Reasoning on Integrated Data and Some Applications in Agronomy

Federico Ulliana
BOREAL Team
Inria & Montpellier University

BOREAL Team

"Knowledge-Representation and Rule-based Languages for Reasoning on Data"

Joint Team with **Inria**, **Univ. Montpellier**, and **INRAe**.

Topics : Symbolic AI, Database Theory, Data Integration, Logic & Graphs

Applications: Agronomy (but not only)

Software: InteGraal (Reasoning on Integrated data)

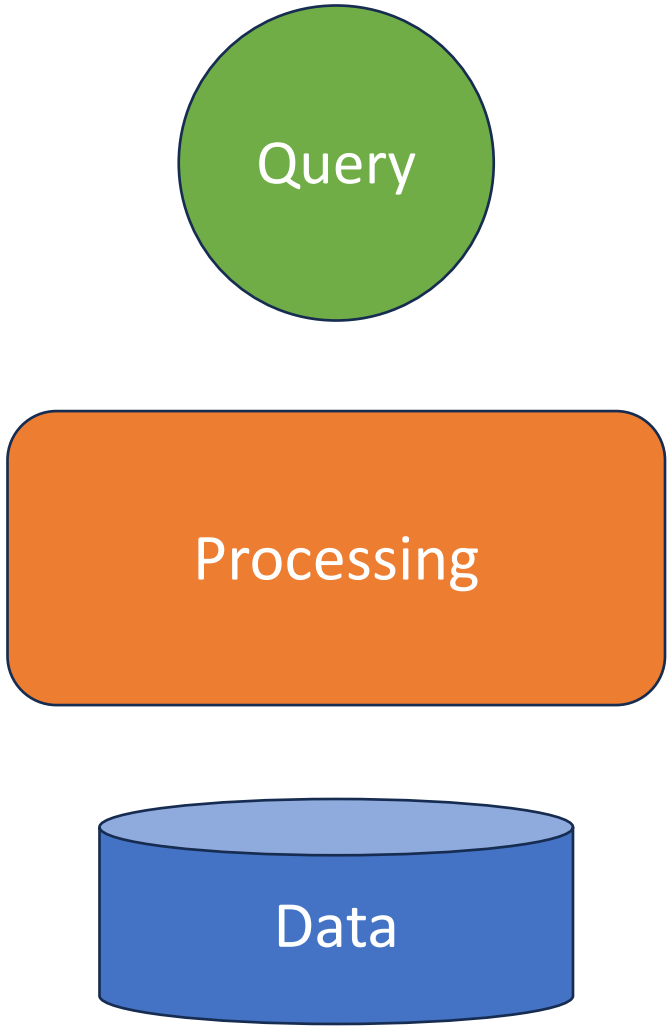
Exploiting Data

Data-driven society: organizations constantly looking for new insights to support their next decision

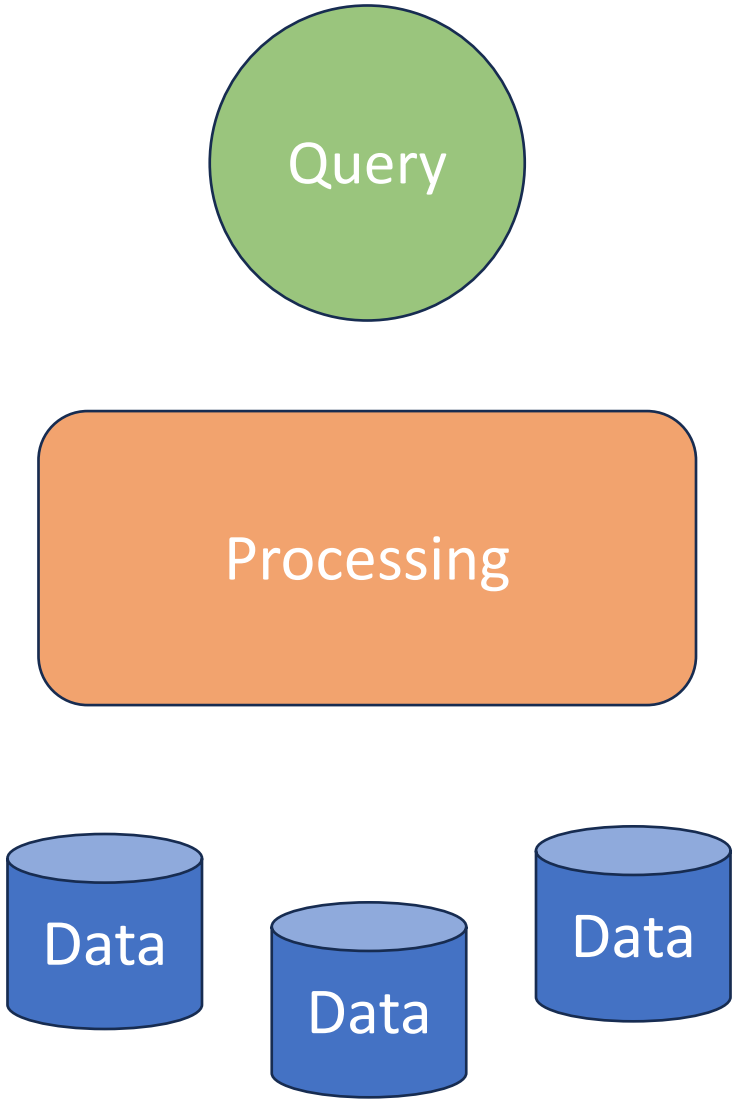
- Healthcare
  - Disease risk prediction, clinical trials
- Agronomy and Agroecology
  - Cultivation, transformation, land management
- Manufacturing, Industry
  - Accountancy, production, assets, clients
- Sciences
  - Scientific datasets, all domains
- Government and PA
  - Social and economic
- National Security
  - Monitor positional data
- Smart Cities
  - Sensor, open data on public services

Yet, data processing and data governance still remain challenging issues.

