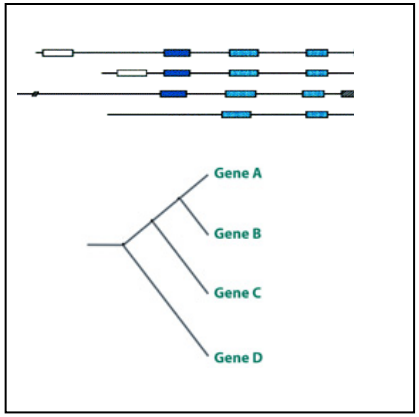
Ontology-based data integration projects at IBC

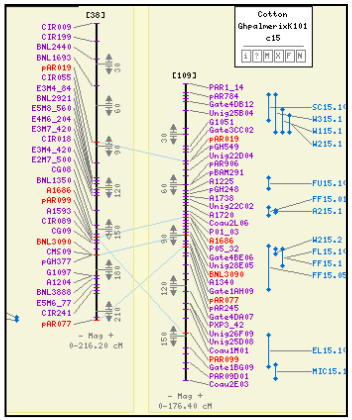
Pierre Larmande

Data and resources

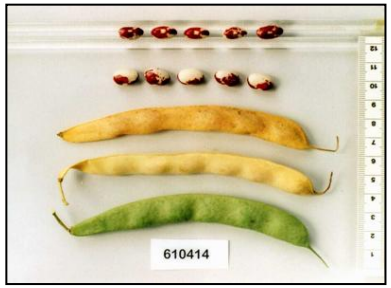
Contexte



Comparative genomics

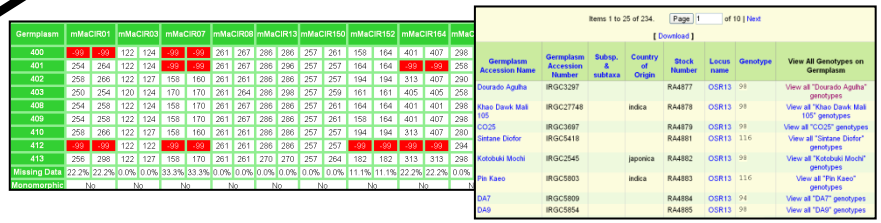


Genetic and physical maps, QTL



Genetic ressources

Markers

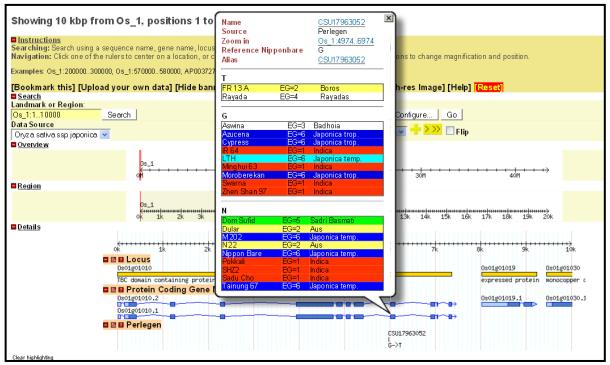


Genotyping studies

Phenotypes

Individuals

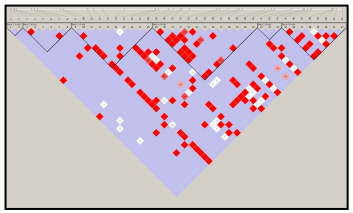
Genomic annotations



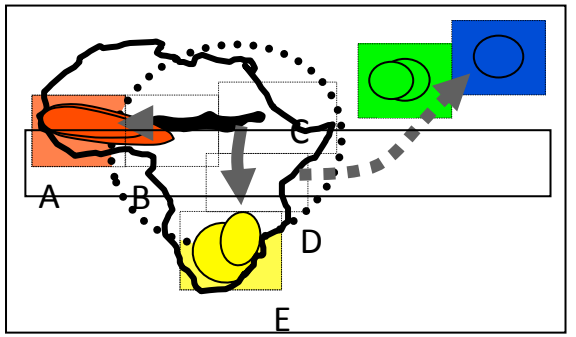
From M. Ruiz

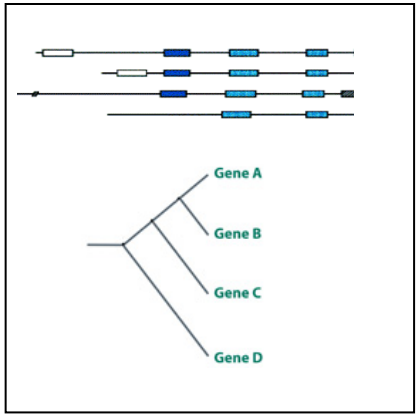
Functional annotations

Analysis Workflows

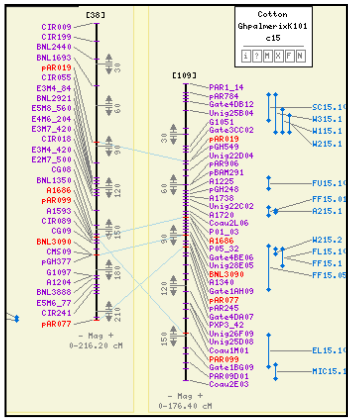


Geographic data

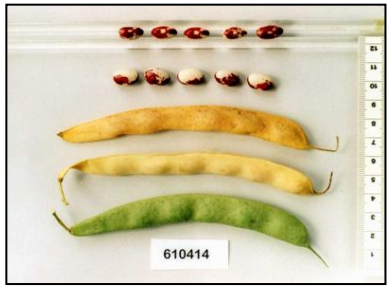




Comparative genomics



Genetic and physical maps, QTL



Genetic ressources

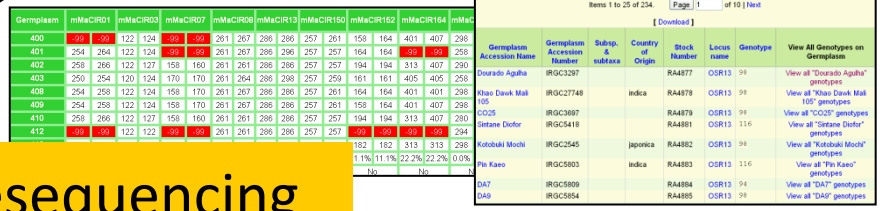
high-throughput phenotyping

Phenotypes

Individuals

Markers

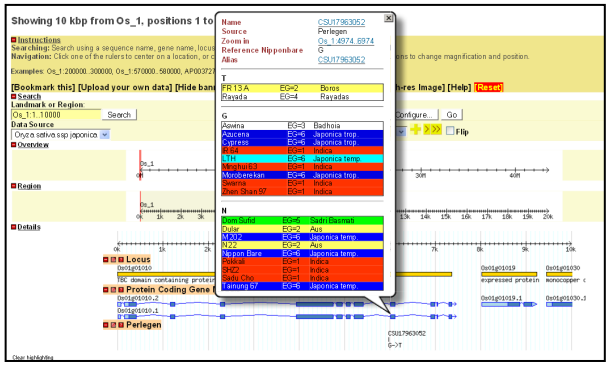
large genome resequencing



Genotyping studies

Geographic data

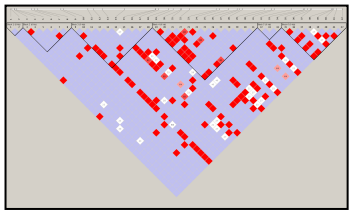
Genomic annotations

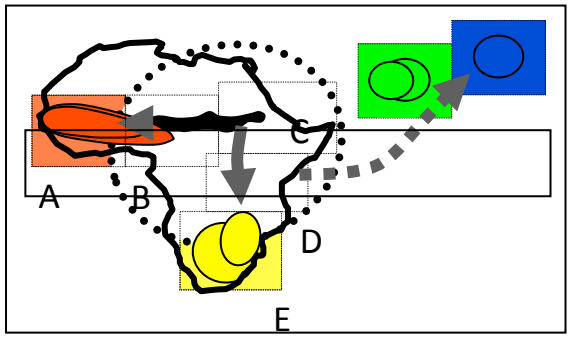


