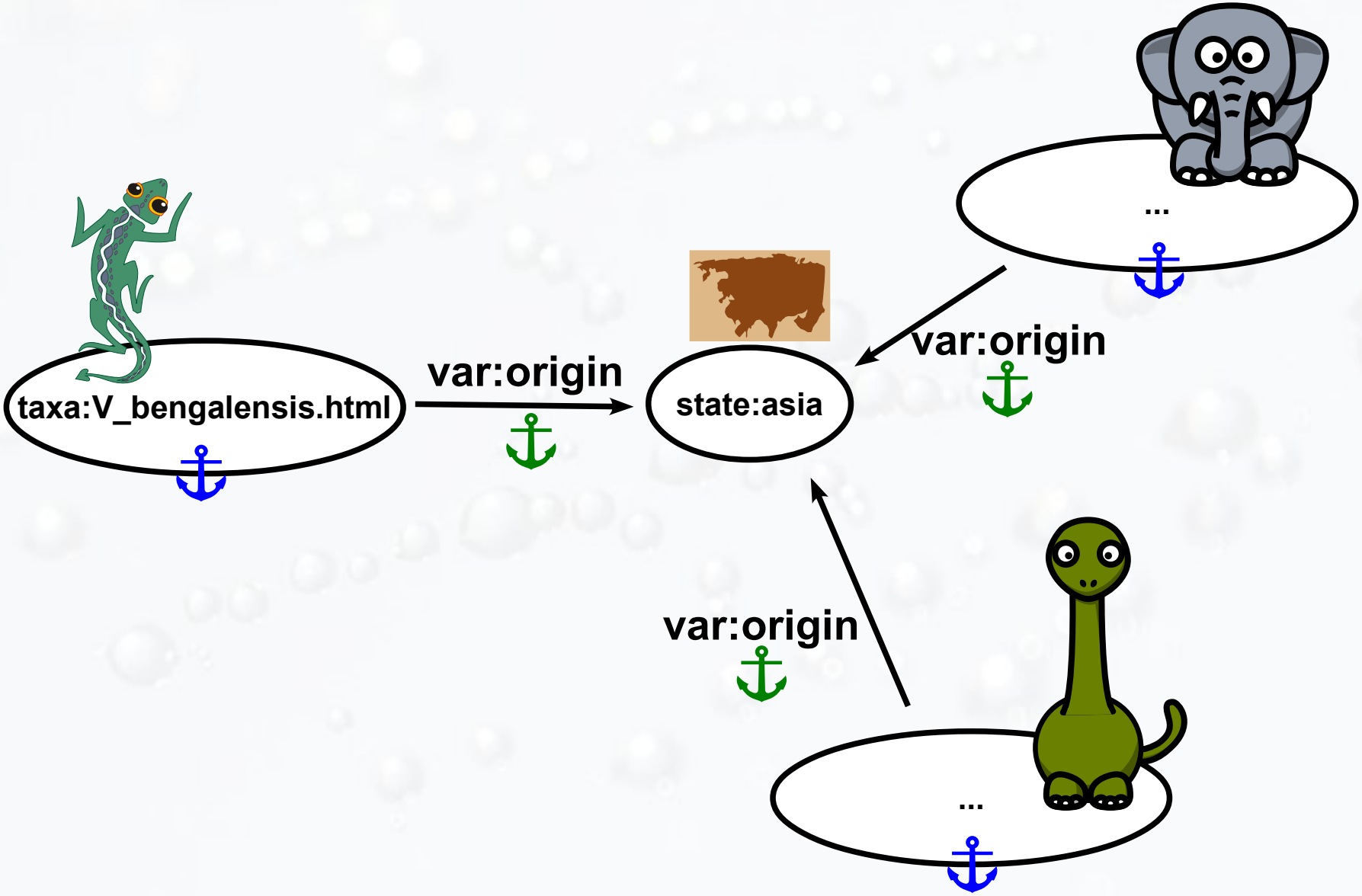
# Linked Data



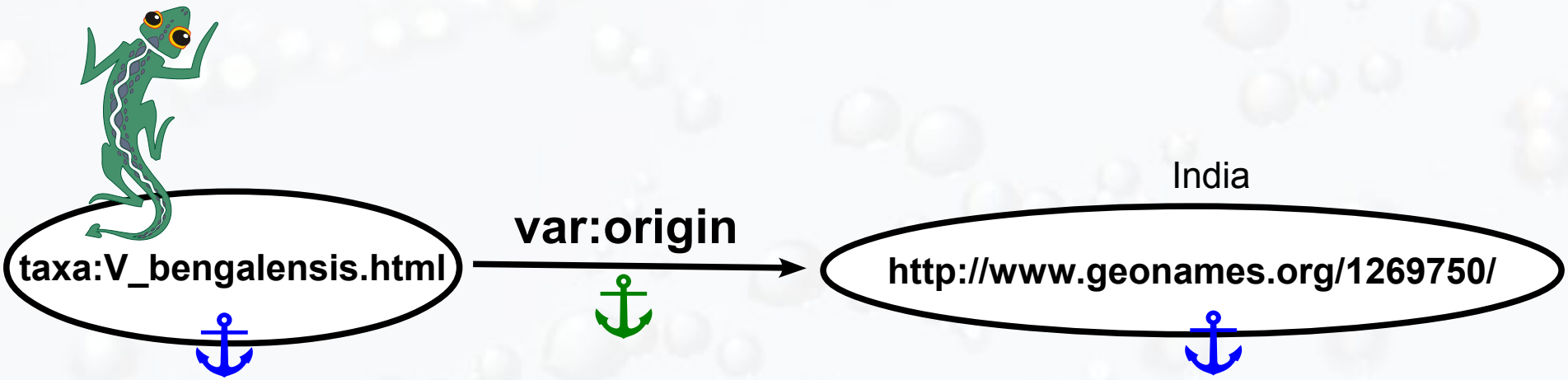
DBpedia
Freebase
Geo Names
Linked Data for Intervals
Ordnance Survey
Linked Sensor Data (Kno.e.sis)
Linked Open Numbers
Linked GeoData
Linked CT
Drug Bank
UniProt
Taxonomy
Affymetrix
PubMed
Gene Ontology
GeneID
MGI
InterPro
OMIM
Gen Bank
UniSTS
UniParc
UniRef
Pfam
PDB
CAS
SGD
Chem2 Bio2RDF
Homolo Gene
HGNC
KEGG Enzyme
KEGG Reaction
KEGG Pathway
KEGG Glycan
KEGG Cpd
KEGG Drug
Pub Chem
ChEBI
OBO
STITCH
SIDER
Medi Care
Disease-some
Daily Med
TCM Gene DIT
Uni Pathway
Reactome
Pro-Site
ProDom
Taxonomy
YAGO
Uberblic
dbpedia lite
Geo Species
Fishes of Texas
Geo Linked Data (es)
Eurostat
TWC LOGD
GovTrack
rdfabout US SEC
Semantic XBRL
rdfabout US Census
EUNIS
Twarql
WordNet (VUA)
WordNet (W3C)
Cornetto
Product DB
Airports
Linked GeoData
Linked Sensor Data (Kno.e.sis)
riese
Open Cyc
Lexvo
totl.net
Linked CT
Drug Bank
Medi Care
Daily Med
Uberblic
Geo Species
Fishes of Texas
Taxon Concept
LOIUS
Geo Names
World Fact-book (FUB)
Freebase
URI Burner
New York Times
Telegraphis
Rechtspraak.nl
BBC Wildlife Finder
Openly Local
Ordnance Survey
legislation.gov.uk
UK Post-codes
statistics data.gov.uk
transport data.gov.uk
reference data.gov.uk
The London Gazette
ESD standards
Energy (En-AKTing)
CO2 (En-AKTing)
NHS (BNAKTing)
Mortality (En-AKTing)
Population (En-AKTing)
biz.data.gov.uk
John Peel (DB Tune)
FanHubz
DB Tropes
Magnatune
GTAA
Moseley Folk
MySpace (DBTune)
Audioscrobbler (DBTune)
Music Brainz (Data Incubator)
Music Brainz (DBTune)
Music Brainz (zitgist)
Surge Radio
Discogs (Data Incubator)
Last.fm Artists (DBTune)
Last.FM (rdfize)
EUTC Productions
BBC Programmes
BBC Music
OpenEI
Jamendo
Pokédex
Poképédia
classical (DB Tune)
Chronicling America
Event-Media
Linked MDB
NSZL Catalog
PBAC
Linked LCCN
MARC Codes List
Open Calais
flickr wrappr
Sussex Reading Lists
St. Andrews Resource Lists