The Workflow



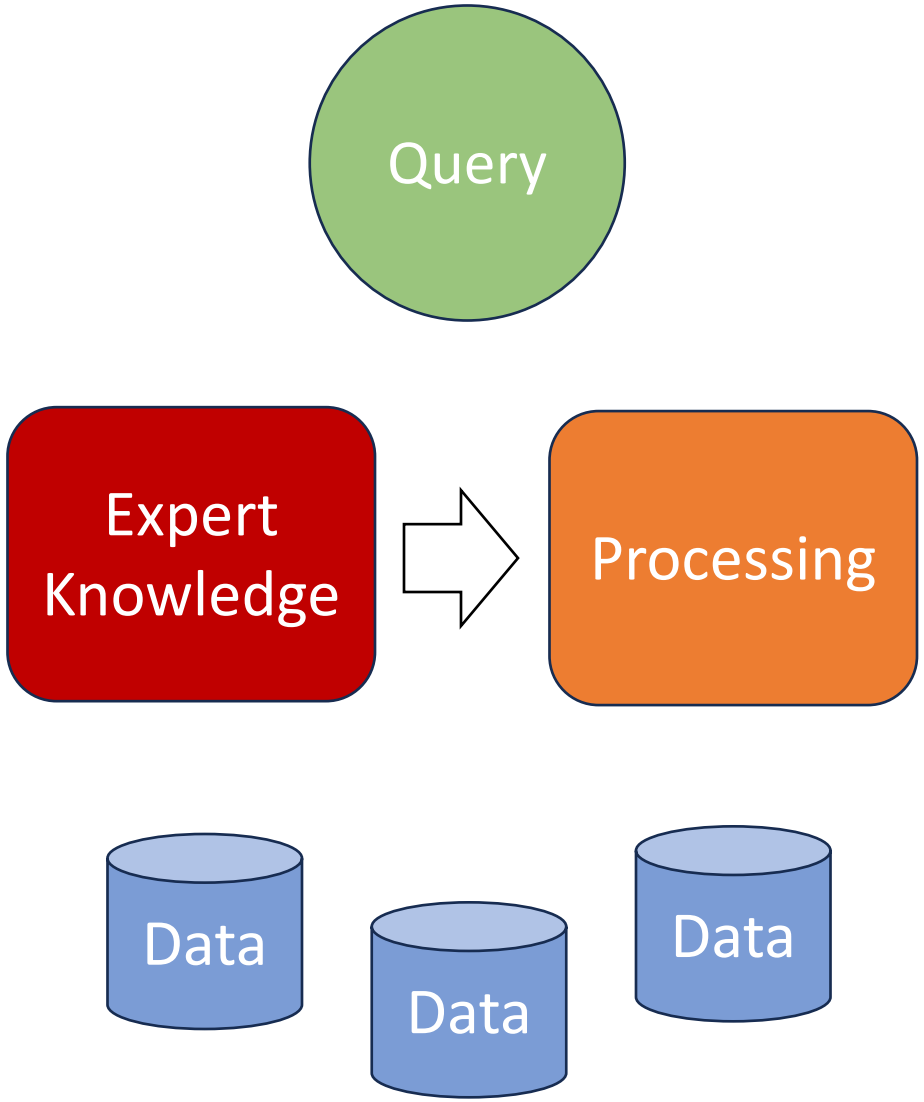
In Reality



Heterogeneous & Federated Data

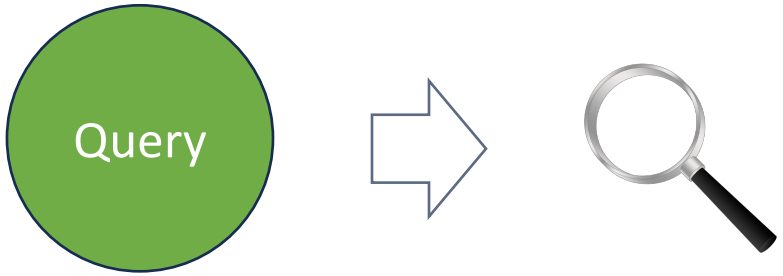
In Reality

Exploiting Data requires Domain Knowledge



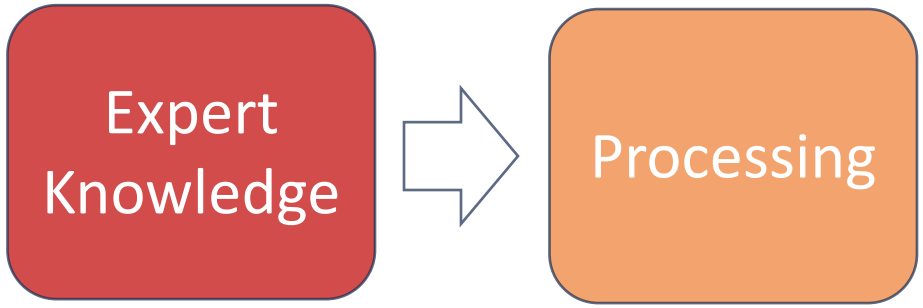
Heterogeneous & Federated Data

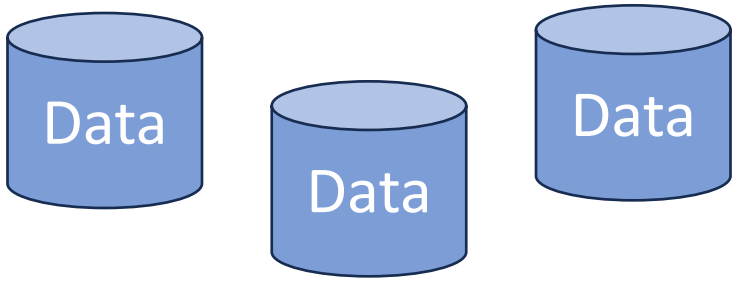
In Reality



Explanations for queries and errors

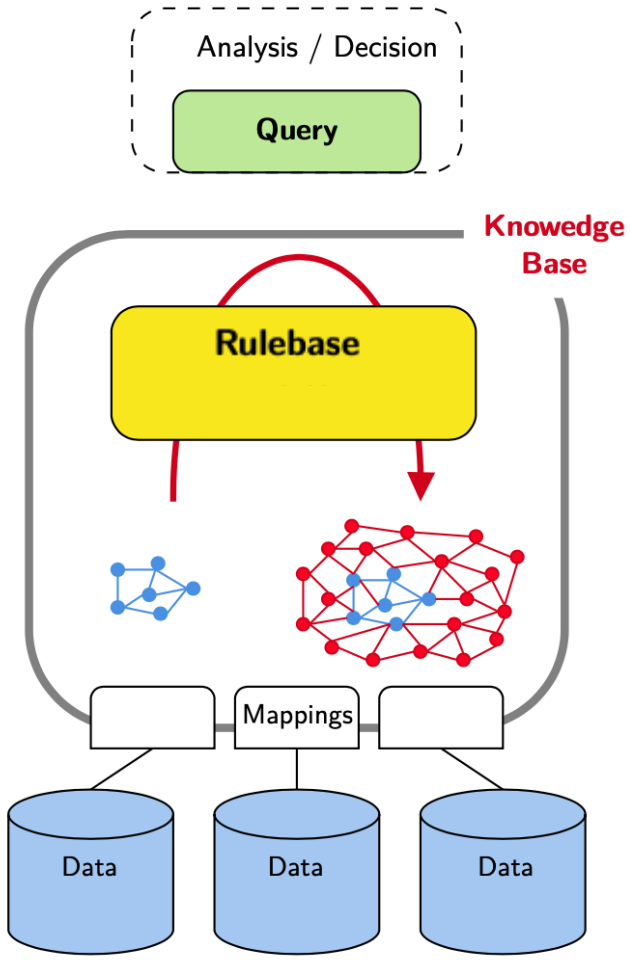
Exploiting Data requires Domain Knowledge





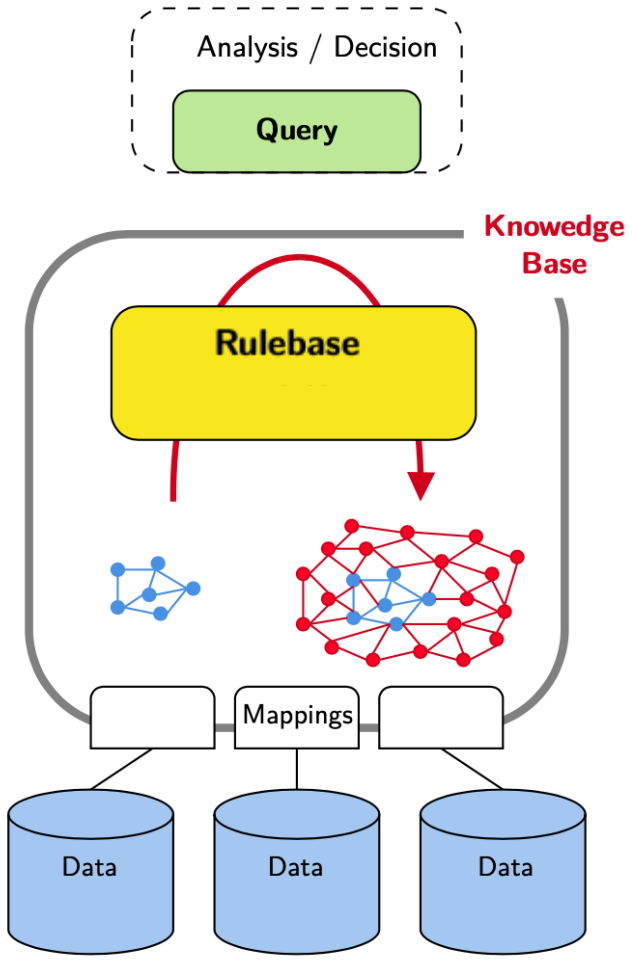
Heterogeneous & Federated Data

Knowledge-Based Data Management (KBDM)



1. Three layer architecture: sources, mappings, KB
   - common to Data-Integration and Ontology-Based Data-Access
2. Domain Knowledge "the glue"
   - Semantics: (quality) KB ← map (raw) sources
   - Abstraction: KB processing → translated to sources
3. Rule-based Languages "the driver"
   - Domain expert knowledge (eg. Ontologies) && data processing
   - Declarativity, Modularity, Explainability

Knowledge-Based Data Management (KBDM)



(Query)

?(X):- Patient(X) $\xrightarrow{ans}$ S1_Alice S_1_Bob

(Facts/Data)

Patient(S_1_Alice) Patient(S_1_Bob)