From M. Ruiz

Functional annotations

Analysis Workflows





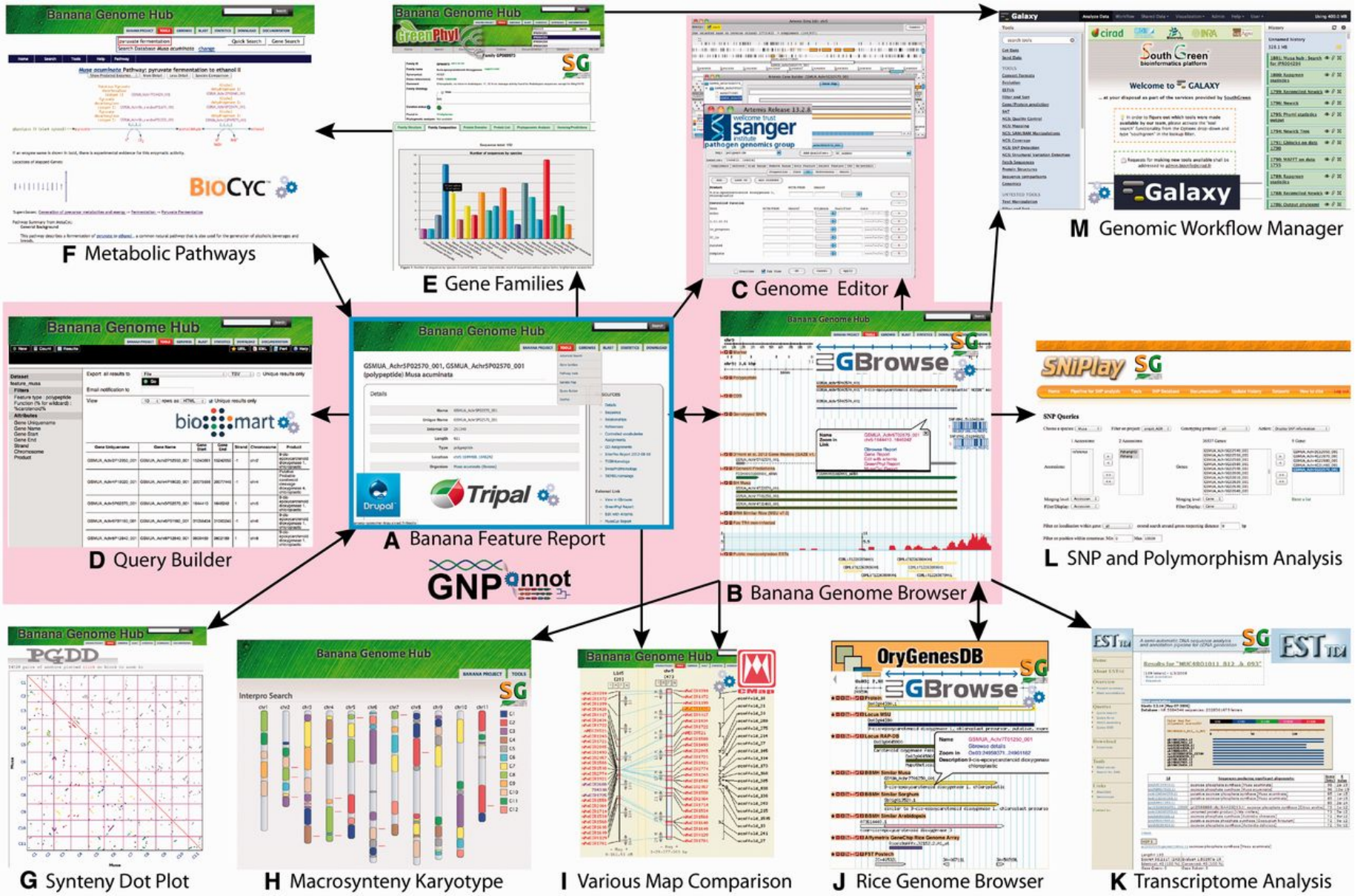
- **Annotation and comparative genomics**
  - GNPAnnot
  - GreenPhyl
  - Analysis of genome sequences
  - Comparative population genomics

- **Information systems**
  - TropGene (genetic)
  - Integrated rice functional genomics (phenotypic)
  - Genome

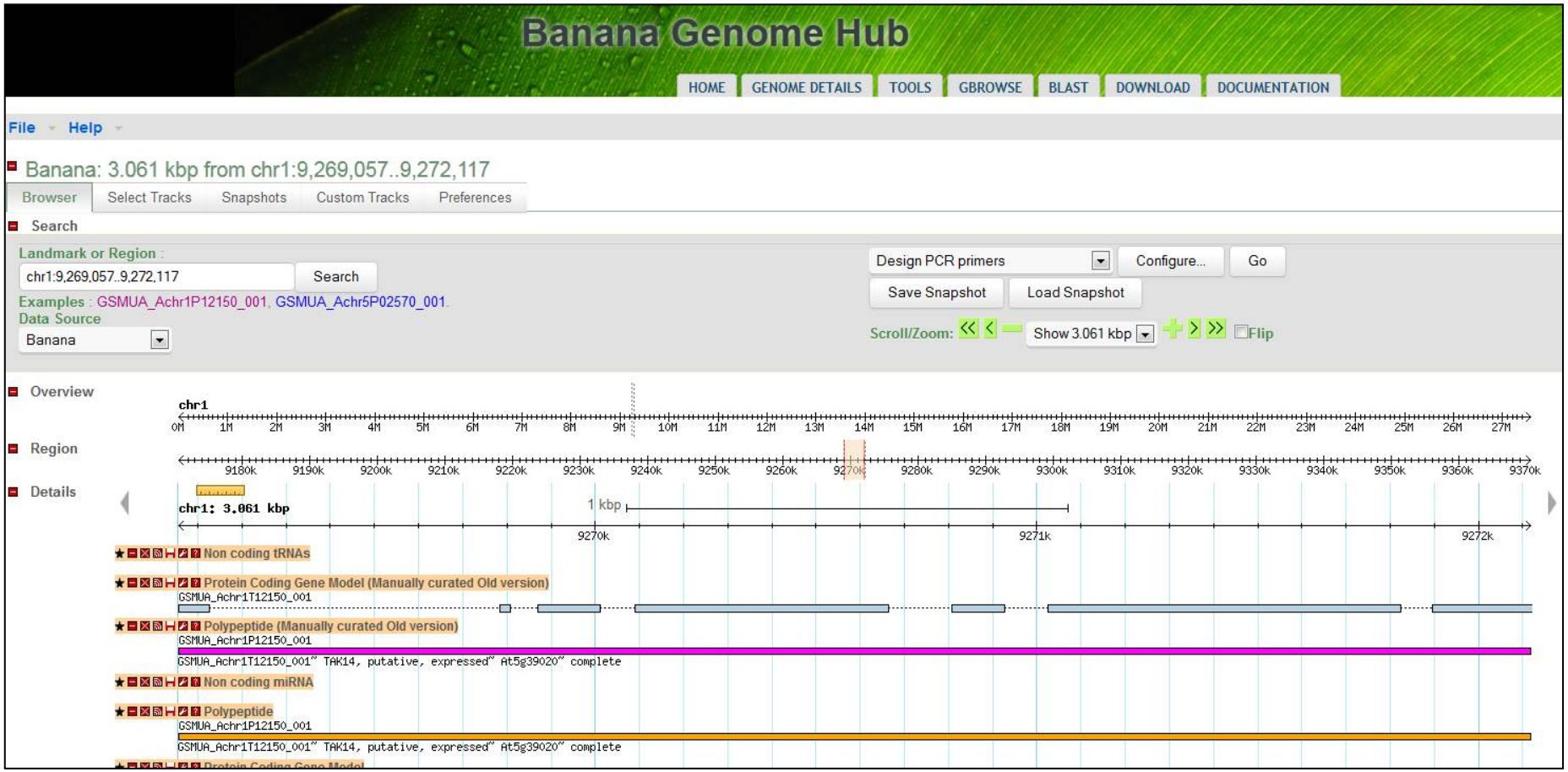
- **Integrated worflows**
  - ESTtik
  - SNiPlay
  - Galaxy

From M. Ruiz

The banana genome hub



Droc, G., Larivière, D., et al. Database, 2013



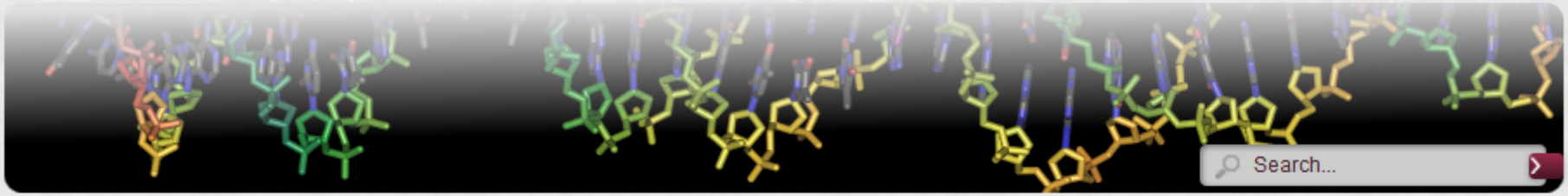
- Génome et espèces centré
- Intégration légère : pas de requêtes complexes multi-bases
- Intégration de nouvelles ressources, de nouveaux types de données : besoin de développement, intégration manuelle
- Pas de raisonnement, inférence, fouille de données automatique

From M. Ruiz

The Computational Biology Institute (IBC)

Computational Biology Institute

Modeling, processing and analysis of large scale data in biology, health, agronomy and environment



News / Actualités

An article for general audience in June ...