Plymouth Reading Lists
Manchester Reading Lists
The Open Library
The Open Library (Talis)
NTU Resource Lists
NDL subjects
t4gm
RAMEAU SH
LIBRIS
LCSH
Iobid Organisations
Iobid Resources
PSH
Gem. Norm-datei
UB Mannheim
VIAF
P20
RDF Book Mashup
semantic web.org
Semantic Crunch Base
RDF ohloh
Lotico
Revyu
SW Dog Food
Good-win Family
VIVO UF
VIVO Indiana
VIVO Cornell
iServe
BibBase
DBLP (L3S)
DBLP (FU Berlin)
DBLP (RKB Explorer)
ACM
OAI
IRIT
IBM
New-castle
RAE2001
CiteSeer
IEEE
Roma
dotAC
CORDIS
ePrints
Course-ware
GESIS
STW
ERA
UN/LOCODE
Project Gutenberg (FUB)
data dcs
KISTI
JISC
LAAS
NSF
Pisa
Wiki
Eurécom
Budapest
DEPLOY
RESEX
RISKS
Ulm
ECS South-ampton EPrints
ECS South-ampton
ECS (RKB Explorer)
Media
Geographic
Publications
User-generated content
Government
Cross-domain
Life sciences
As of September 2010


Surge Radio
Discogs (Data Incubator)
Music Brainz (DBTune)
Jamendo
Poké-pédia
The Open Library (Talis)
PSH
Gem. Norm-datei
UB Mann-heim
Last.fm Artists (DBTune)
Last.FM (rdfize)
Linked LCCN
VIAF
P20
RDF Book Mashup
semantic web.org
classical (DB Tune)
Pokedex
PBAC
MARC Codes List
BBC Programmes
BBC Music
Chronicling America
Event-Media
Linked MDB
NSZL Catalog
Semantic Crunch Base
Lotico
Revyu
SW Dog Food
RDF ohloh
BBC Wildlife Finder
Recht-spraak.nl
Tele-graphis
New York Times
URI Burner
flickr wrappr
Open Calais
Good-win Family
VIVO UF
BibBase
DBLP (L3S)
VIVO Indiana
DBLP (FU Berlin)
IEEE
Taxon Concept
Geo Names
World Fact-book (FUB)
Freebase
DBpedia
iServe
VIVO Cornell
OS
data dcs
NASA (Data Incubator)
Project Guten-berg (FUB)
STW
GESIS
Fishes of Texas
Geo Species
Uberblic
dbpedia lite
ERA
UN/ LOCODE
Geo Linked Data (es)
UMBEL
lingvoj
YAGO
TCM Gene DIT
SIDER
STITCH
OBO
Daily Med
Medi Care
Disea-some
ChEBI
Linked CT
Drug Bank
Reactome
riese
Open Cyc
Lexvo
totl.net
Uni Path-way
Pfam
PDB
Twarql
WordNet (VUA)
Linked Open
Taxo-nomy
UniProt
CAS