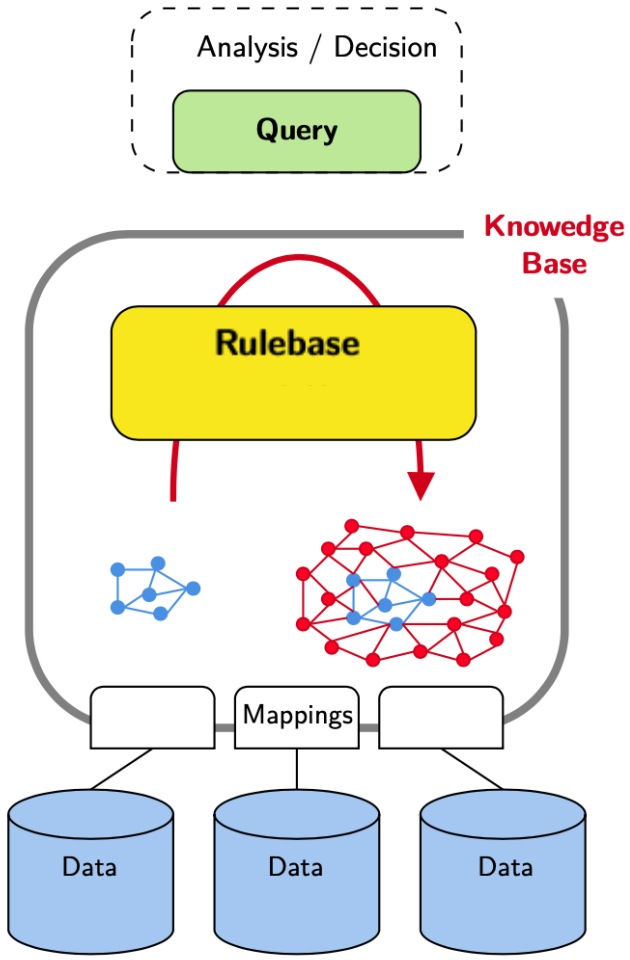
(Mapping)

patients.name.$val $\rightsquigarrow$ Patient(f($val))

(Source JSON)

```
{patients:[{name:Alice},
 {name:Bob,diagn:high_oxytocin}] }
```

Knowledge-Based Data Management (KBDM)



(Query)

?(X):- ObesityRisk(X) $\xrightarrow{ans}$ S_1_Bob

(Rules)

$\forall X.Patient(X)$
$Diagnosis(X, SN{:}high\_oxytocin) \rightarrow ObesityRisk(X)$

(Facts/Data)

$Patient(S_1\_Alice)$ $Patient(S_1\_Bob)$
$Diagnosis(S_1\_Bob, SN{:}high\_oxytocin)$

(Mapping)

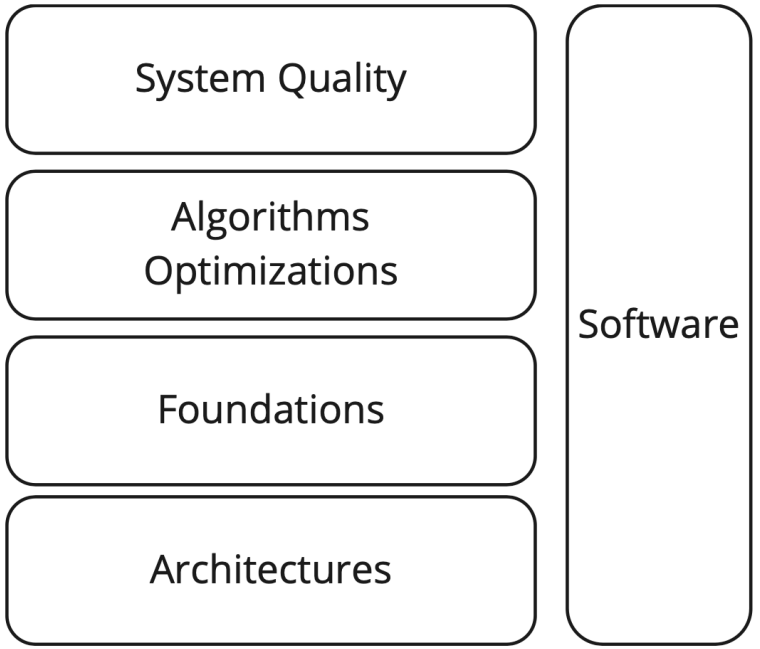
$patients.name.\$val \rightsquigarrow Patient(f(\$val))$
$patients[name.\$n, diagn.\$d] \rightsquigarrow Diagnosis(f(\$n), g(\$d))$

(Source JSON)

```
{patients:[{name:Alice},
 {name:Bob,diagn:high_oxytocin}] }
```

Boreal Project

1. Architectures of KBDM Systems
   - Multi-source/level, Mapping Languages, Applications
2. Foundations of Reasoning and Query Answering
   - Decidability & (fine-grained) Complexity, Extended Existential Rules
3. Algorithms and Optimizations for Query Answering
   - Combined Approaches, Leveraging Database Technology
4. Evaluating and Improving the Quality of KBDM Systems
   - Explanations, Provenance, Inverse Rewriting
5. Software



Some Applications in Agronomy

1. Data Preparation
   - Training-Set Preparation for ML and Food Package Selection

2. Explanations
   - Enforcing Legislation on Chemical Spraying

3. Domain Expert Knowledge
   - Agroecosystems Design for Vine Grassing

(disclaimer)
high-level view
of the KDBM model
behind the concrete problem

(1) Training-Set Preparation for ML and Food Package Selection

What is the “best” package for a specific type of food ?