June, 2013 issue of the magazine "Pour la Science" published an article of A. Kajava about recent development in the understanding the molecular mechanisms of neurodegenerative diseases. This article is...

The Computational Biology Institute (IBC) aims at the development of innovative methods and software to analyze, integrate and contextualize large-scale biological data in the fields of health, agronomy and environment. Scalable computational solutions able to handle this ever-increasing volume of data constitute the present and future bottleneck that may limit their economic impact. Several branches of research will thus be combined: algorithmics (combinatorial, numerical, highly parallel, stochastic), modeling (discrete, qualitative, quantitative, probabilistic), and data management and information retrieval (integration, workflows, cloud). Concepts and tools will be validated using key applications in fundamental biology (transcriptomics, structure and function of proteins, development and morphogenesis), health (pathogens, cancer, stem cells), agronomy (plant genomics, tropical agriculture), and environment (population dynamic, biodiversity).The project is divided into five complementary work-packages that include the main aspects of processing biological data on a large scale:

Next seminar

CRAC: an integrated approach to the analysis of...

Plenary sessions

Nicolas Philippe,
Institut de Recherche en Biothérapie (IRB), Montpellier, France.
Friday, June 14 2013,
2pm, room 127 (see plan)

Read more

Suscribe to ibo-seminar list

- WP1-HTS: Methods for high-throughput sequencing analysis
- WP2-Evolution: Scaling-up evolutionary analyses
- WP3-Annotation: Structural and functional annotation of proteomes
- WP4-Imaging: Integrating cell and tissue imaging with Omics data
- WP5-Databases: Biological data and knowledge integration

The Rice Semantic Hub : use case

L'objectif principal est l'extraction de connaissances issues des grands projets génomes et de phénotypage à haut débit.

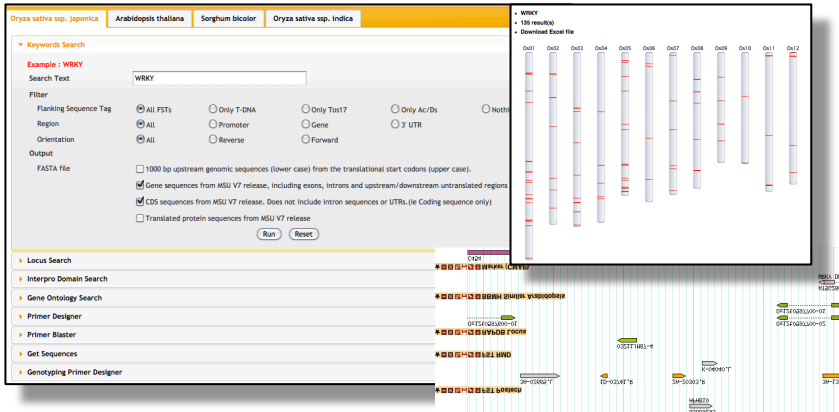
Nous allons lier les données « multi-échelle » : environnements, phénotypes, ressources génétiques, données de génotypage, QTL et séquence génomique.

Développement d'un environment de travail avec de multiples fonctionnalités pour la visualisation de données, requêtes et analyses.

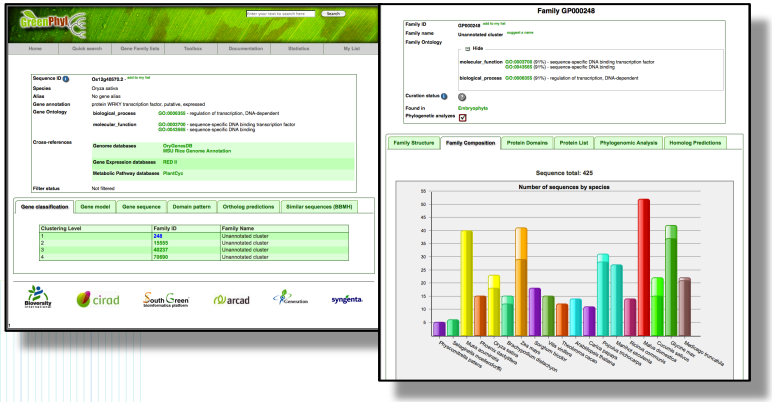
From M. Ruiz

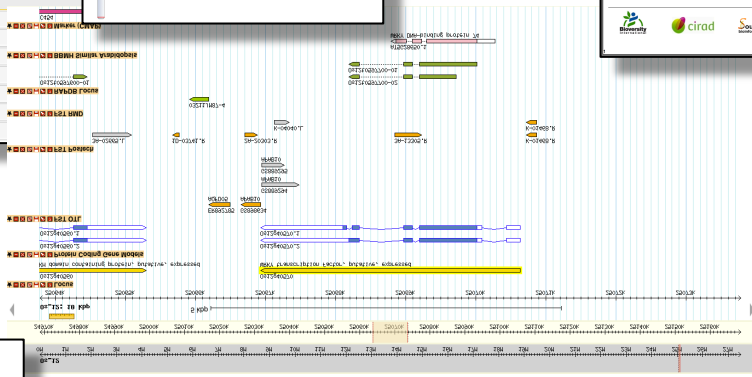
The Rice South Green Hub

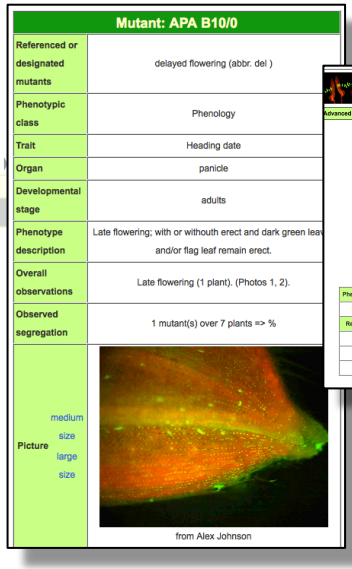
OryGenesDB

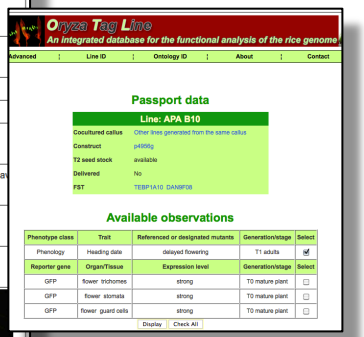


GreenPhylDB

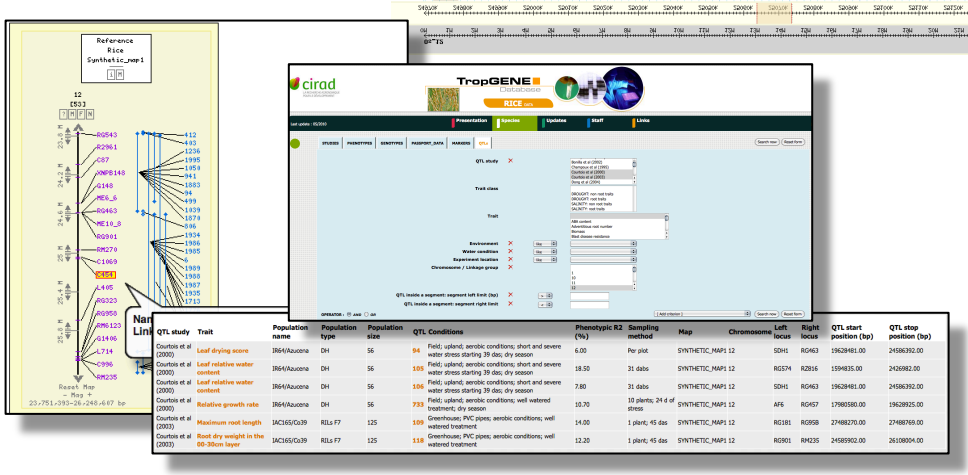








Oryza Tag Line



TropGeneDB

Resources and tools useful for rice genomics

IRIGIN Project : Contribution française au projet de séquençage massif de la GRiSP pour l'amélioration du riz dans les pays du Sud.