# Connected Graph

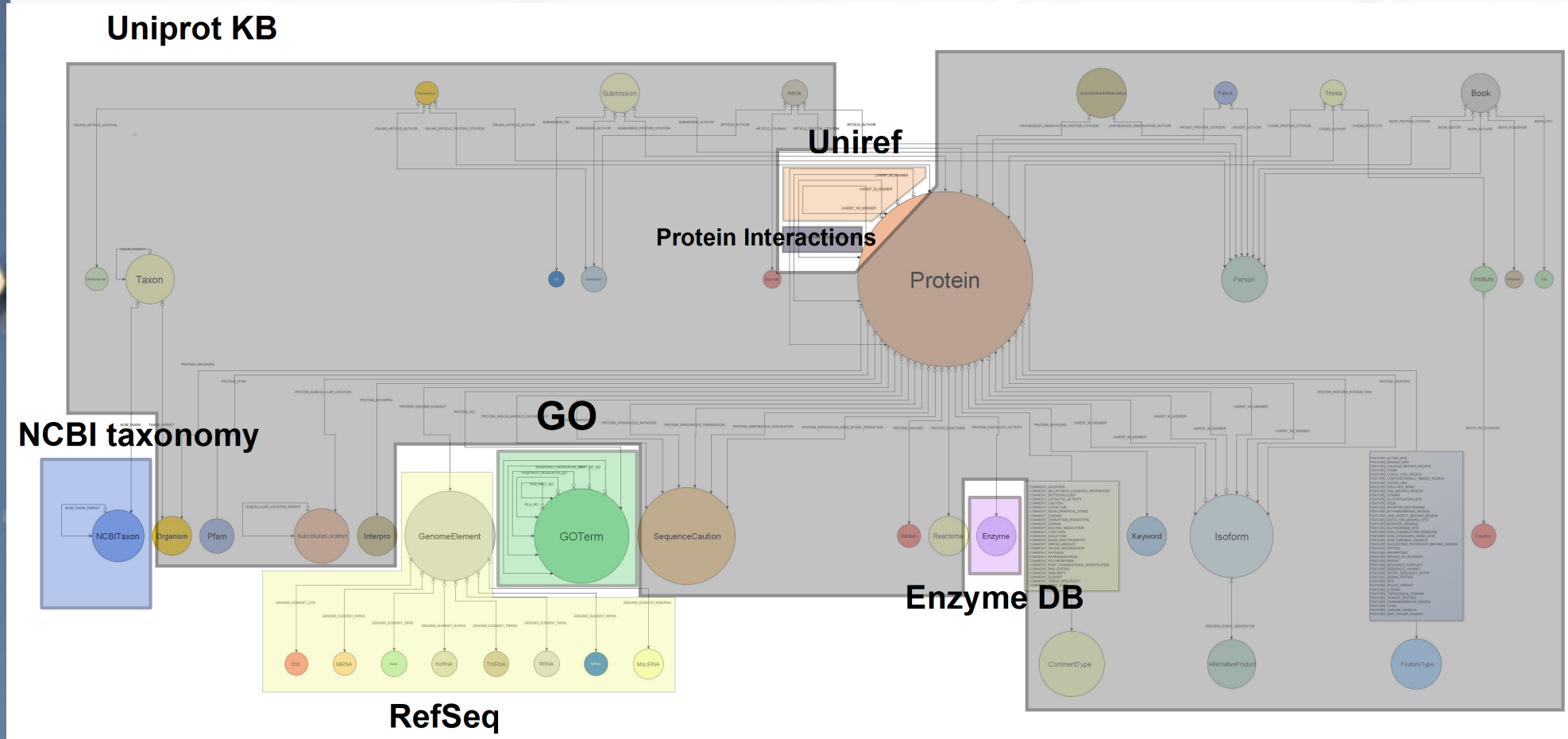


# Origin in GeoNames



# Graph Databases

# Graph-based Integration bio4j

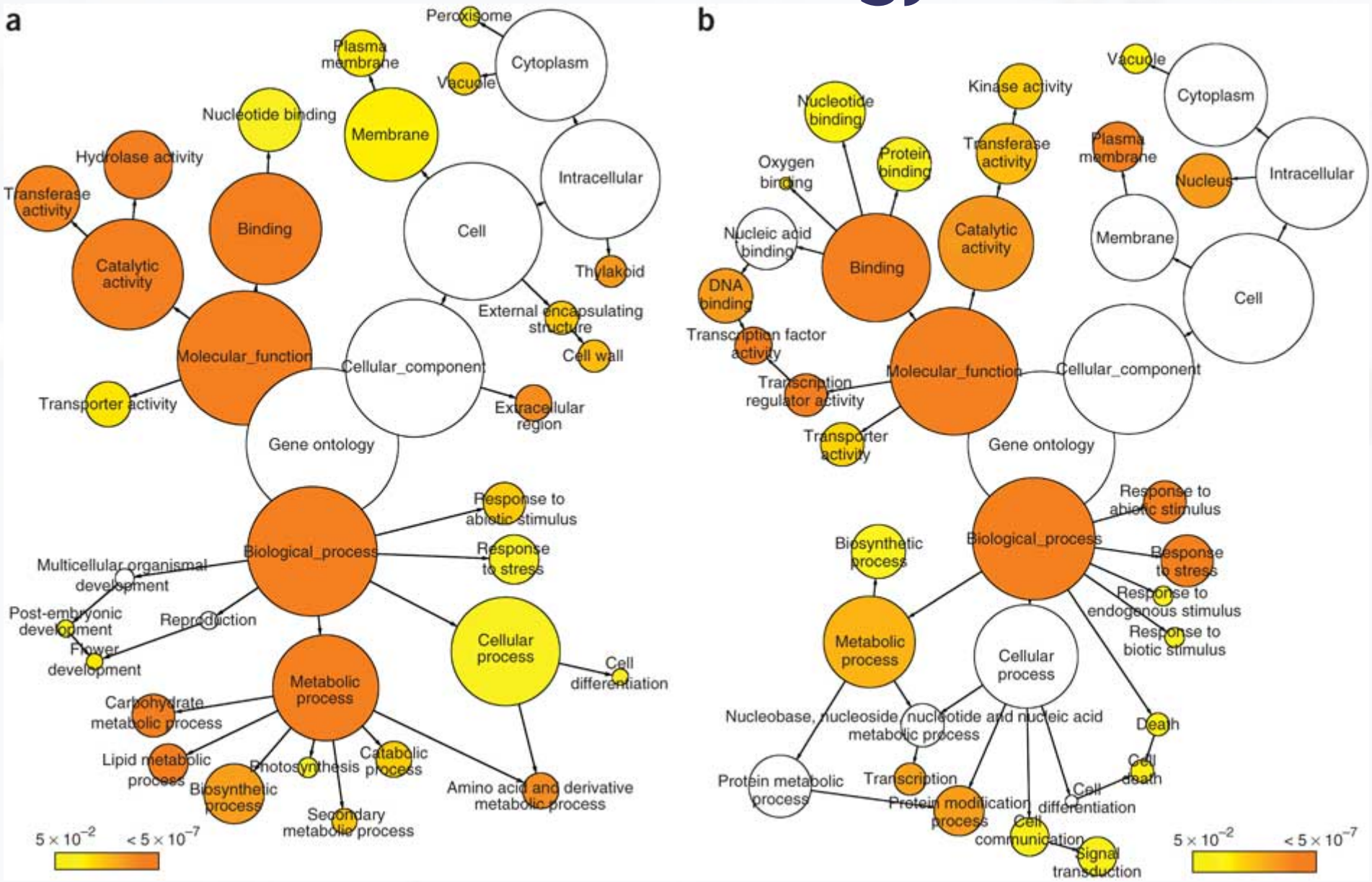



http://bio4j.com/

# From Graphs to Ontologies

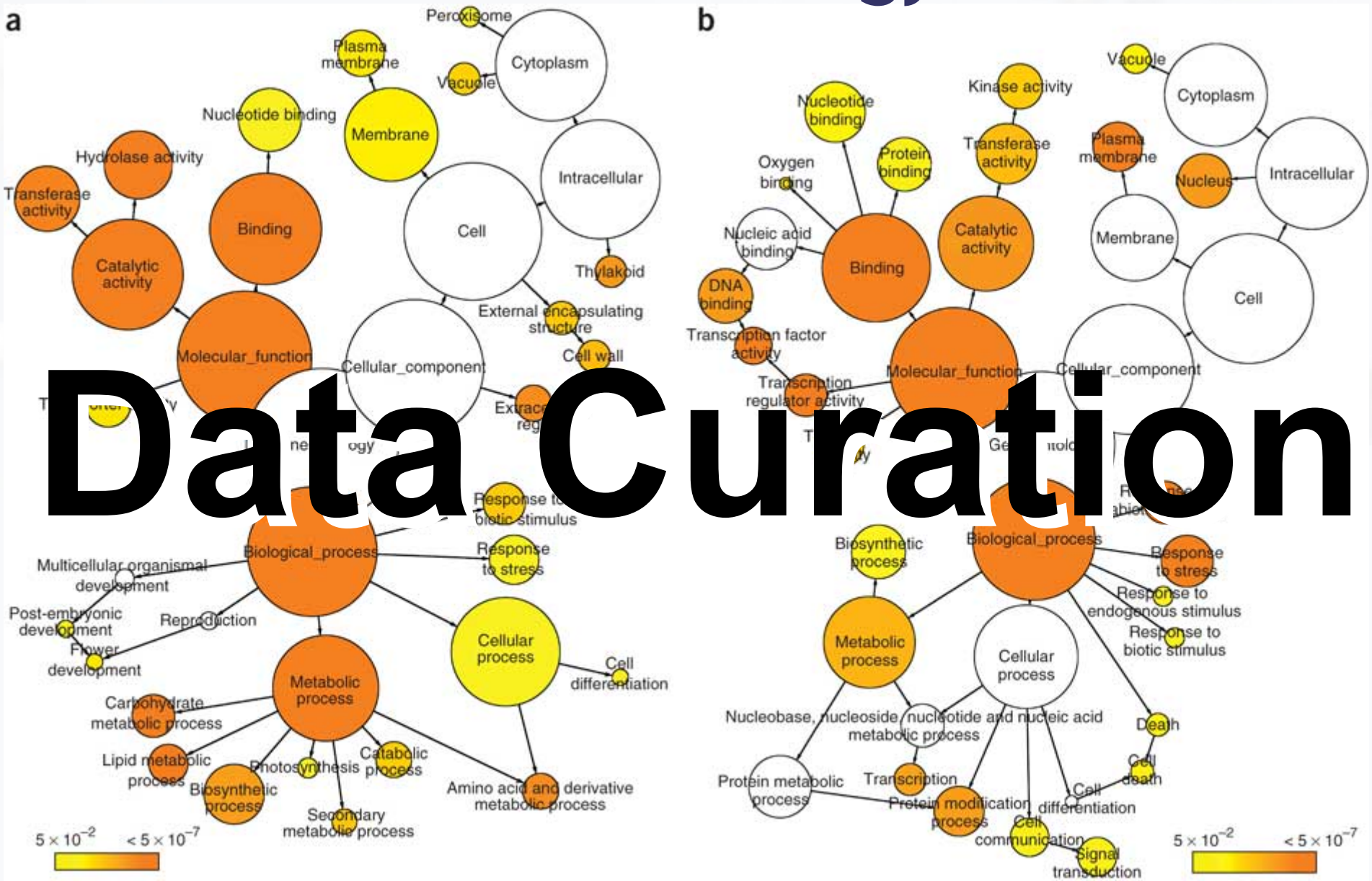
- Graph database
  - minimum common denominator
  - “freedom” of making relations beyond the schema
  - nodes/edges → raw data → knowledge implicit
- Ontology
  - graph → can be stored in a graph database
  - nodes/edges → formal knowledge

# Geneontology




The genome of woodland strawberry
Nature Genetics 43, 109–116 (2011) doi:10.1038/ng.740