- Shelf-life vs. environmental impact

w/ Pierre Bisquert, Patrice Buche, David Camarazo, Julien Cufi, Michel Leclère, Florent Tornil

(1) Training-Set Preparation for ML and Food Package Selection

- Use a ML model to predict o2-solubility
- Prepare training-set data **(our task)**

Training set

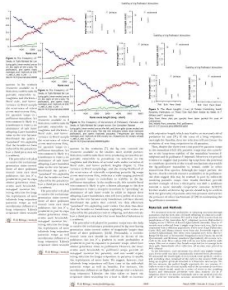
Y

food product	o2-solubility	nutritional composition

(1) Training-Set Preparation for ML and Food Package Selection

nom	solubilité	matière grasse	proteine	sel	eau	sucre
Fish - controlW	0.8 (@web)	6 (@web)	21 (Me@tyLab)	0.5 (@web)	66 (@web)	– *(if missing set to 0)*

(1) Training-Set Preparation for ML and Food Package Selection



data from scientific literature

default values (rented data)

RDF Triplestore

@Web

Web API

MeatyL@B

Knowledge

Nutritional Composition

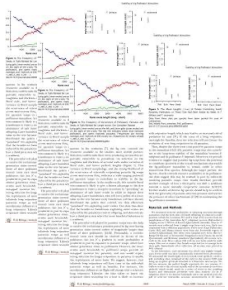
Processing

@Web
better-than
MeatyL@b

Exploitation

Export

(1) Training-Set Preparation for ML and Food Package Selection



data from scientific literature

RDF Triplestore

@Web

default values (rented data)

Web API

MeatyL@Bx

Knowledge

Nutritional Composition

Processing

@Web
better-than
MeatyL@b

Exploitation

Export

Mapping

Q_{SPARQL} ⇝ **@Web**(Product,Comp)

Rules

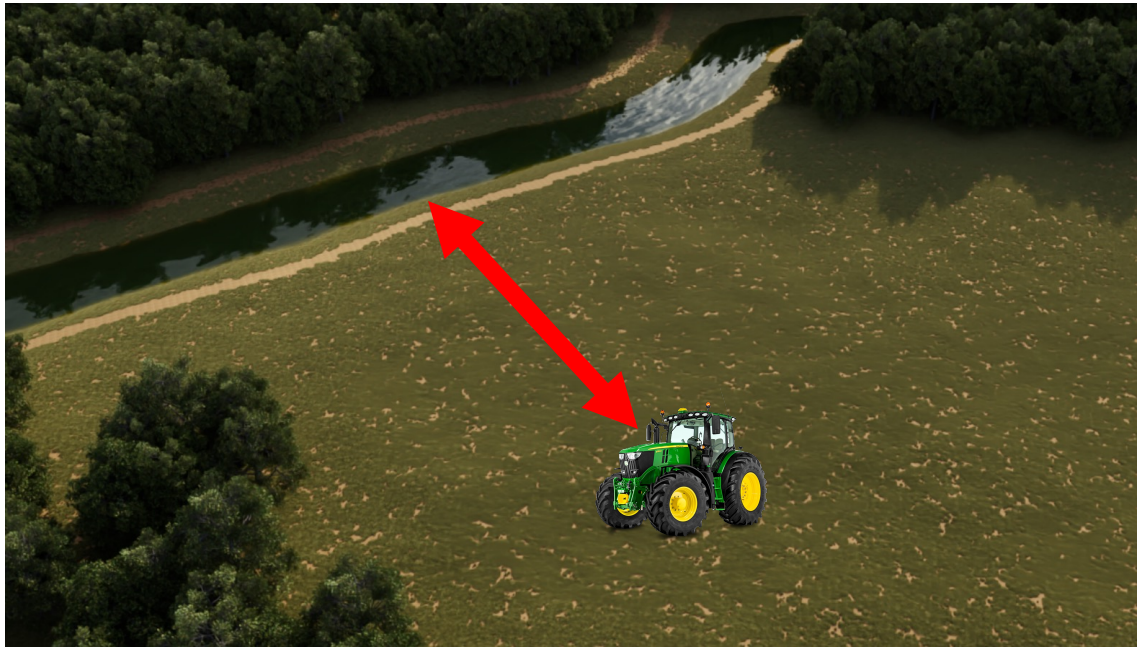
MeatyL@b(Product,Comp)
and NOT **@Web**(Product,C)
→ **Composition**(Product,Comp)

(2) Enforcing Legislation on Chemical Spraying

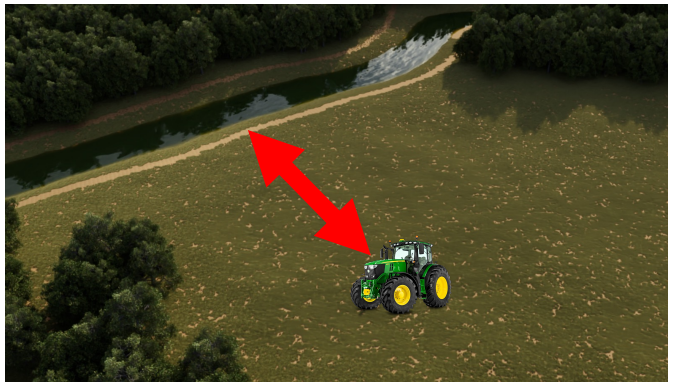
German Research Center for Artificial Intelligence

"Remove harmful weeds on a field,

but **don't use chemical agents close to a water body**!"