7,000 individus
17'755 fois le genome du riz
25 FlowCell Illumina soit 7'102 Gbases

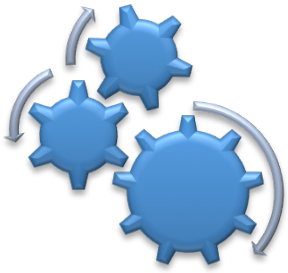
Problème de Gestion des BIG DATA
Des dizaines de Térabytes de Séquences

Equipe GDR produit et centralise les données
GDR (IRD) et ID (CIRAD) vont analyser les données

Types de données (SNPs, InDels, SV) **seront disponibles directement après séquençage pour la communauté**

From François Sabot

Les composants

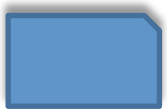


Workflows

Galaxy, Taverna, Open Alea

SGBD (locales et distantes)

Bases de données relationnelles (MySQL, PostGres)
NoSQL (MongoDB)

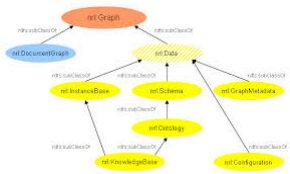


Formats semi-structurés

SAM, VCF, tableurs excel, etc.

Ressources distantes

Web Services, Linked Open Data

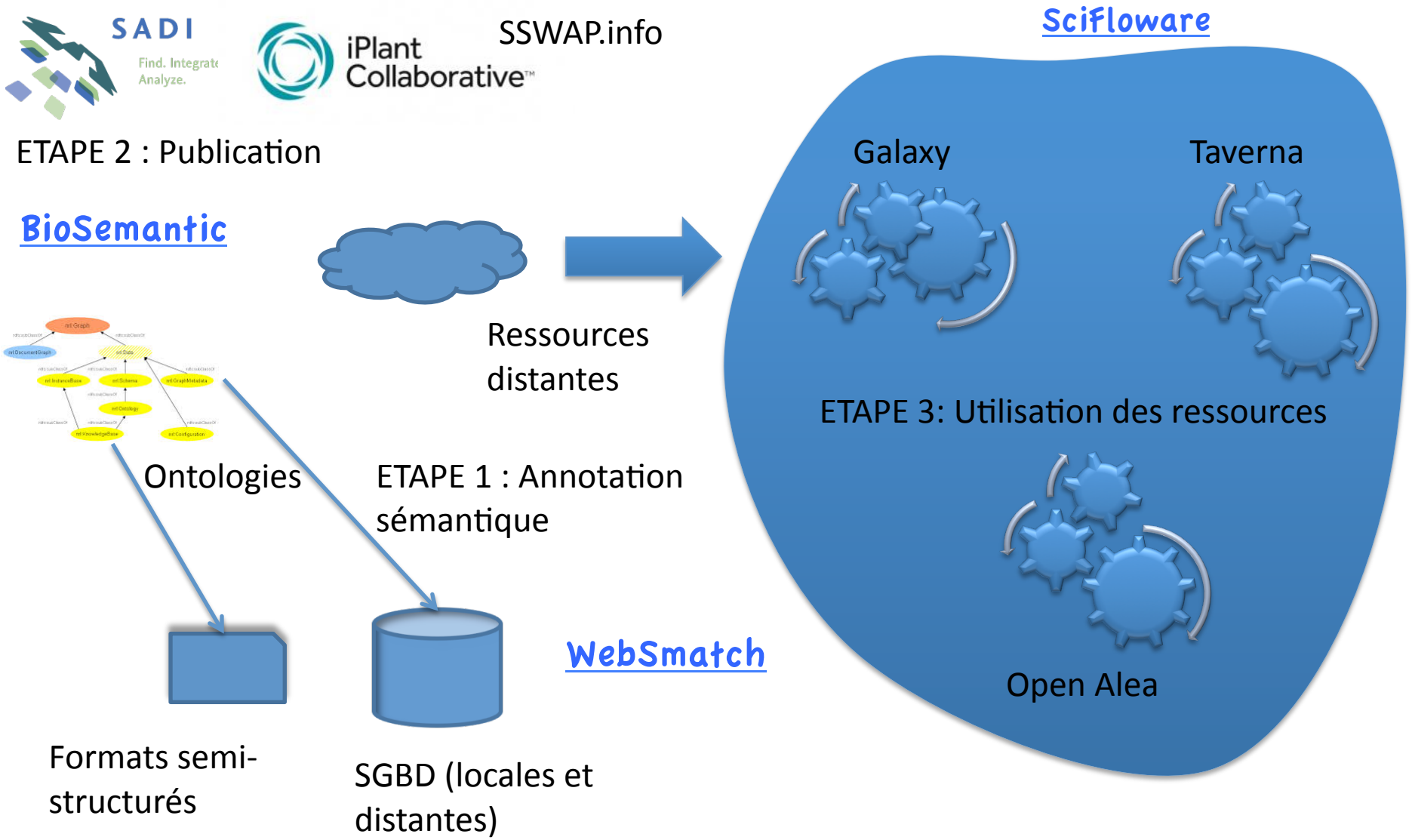


Ontologies, thésaurus

Bio-ontologies

From M. Ruiz

Méthodologies utilisées



ENRICHISSEMENT SÉMANTIQUE DE VUES RDF D2RQ DANS LE BUT D'AUTOMATISER L'INTÉGRATION DE BASES DE DONNÉES RELATIONNELLES DISTRIBUÉES

Julien WOLLBRETT
Pierre LARMANDE
Manuel RUIZ

Mise en correspondances entre base de données et ontologies

- Sujet vaste regroupant différentes méthodes, différents objectifs [1]

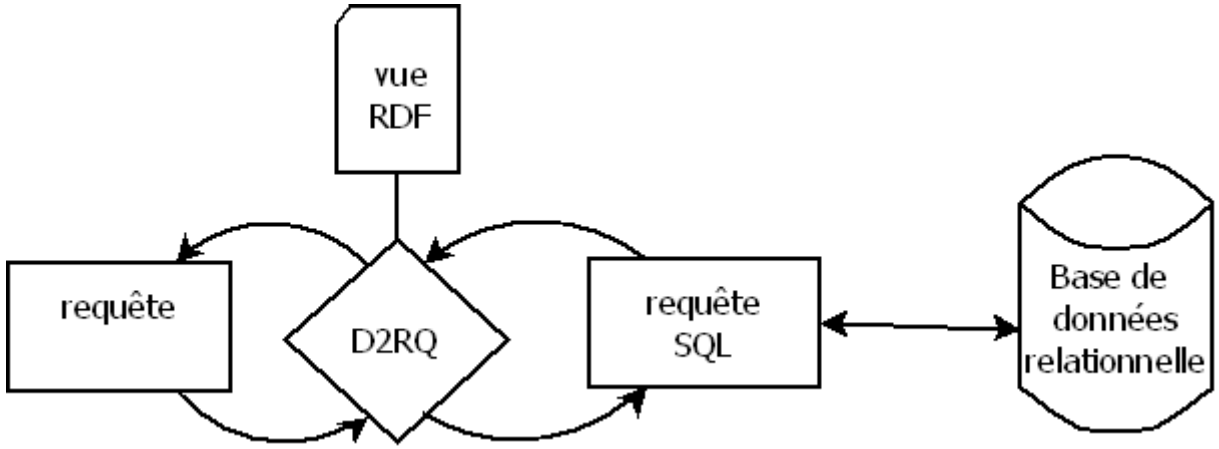
- OBDA (ontology based data access)
  - Utilisation d'une ontologie comme schéma global
  - Extraction d'une vue du schéma de base de données
  - Annotation de la vue à l'aide de termes ontologiques

[1] D.-E. Spanos, P. Stavrou, et N. Mitrou, « Bringing Relational Databases into the Semantic Web: A Survey », *Semantic Web*, vol. 3, n° 2, p. 169-209, 2012