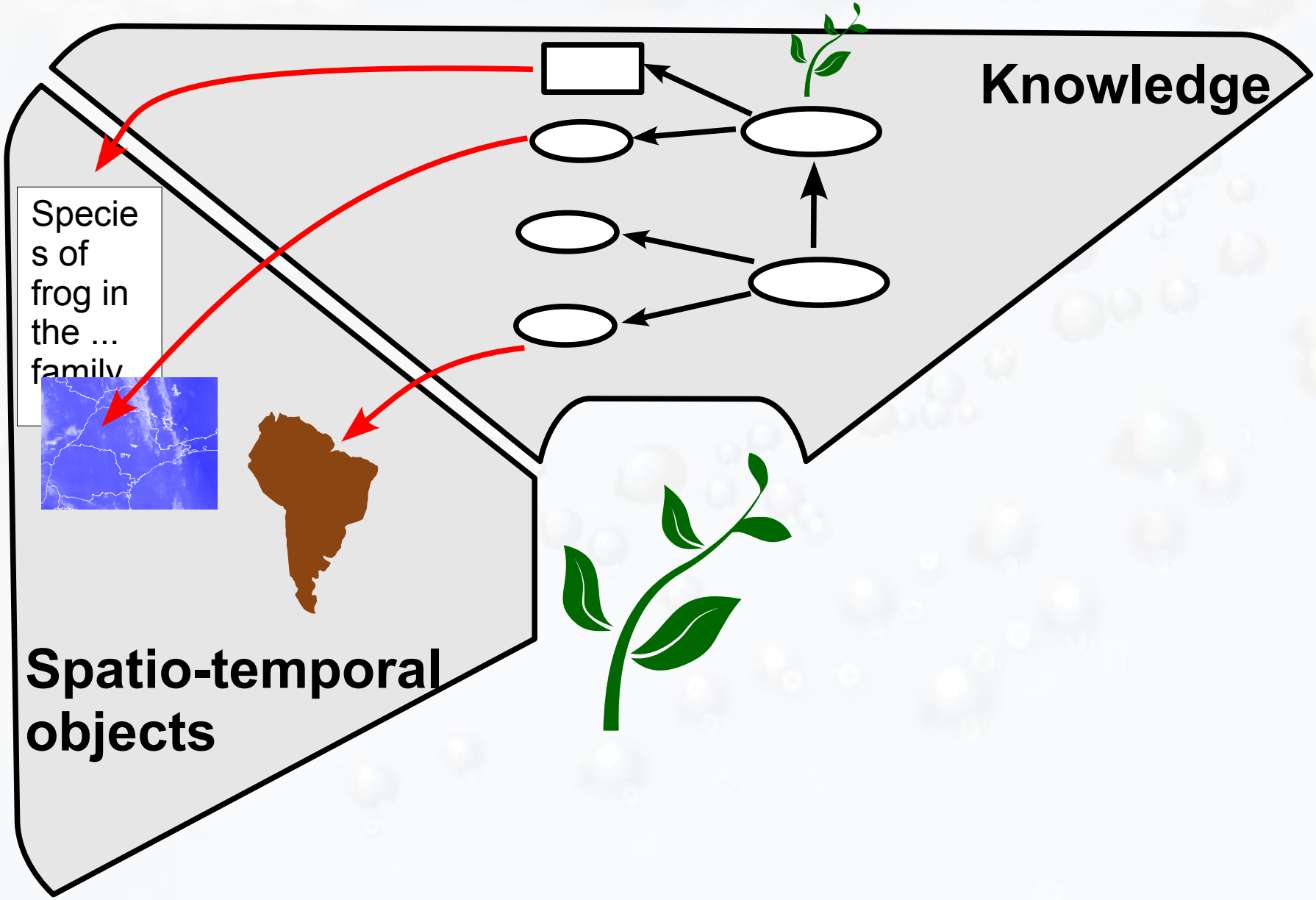
# Geneontology

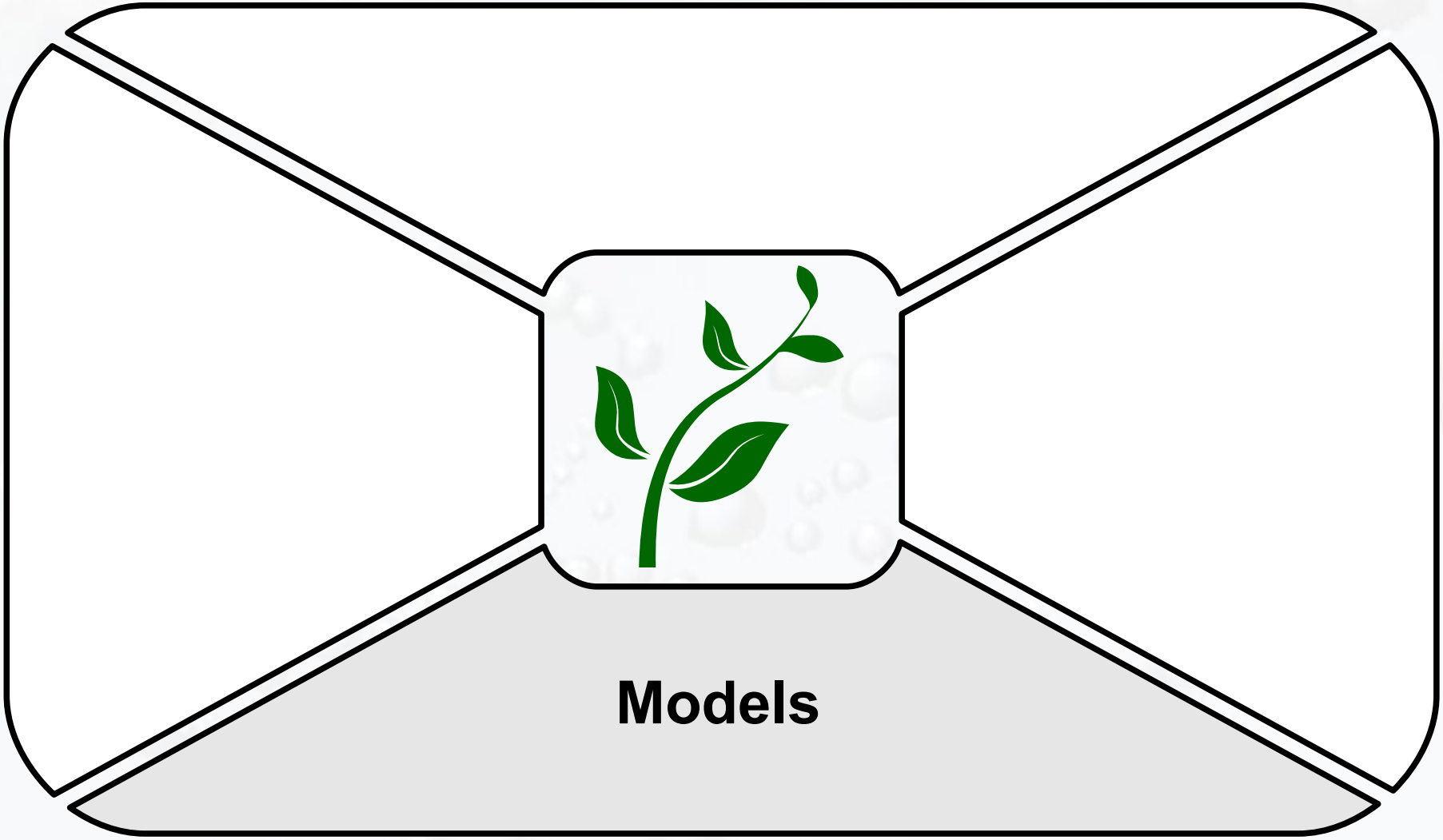



# Data Curation

The genome of woodland strawberry
Nature Genetics 43, 109–116 (2011) doi:10.1038/ng.740