(2) Enforcing Legislation on Chemical Spraying



Explanations

There is a violation at (position ***P***, time ***T***).

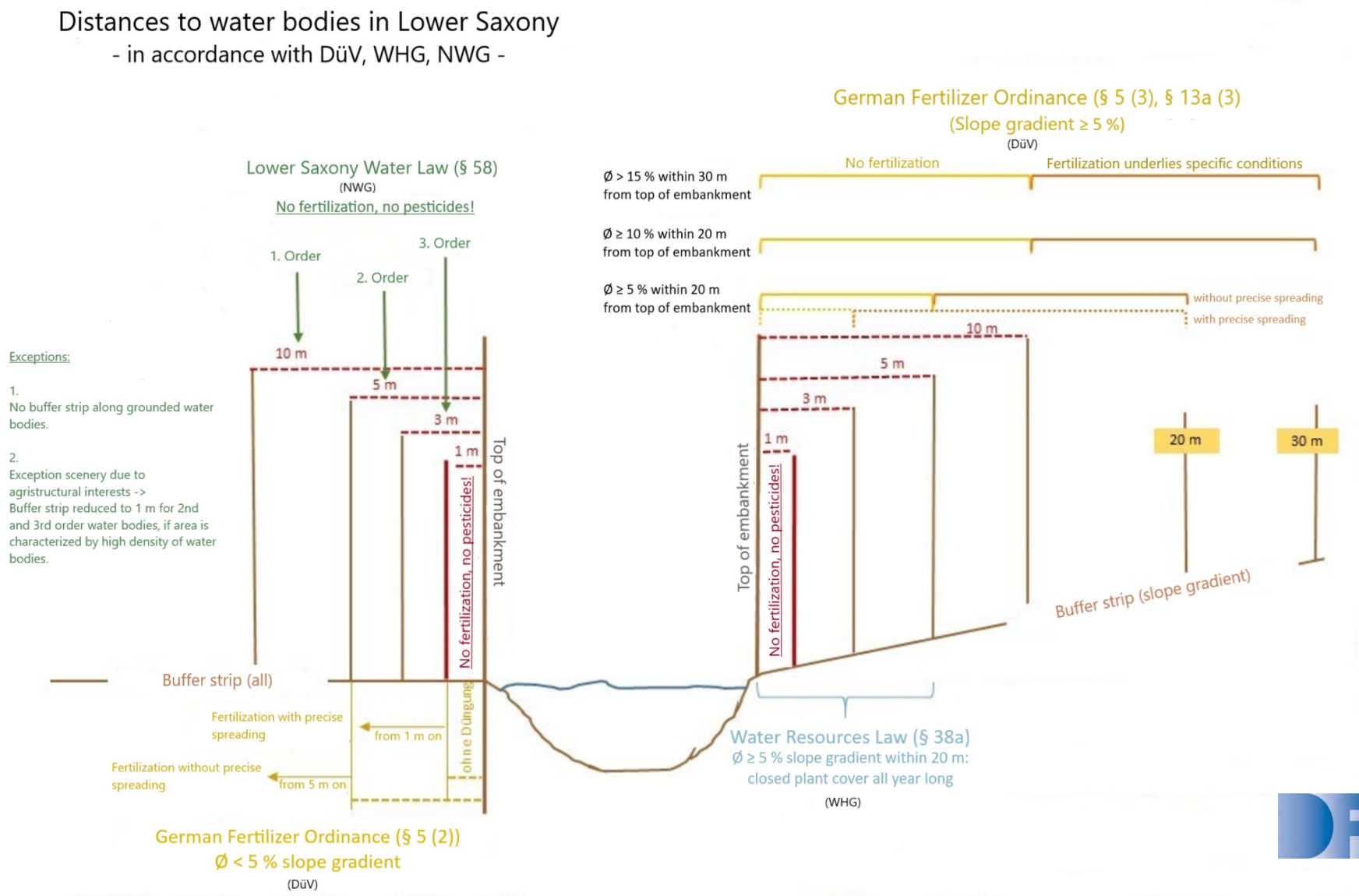
The tractor is at a distance ***D*** from the river, on a field with a slope of **S%**, and is spraying a chemical of type **Y** using a precise spreading device **PS**.

In this case the legislation article **L** states that the minimal required distance must be at least **M** meters, whereas **D** < **M**.

The legislation article **L** defines a “relevant buffer constraint” which imposes the minimal distance **M** according to the slope of the field **S%** and the use of a precise spreading device **PS** and chemicals of type ***Y***.

German Research Center for Artificial Intelligence

(2) Enforcing Legislation on Chemical Spraying



German Research Center for Artificial Intelligence

(2) Enforcing Legislation on Chemical Spraying

Land (static)

Vehicle (dynamic)

Meteo (Web API)

Legislation(s)

Knowledge	Processing	Exploitation
Land, Spraying Vehicule, Legislation	Law Violations	Explain Why

(2) Enforcing Legislation on Chemical Spraying

Land (static)

Vehicle (dynamic)

Meteo (Web API)

Legislation(s)

Knowledge	Processing	Exploitation
Land, Spraying Vehicule, Legislation	Law Violations	Explain

Rules

LawConstraint(Law, Params…, SafeDistance) &

Spread(time,position,Params…,Distance) &

Distance < SafeDistance

→ **Violation**(time, position,Law)

(3) Agroecosystems Design for Vine Grassing

Selecting plant species for targeted ecosystem-services

Elie Najm PhD 2019-2022 co-directed by Christian Gary (Inrae, ABSys) and Marie-Laure Mugnier
Joint work with Jean-François Baget (BOREAL), Raphael Metral (ABSys) and Léo Garcia (ABSys)