From Julien Wollbrett

D2RQ

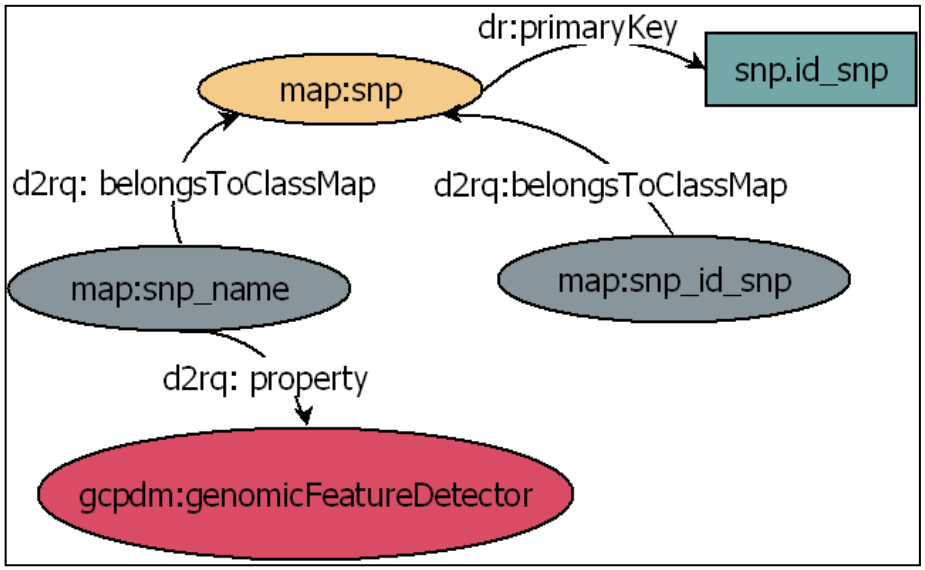
- Outil permettant la mise en correspondances des BDR et des ontologies
  - Création d'une vue RDF du schéma de la BDR
  - Annotation d'éléments de la vue RDF à l'aide de concepts ontologiques
  - Utilisation de la vue RDF pour interroger la BDR

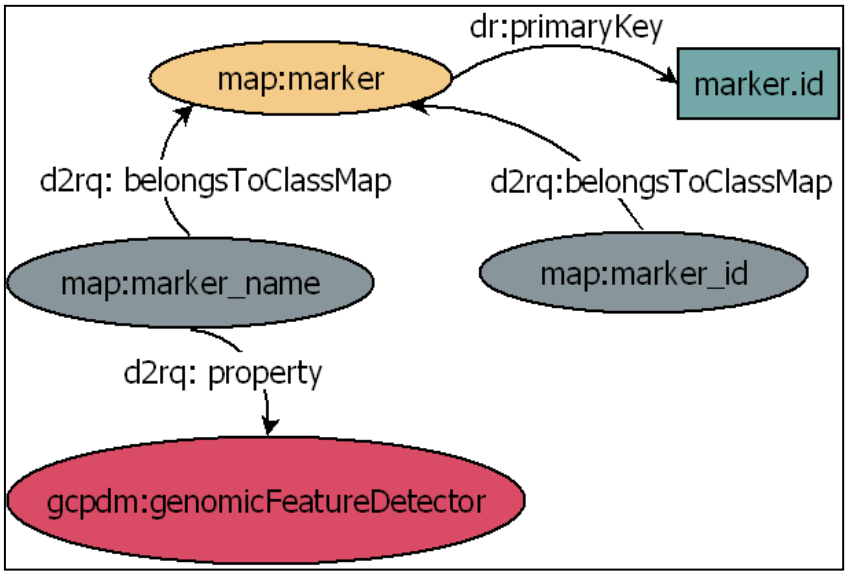


From Julien Wollbrett

Annotation sémantique de la vue RDF D2RQ

- Ajout d'une annotation sémantique à des éléments du schéma relationnel
- Permet la détection de concepts identiques dans des schémas relationnels hétérogènes
- Réalisée manuellement





snp(<u>id_snp</u>, name)

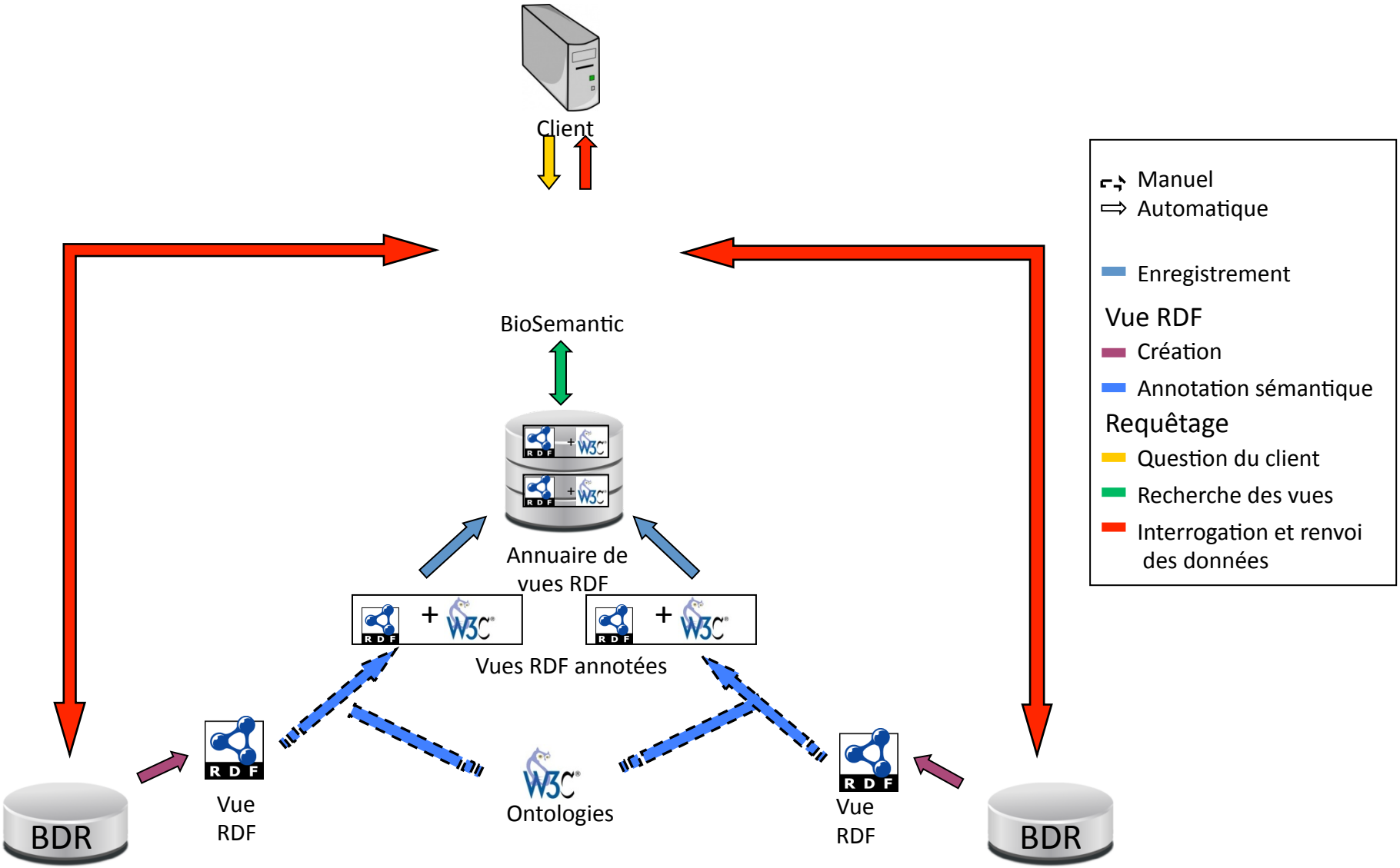
marker(<u>id</u>, name)

Plan

- Contexte
- **Approche proposée par BioSemantic**
- Enrichissement des vues D2RQ
- Résultats
- Conclusion/ perspectives

From Julien Wollbrett

Architecture de BioSemantic



Julien Wollbrett, Pierre Larmande, Frédéric de Lamotte and Manuel Ruiz, **Clever generation of rich SPARQL queries from annotated relational schema: application to Semantic Web Service creation for biological databases, BMC Bioinformatics, 2013**

Création automatique de requêtes

- Le client sélectionne un terme ontologique d'entrée et un terme ontologique de sortie
  - Ex: trouver tous les **germplasms** utilisés dans une **étude** donnée
- Vues RDF des Bases de données relationnelles
- Parcours des vues RDF pour rechercher automatiquement le plus court chemin reliant l'entrée à la sortie