# Bridging

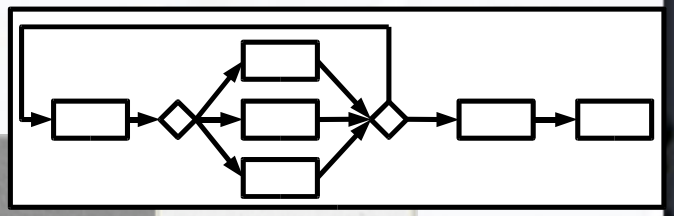

Knowledge
Specie
s of
frog in
the ...
family
Spatio-temporal
objects

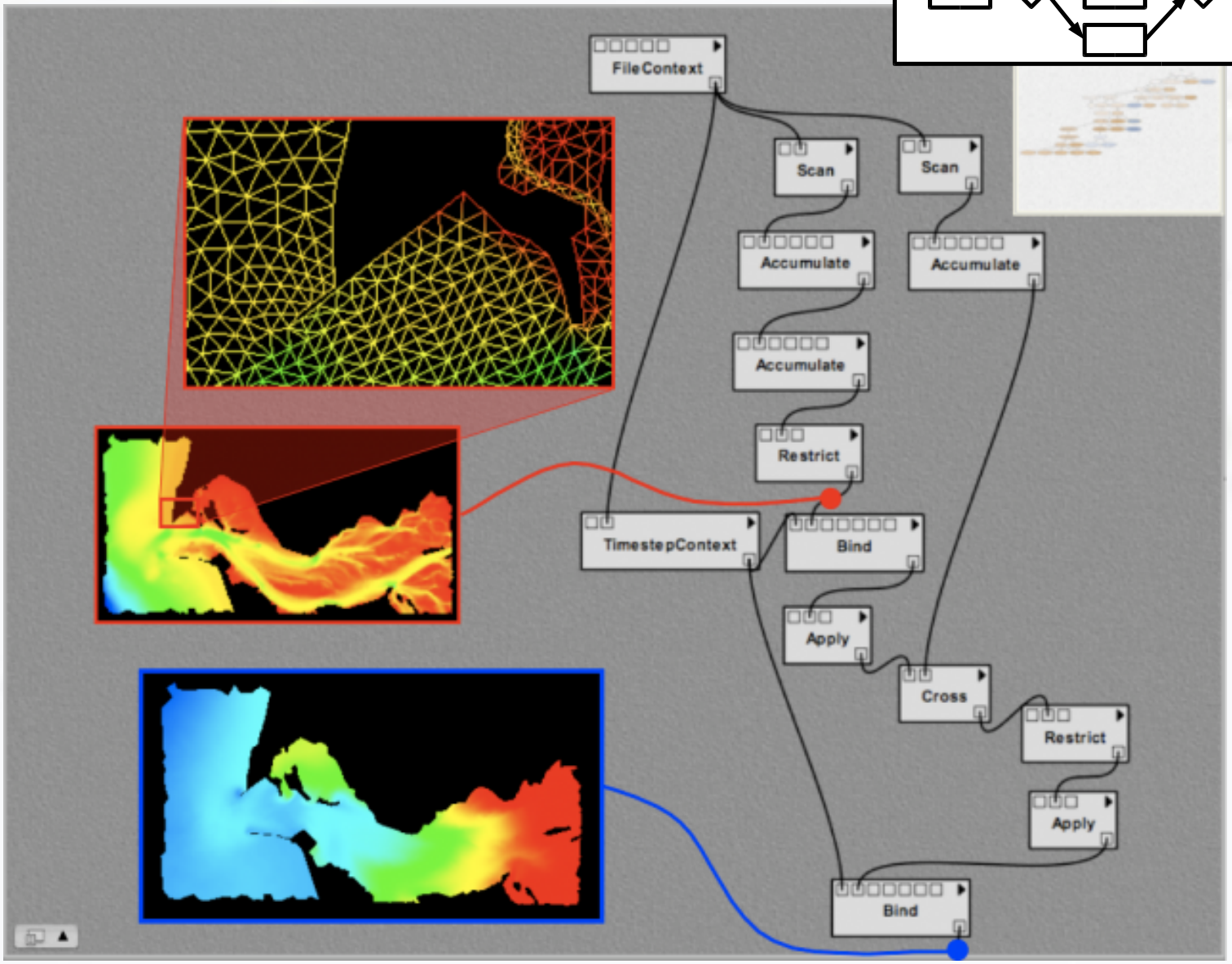

Models

# Open Modeling

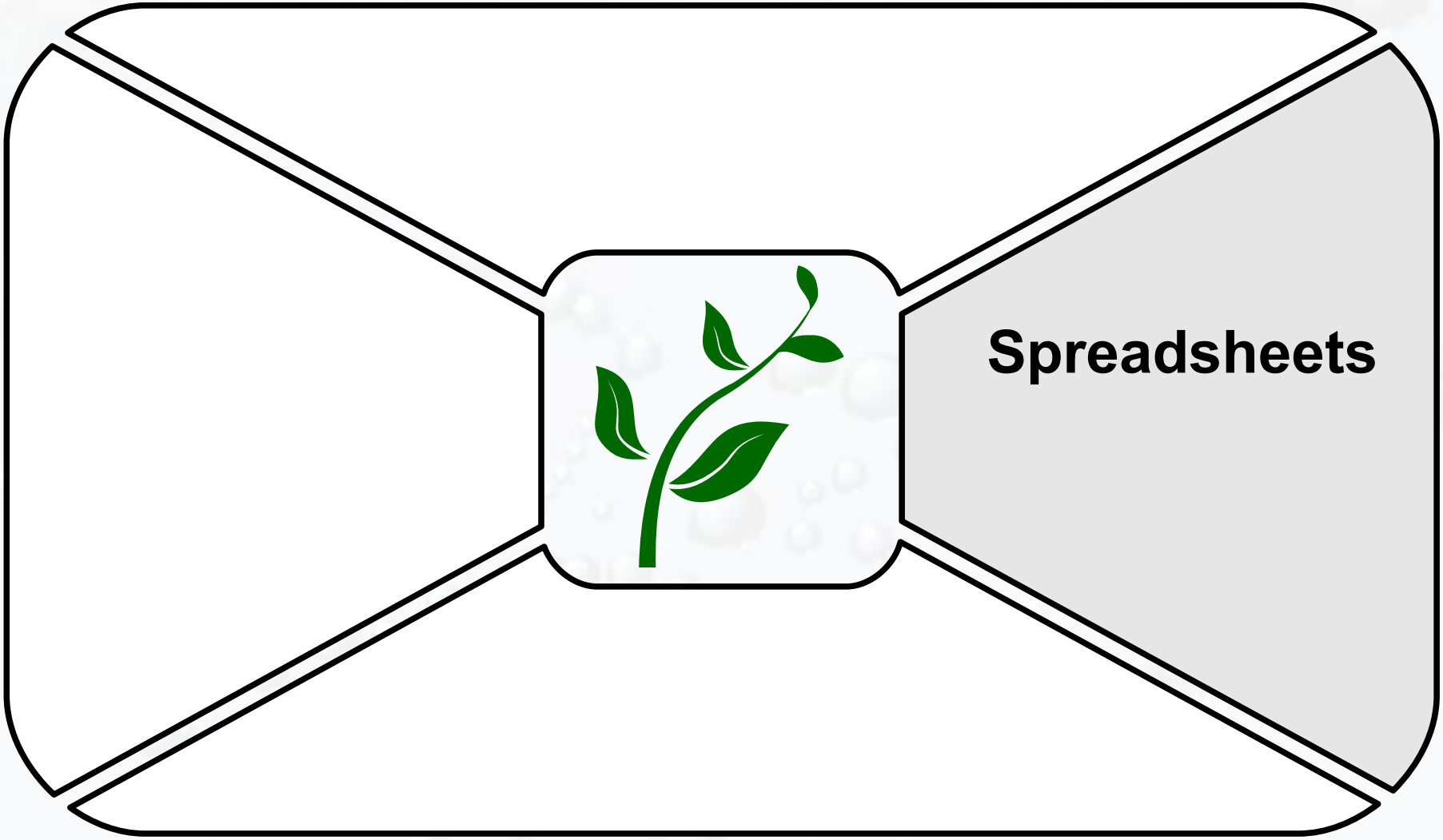
- Share and reuse models
- Confront or connect models
- Reproducibility

# Open Modeling & Workflows






A Science Cloud on your Desktop: VisTrails + GridFields
http://clue.cs.washington.edu/node/1