(3) Agroecosystems Design : Recommendations

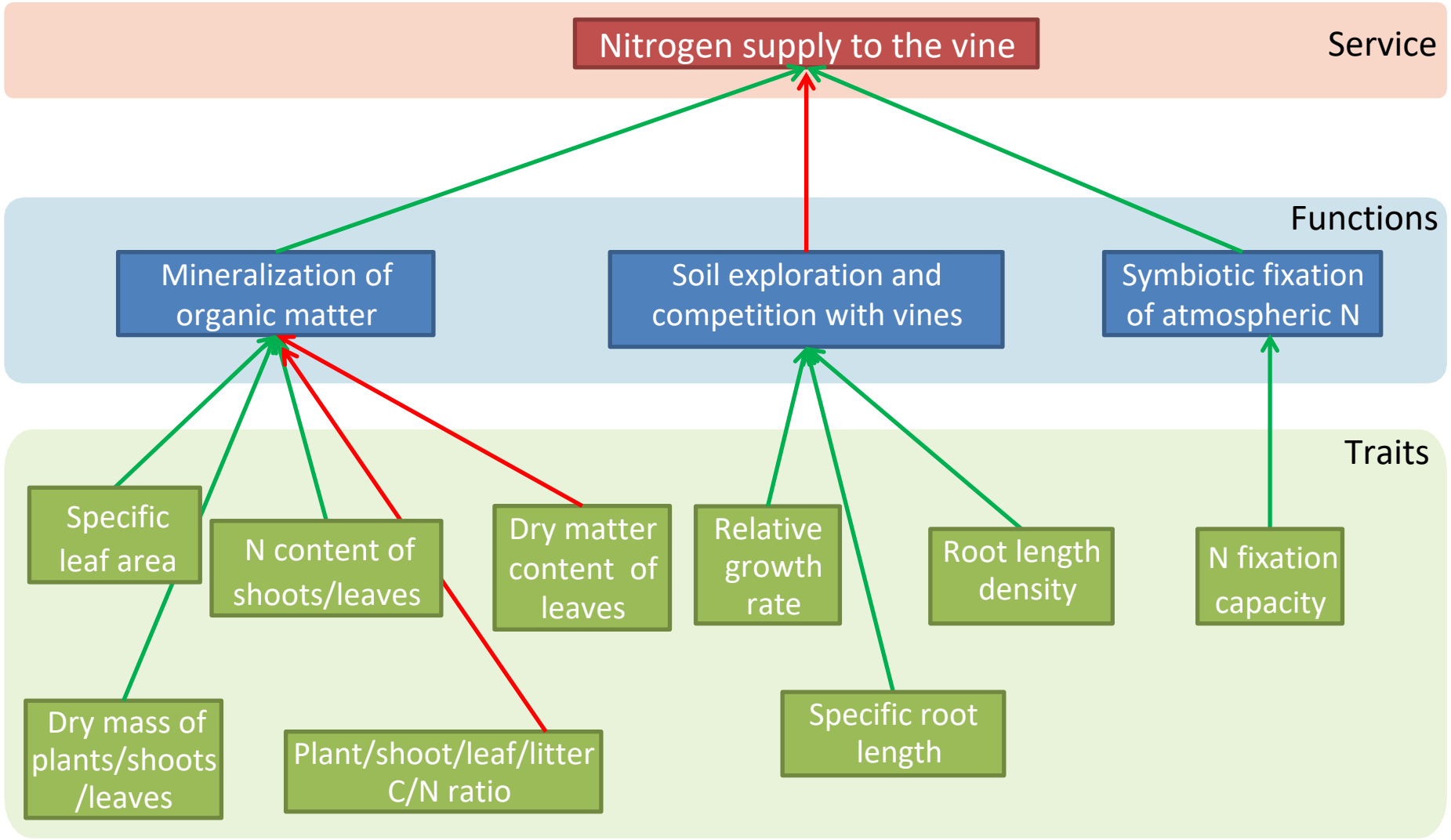
What are the 10 best species for nitrogen supply to the vine
(with a reliability of at least 50%?)

species	value 1	reliability
lotus corniculatus	0.77680105	89
trifolium repens	0.76497017	78
trifolium pratense	0.75703032	78
trifolium hybridum	0.72317251	55
trifolium dubium	0.72164679	67
lathyrus pratensis	0.71902614	55
anthyllis vulneraria	0.71794294	67
trifolium campestre	0.70497417	55
medicago lupulina	0.70018017	67
onobrychis viciifolia	0.69490763	55

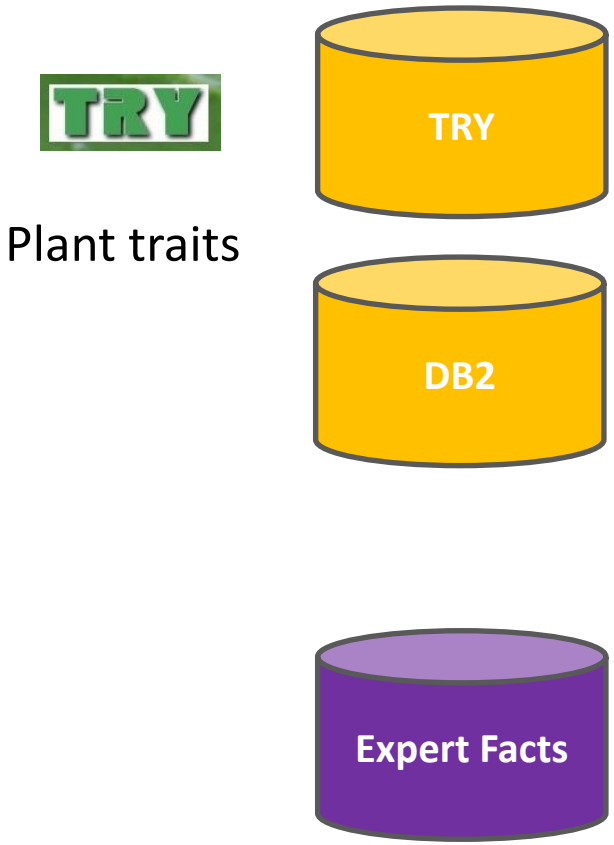


...