From Julien Wollbrett

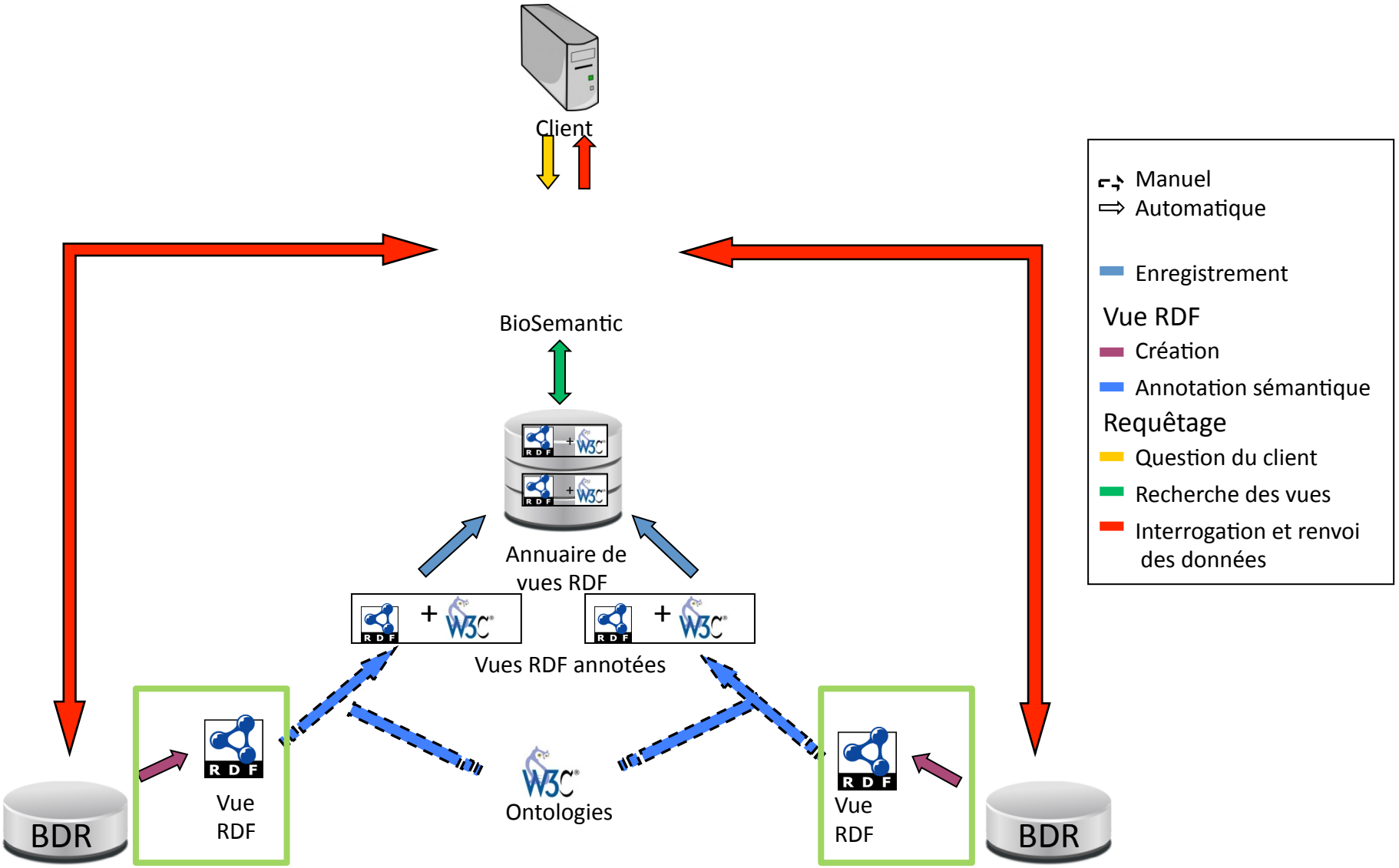
Limites de D2RQ pour notre approche

- Pas implémenté pour créer automatiquement des requêtes
- Pas suffisamment expressif pour notre utilisation
  - Présence de métadonnées permettant de transformer une requête SPARQL en requête SQL

➔ Nécessité d'enrichir le langage D2RQ

From Julien Wollbrett

Architecture de BioSemantic



Julien Wollbrett, Pierre Larmande, Frédéric de Lamotte and Manuel Ruiz, **Clever generation of rich SPARQL queries from annotated relational schema: application to Semantic Web Service creation for biological databases, BMC Bioinformatics, 2013**

Plan

- Contexte
- Approche proposée par BioSemantic
- **Enrichissement des vues D2RQ**
- Résultats
- Conclusion/ perspectives

From Julien Wollbrett

Contexte spécifique de BioSemantic

- Approche OBDA
- Recherche de plus court chemin dans un graphe
  - Création automatique de requête
  - Graphe représentant un schéma de base de données relationnelle

From Julien Wollbrett

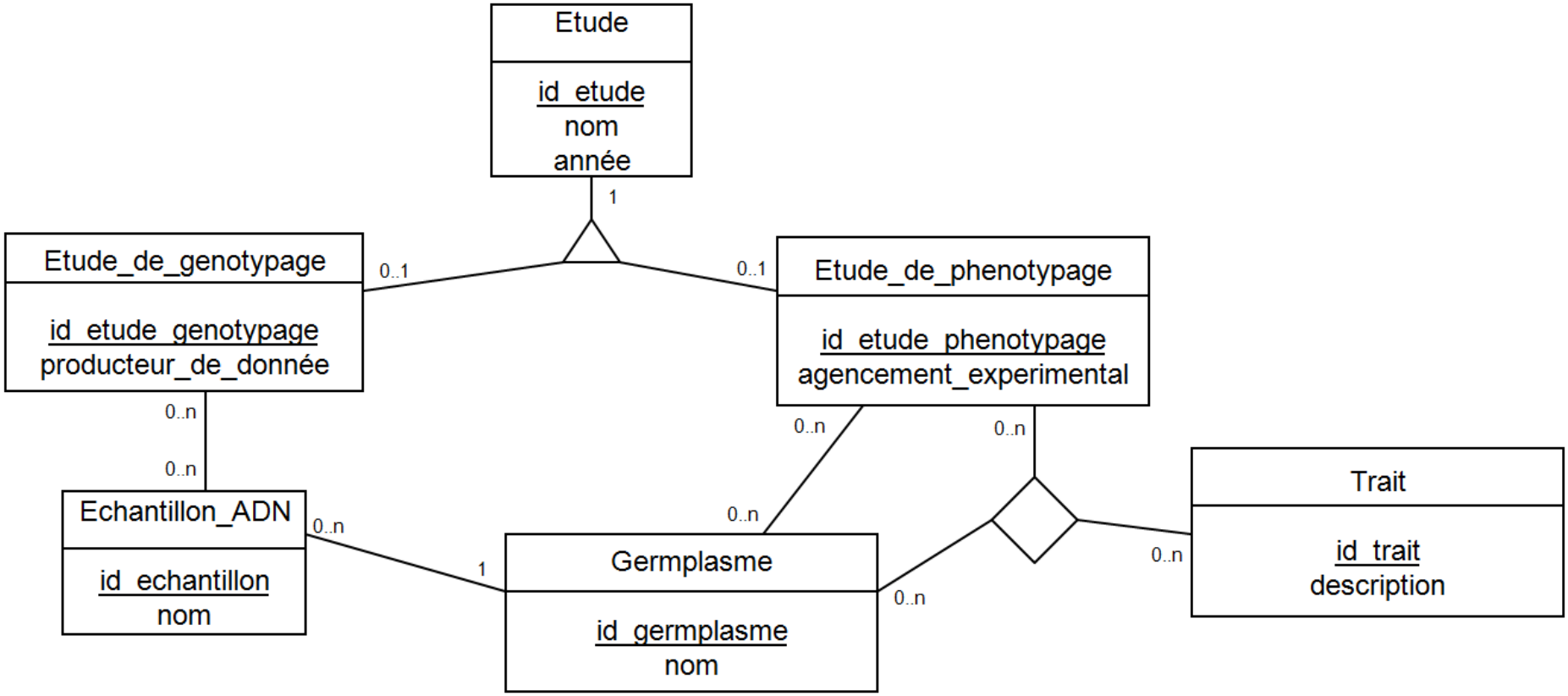
Plan

- Contexte
- Approche proposée par BioSemantic
- Enrichissement des vues D2RQ
  - **Combinaison de chemins**
  - Pondération des nœuds des chemins
- Résultats
- Conclusion/ perspectives