Spreadsheets

# Implicit Schemas

ID

Schema

Instances

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Espécie | Área | Data | Hora | Isca | Amostra | coleta |
| 22 | | | | | | | |
| 23 | Eulaema cingulata | Bom Jardim | 8/7/2008 | 11:01 | S.M. | 1 | 2 |
| 24 | Eulaema meriana | Bom Jardim | 8/7/2008 | 13:01 | S.M. | 1 | 2 |
| 25 | Eulaema meriana | Bom Jardim | 8/7/2008 | 12:55 | A.B. | 1 | 2 |
| 26 | Eulaema meriana | Bom Jardim | 8/7/2008 | 13:11 | S.M. | 1 | 2 |
| 27 | Eulaema meriana | Bom Jardim | 8/7/2008 | 16:10 | A.B. | 1 | 2 |
| 28 | Eulaema meriana | Bom Jardim | 8/7/2008 | 12:09 | S.M. | 1 | 2 |
| 29 | Eulaema meriana | Bom Jardim | 8/7/2008 | 14:44 | A.B. | 1 | 2 |
| 30 | Eulaema meriana | Bom Jardim | 8/7/2008 | 11:51 | A.B. | 1 | 2 |
| 31 | Eulaema meriana | Bom Jardim | 8/7/2008 | 14:44 | A.B. | 1 | 2 |
| 32 | Eulaema meriana | Bom Jardim | 8/7/2008 | 12:04 | A.B. | 1 | 2 |
| 33 | Exaerete frontalis | Bom Jardim | 8/7/2008 | 11:27 | S.M. | 1 | 2 |
| 34 | Eufriesea flaviventr | Bom Jardim | 8/7/2008 | 12:40 | S.M. | 1 | 2 |
| 35 | Eufriesea flaviventr | Bom Jardim | 8/7/2008 | 13:25 | S.M. | 1 | 2 |
| 36 | Euglossa cognata | Bom Jardim | 8/7/2008 | 9:20 | S.M. | 1 | 2 |
| 37 | Euglossa ignita | Bom Jardim | 8/7/2008 | 11:08 | A.B. | 1 | 2 |
| 38 | Euglossa ignita | Bom Jardim | 8/7/2008 | 10:42 | A.B. | 1 | 2 |
| 39 | Euglossa imperiali | Bom Jardim | 8/7/2008 | 14:32 | S.M. | 1 | 2 |

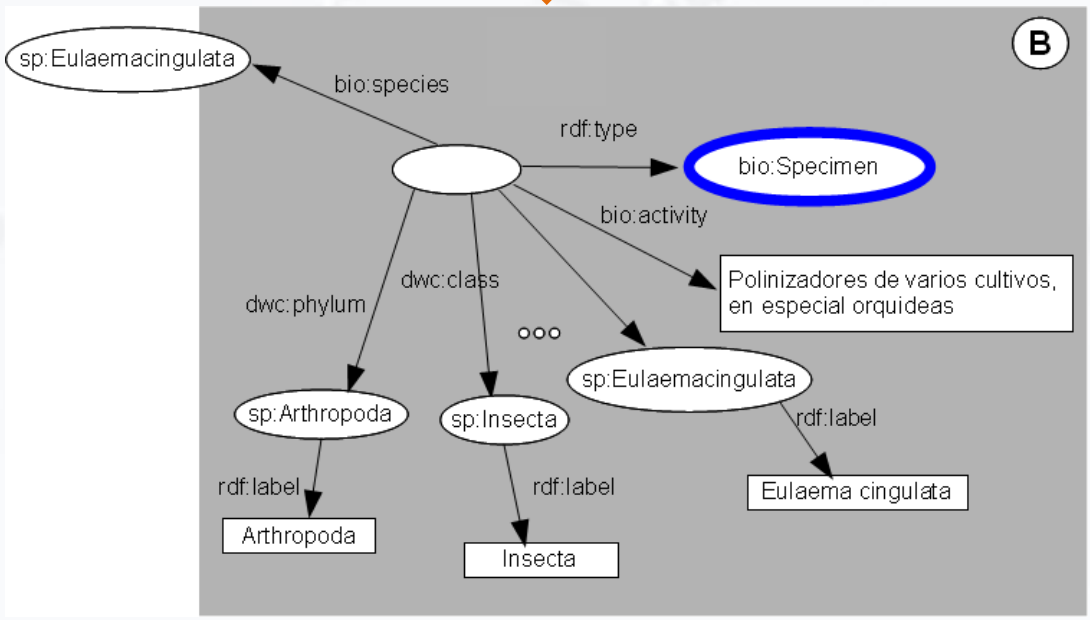
when what where what

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3 | Data | N. da Estação | Latitude | | Longitude | | Prof. | Classificação |
| 4 | | | Graus | Minutos | Graus | Minutos | (metros) | Larsonneur et al.(1982) |
| 5 | 14/12/1997 | 6644 | 25 | 45.80´ | 45 | 11.77´ | 485 | Litoclástico |
| 6 | 14/12/1997 | 6645 | 25 | 44.09´ | 45 | 13.93´ | 256 | Litobioclástico |
| 7 | 14/12/1997 | 6646 | 25 | 43.78´ | 45 | 16.06´ | 198 | Biolitoclástico |

# From Expression to Knowledge

# Schema → Ontology

| | C | D | E | F | G | H | I | J | K | M | O |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | POLINIZADOR | | | | | | | | | | |
| 2 | Filo | Subfilo | Clase | Orden | Superfamilia | Familia | Subfamilia | Género | Especie | Etapa de Vida | Función / Act |
| 73 | Arthropod | Hexapoda | Insecta | Hymenoptera | Apoidea | Apidae | Apinae | *Eulaema* | *Eulaema bombiformis* | Adulto | Polinizadores de varios |
| 74 | Arthropod | Hexapoda | Insecta | Hymenoptera | Apoidea | Apidae | Apinae | *Eulaema* | *Eulaema cingulata* | Adulto | Polinizadores de varios |
| 75 | Arthropod | Hexapoda | Insecta | Hymenoptera | Apoidea | Apidae | Apinae | *Eulaema* | *Eulaema meriana* | Adulto | Polinizadores de varios |



# Building Patterns

ID

Schema

Instances

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Espécie | Área | Data | Hora | Isca | Amostra | coleta |
| 22 | | | | | | | |
| 23 | *Eulaema cingulata* | Bom Jardim | 8/7/2008 | 11:01 | S.M. | 1 | 2 |
| 24 | *Eulaema meriana* | Bom Jardim | 8/7/2008 | 13:01 | S.M. | 1 | 2 |
| 25 | *Eulaema meriana* | Bom Jardim | 8/7/2008 | 12:55 | A.B. | 1 | 2 |
| 26 | *Eulaema meriana* | Bom Jardim | 8/7/2008 | 13:11 | S.M. | 1 | 2 |
| 27 | *Eulaema meriana* | Bom Jardim | 8/7/2008 | 16:10 | A.B. | 1 | 2 |
| 28 | *Eulaema meriana* | Bom Jardim | 8/7/2008 | 12:09 | S.M. | 1 | 2 |
| 29 | *Eulaema meriana* | Bom Jardim | 8/7/2008 | 14:44 | A.B. | 1 | 2 |
| 30 | *Eulaema meriana* | Bom Jardim | 8/7/2008 | 11:51 | A.B. | 1 | 2 |
| 31 | *Eulaema meriana* | Bom Jardim | 8/7/2008 | 14:44 | A.B. | 1 | 2 |
| 32 | *Eulaema meriana* | Bom Jardim | 8/7/2008 | 12:04 | A.B. | 1 | 2 |
| 33 | *Exaerete frontalis* | Bom Jardim | 8/7/2008 | 11:27 | S.M. | 1 | 2 |
| 34 | *Eufriesea flaviventr* | Bom Jardim | 8/7/2008 | 12:40 | S.M. | 1 | 2 |
| 35 | *Eufriesea flaviventr* | Bom Jardim | 8/7/2008 | 13:25 | S.M. | 1 | 2 |
| 36 | *Euglossa cognata* | Bom Jardim | 8/7/2008 | 9:20 | S.M. | 1 | 2 |
| 37 | *Euglossa ignita* | Bom Jardim | 8/7/2008 | 11:08 | A.B. | 1 | 2 |
| 38 | *Euglossa ignita* | Bom Jardim | 8/7/2008 | 10:42 | A.B. | 1 | 2 |
| 39 | *Euglossa imperiali* | Bom Jardim | 8/7/2008 | 14:32 | S.M. | 1 | 2 |