(3) Agroecosystems Design : Expert-Knowledge



(3) Agroecosystems Design



Knowledge

Traits, Ecosystem Functions & Services

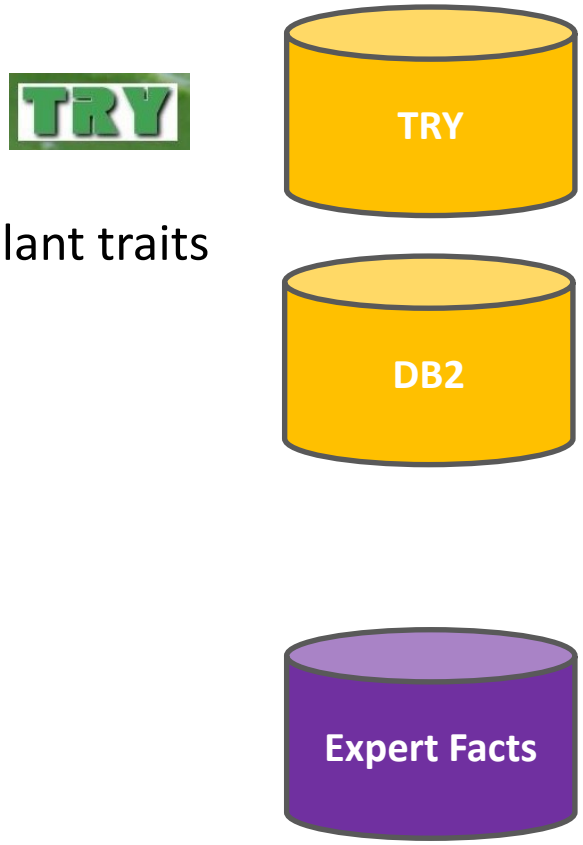
Processing

Aggregate Traits into Functions then into Services

Exploitation

Explain

(3) Agroecosystems Design



Knowledge

Traits, Ecosystem Functions & Services

Processing

Aggregate Traits into Functions then into Services

Exploitation

Rank

Explain

Rules

Traits(Species, T_1:V_1) … **Traits**(Species, T_N:V_N) &

Contributing(Function, T_1 ,…,T_N)

→

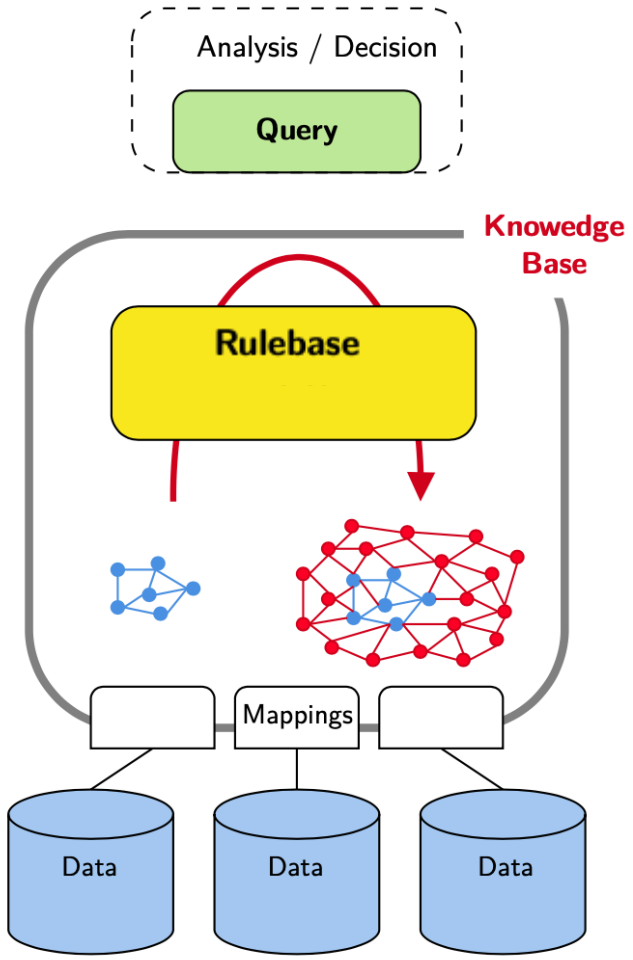
FunctionValue(Function, Species, **fct:aggregate**(V_1…V_N))

InteGraal

gitlab.inria.fr/rules/integraal

- Java Tool for Reasoning on Integrated Data developed by BOREAL
- Used for all use-cases presented

Conclusions: Reasoning over Integrated Data



- Knowledge-Based Data-Management as a methodology for exploiting heterogeneous and federated data
- Data preparation and cleaning are mandatory
- Domain knowledge is crucial for data quality and data-service mainteinance
- Declarative rule languages for expressing both domain knowledge and data processing
- Explanations help making sense of answers and increase the quality of the integration system