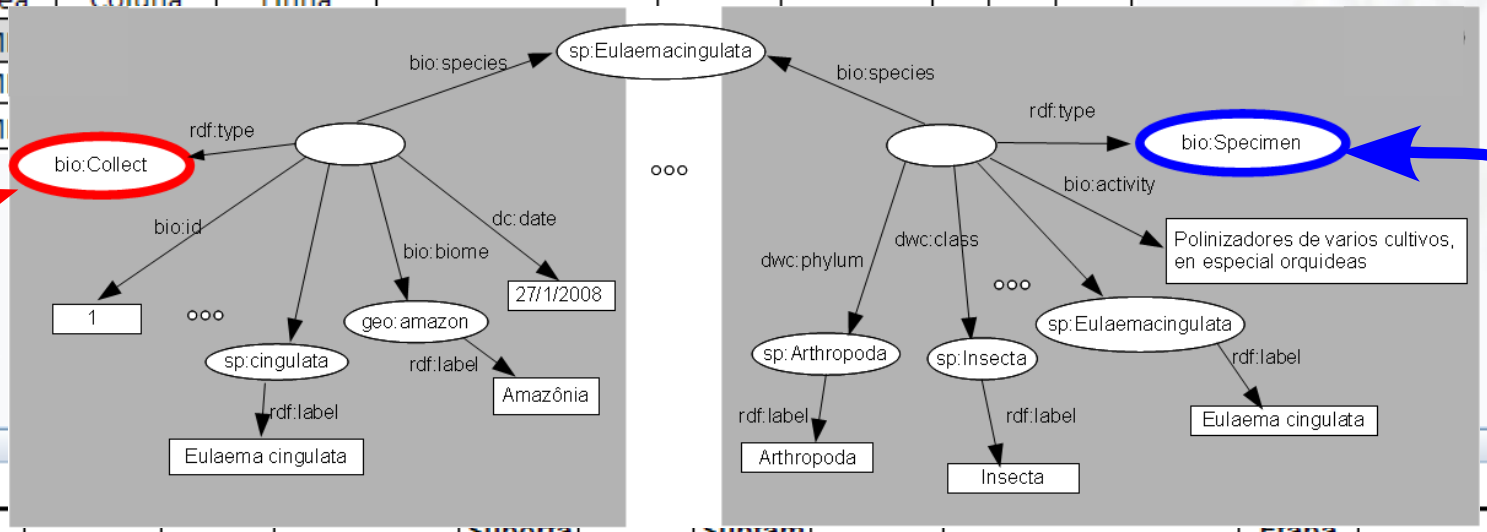
when what where what

Arial 10

N21 $f(x)$ $\Sigma$ =

| | | | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|
| 3 | Data | N. da Estação | Latitude | | Longitude | | Prof. | Classificação |
| 4 | | | Graus | Minutos | Graus | Minutos | (metros) | Larsonneur et al.(1982) |
| 5 | 14/12/1997 | 6644 | 25 | 45.80´ | 45 | 11.77´ | 485 | Litoclástico |
| 6 | 14/12/1997 | 6645 | 25 | 44.09´ | 45 | 13.93´ | 256 | Litobioclástico |
| 7 | 14/12/1997 | 6646 | 25 | 43.78´ | 45 | 16.06´ | 198 | Biolitoclástico |

| | A | B | C | D | E | H | I | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **Planilha Geral dos Dados da Biota** | | | | | | | | | |
| 2 | **Arquivo de origem: Abelhas 2o. Periodo** | | | | | | | | | |
| 3 | ID * | Número da unidade de coleta ** | | | Espécie | Grupo | Bioma *** | dia | mês | ano |
| 4 | | | | | | | | | | |
| 5 | | Área | Coluna | Linha | | | | | | |
| 6 | 1 | A1M | | | | | | | | |
| 7 | 2 | A1M | | | | | | | | |
| 8 | 3 | A1M | | | | | | | | |




| | C | | | | | | | | | | O |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | |
| 2 | Filo | Subfilo | Clase | Orden | Superfamilia | Familia | Subfamilia | Género | Especie | Etapa de Vida | Función / Act |
| 73 | Arthropod | Hexapoda | Insecta | Hymenoptera | Apoidea | Apidae | Apinae | *Eulaema* | *Eulaema bombiformis* | Adulto | Polinizadores de varios |
| 74 | Arthropod | Hexapoda | Insecta | Hymenoptera | Apoidea | Apidae | Apinae | *Eulaema* | *Eulaema cingulata* | Adulto | Polinizadores de varios |
| 75 | Arthropod | Hexapoda | Insecta | Hymenoptera | Apoidea | Apidae | Apinae | *Eulaema* | *Eulaema meriana* | Adulto | Polinizadores de varios |

# Data Exploration Prototype

BioSpread Project - LIS IC UNICAMP - Mozilla Firefox

localhost/biospreadPortal/v2/?task=pages/txnavigator

BioSpread Project

Integrating data from spreadsheet abroad

Home
About
Contact
Examples
Data
Playground
SPARQL Endpoint
License

## A Taxonomy Navigator

**A JavaScript navigator for the extracted data. Navigate through the select boxes and retrieve more information about species**

animalia | chordata | aves | passeriformes | icteridae | icterus | icterus galbula

**Taxonomy Path:** animalia> chordata> aves> passeriformes> icteridae> icterus> icterus galbula

**About: "*icterus galbula*"**

**Associated Events: Records(200)**

Find: paola | Previous | Next | Highlight all | Match case

# IT Agenda

- Data agenda
  - Dataspace
  - Select data to be stored/shared
    - Minimum Information & Minimum Information Model
  - Links
  - Data curation
- Model agenda
  - Open modeling
  - Linking models
- Prototype-driven implementation

# Acknowledgments


- This work was partially financed by:
  - CNPq
  - FAPESP
  - INCT in Web Science
  - NIPE Unicamp
  - Unicamp

University of Campinas – Brazil
LIS – Laboratory of Information Systems
http://www.lis.ic.unicamp.br/

André Santanchè
http://www.ic.unicamp.br/~santanche