From Julien Wollbrett

Relations concernées par la combinaison de chemin

- Héritage, agrégation, composition



From Julien Wollbrett

Passage au modèle relationnel

- Problème de combinaison de chemins pour les relations d'héritages non applaties

etude(id_etude, nom, annee)
echantillon_ADN(id_echantillon, nom, #id_germplasme)
germplasme(id_germplasme, nom)
etude_de_genotypage (id_etude_genotypage, producteur_de_donnee, #id_etude)
etude_de_phenotypage(id_etude_phenotypage, agencement_experimental, #id_etude)
echantillon_genotypage (#id_echantillon, # id_etude_genotypage)
germplasme_phenotypage (# id_germplasme, # id_etude_phenotypage)

From Julien Wollbrett

Détection des relations

Subclass(r,s) <- Rel(r)^Rel(s)^PK(x,r)^FK(x,r,_,s)

avec

Rel(r)r est une relation

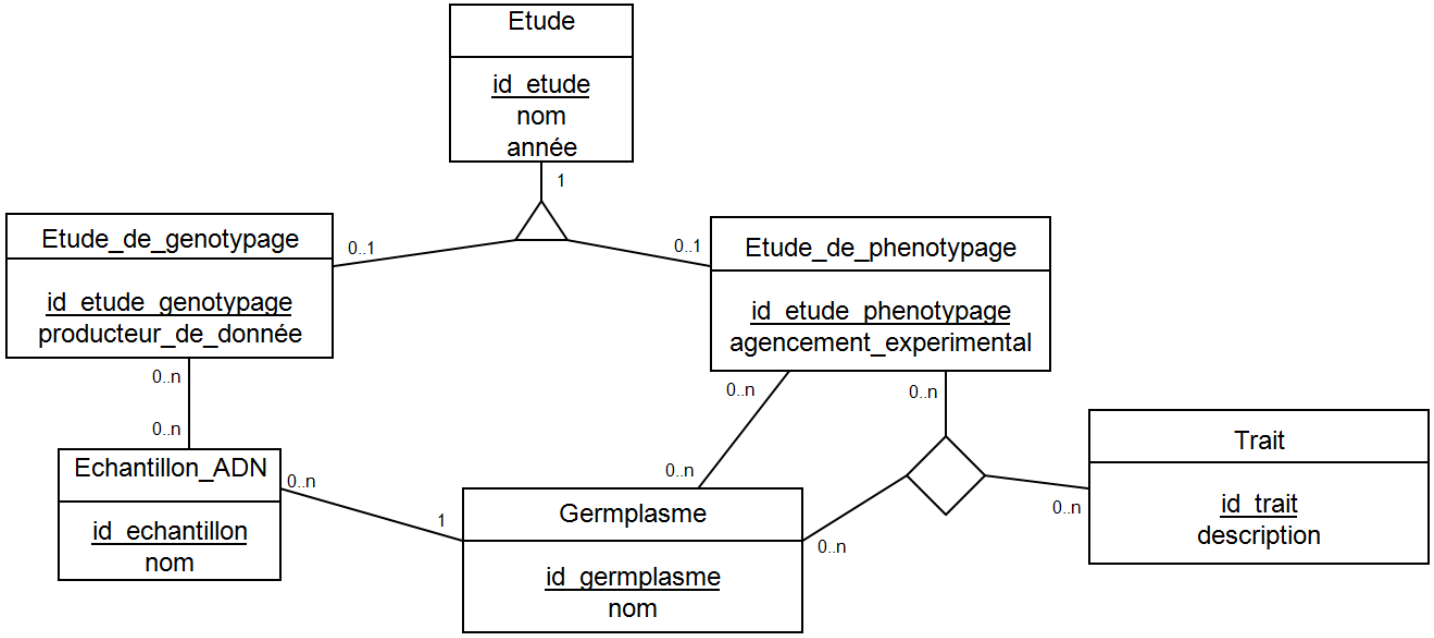
PK(x,r) x est la clé primaire de r

FK(x,r,y,s) x est la clé primaire de la relation r et référence y dans la relation s

From Julien Wollbrett

Enrichissement de la vue D2RQ

- Ajout de métadonnées dans la vue D2RQ
  - *Etude_de_genotypage rdfs:subClassOf Etude*
  - *Etude_de_phenotypage rdfs:subClassOf Etude*



From Julien Wollbrett

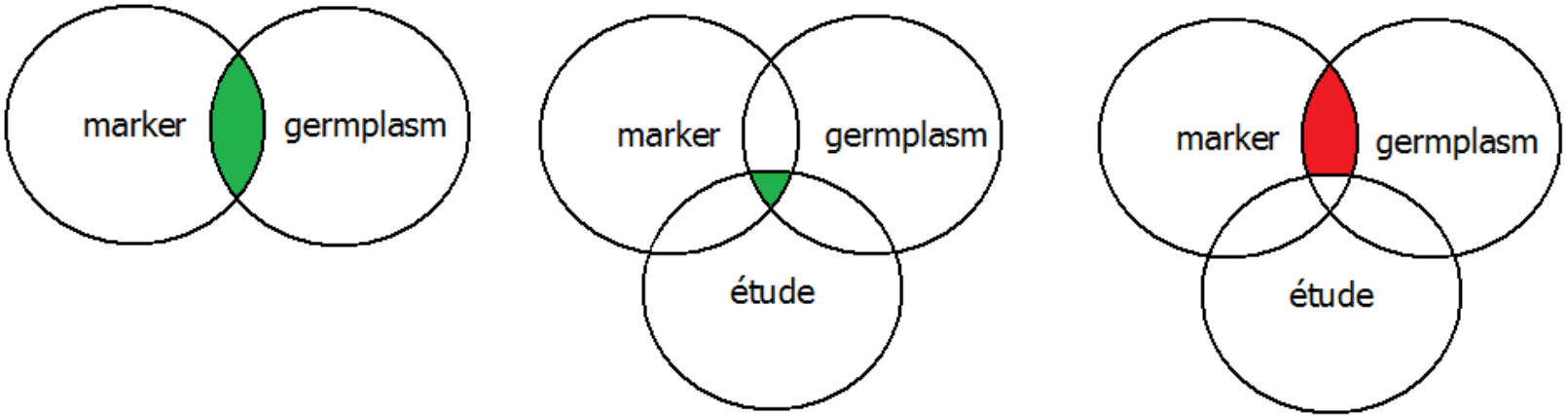
Plan

- Contexte
- Approche proposée par BioSemantic
- Enrichissement des vues D2RQ
  - Combinaison de chemins
  - **Pondération des nœuds des chemins**
- Résultats
- Conclusion/ perspectives

From Julien Wollbrett

Prise en compte de l'arité des tables d'association

- Arité d'une association: nombre d'entités reliées entre elles par une association



From Julien Wollbrett

Détection des tables d'association

Algorithme de détection:

pk= clé primaire de R

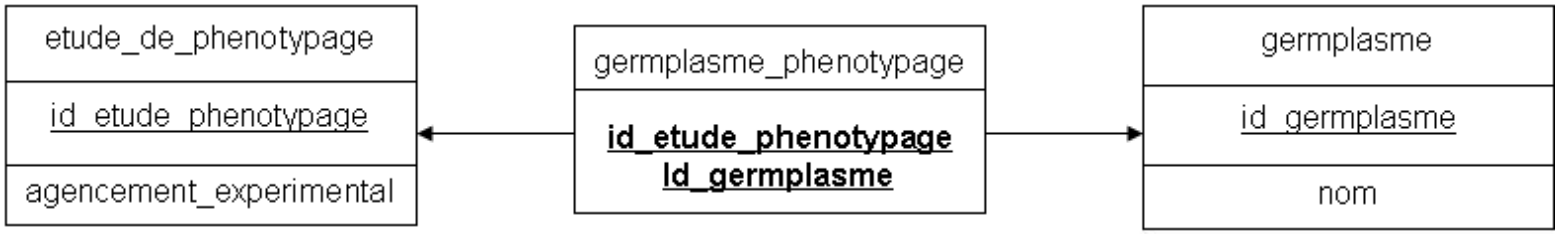
fk=clés étrangères de R

$if((\forall u \in R)(u \in fk => u \in pk))\{$

$if((\forall u \in R)(u \in pk => u \in fk))\{$

R est une table d'association

}}



From Julien Wollbrett

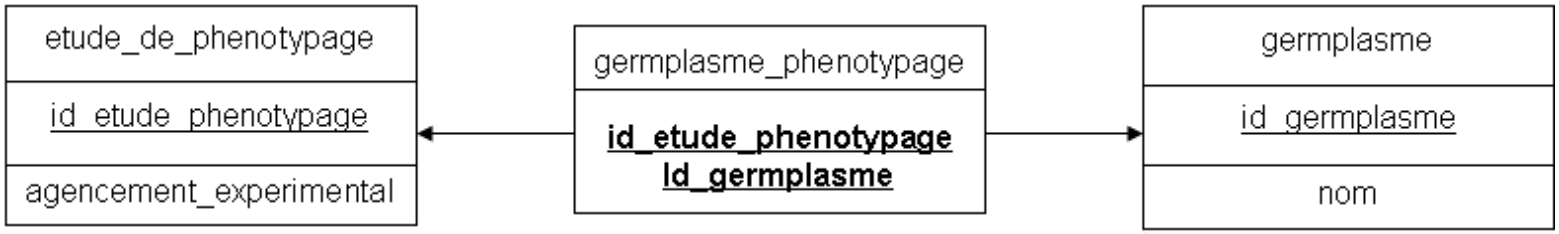
Enrichissement de la vue D2RQ

- Métadonnées ajoutées:

map:germplasme_phenotypage dr:associatedTo map:germplasme

map:germplasme_phenotypage dr:associatedTo map:etude_de_phenotypage

map:germplasme_phenotypage dr:arity "2"^^rdf:int



From Julien Wollbrett

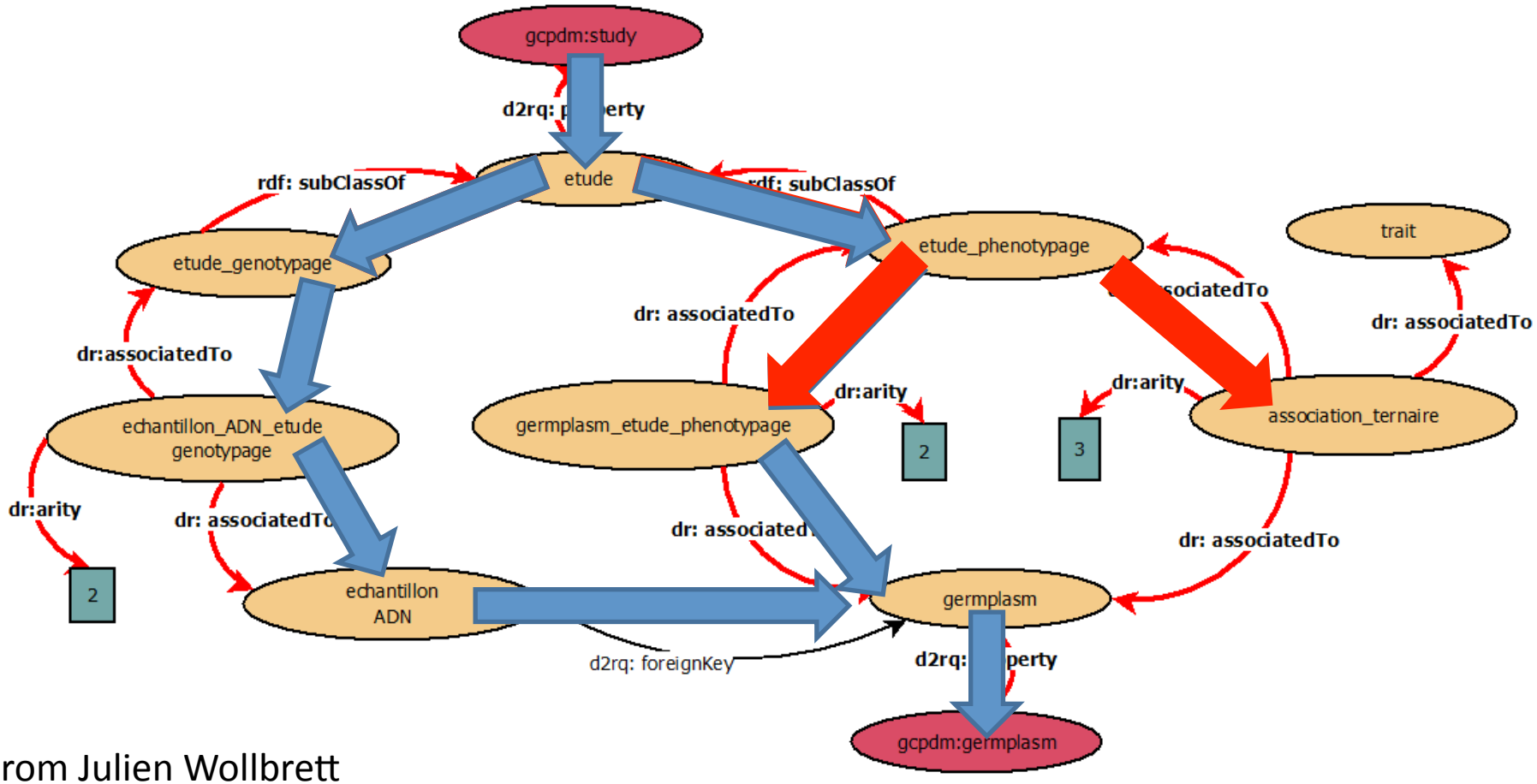
Plan

- Contexte
- Approche proposée par BioSemantic
- Enrichissement des vues D2RQ
- **Résultats**
- Conclusion/ perspectives

From Julien Wollbrett

Utilisation de l'enrichissement dans la recherche de plus court chemin

Requête souhaitée: trouver tous les germplasms utilisés dans une étude donnée



From Julien Wollbrett

Pertinence des requêtes

	héritage	Plusieurs tables d'association d'arité différentes	Dijkstra	BioSemantic	Requête SQL manuelle
Requête type 1	Oui	Non	1595	7212	7212
Requête type 2	Non	Oui	0	12302	12302
Requête type 3	Non	Oui	197	197	197
Requête type 4	Non	Non	2055	2055	2055

From Julien Wollbrett

Plan

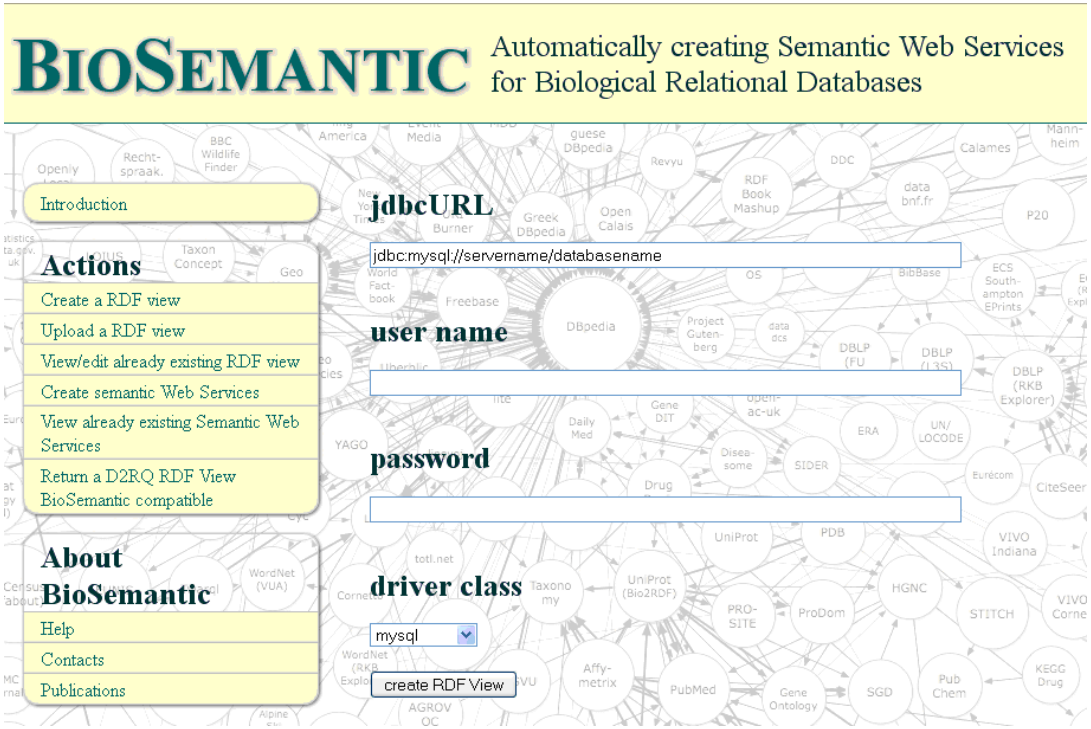
- Contexte
- Approche proposée par BioSemantic
- Enrichissement des vues D2RQ
- Résultats
- **Conclusion/ perspectives**

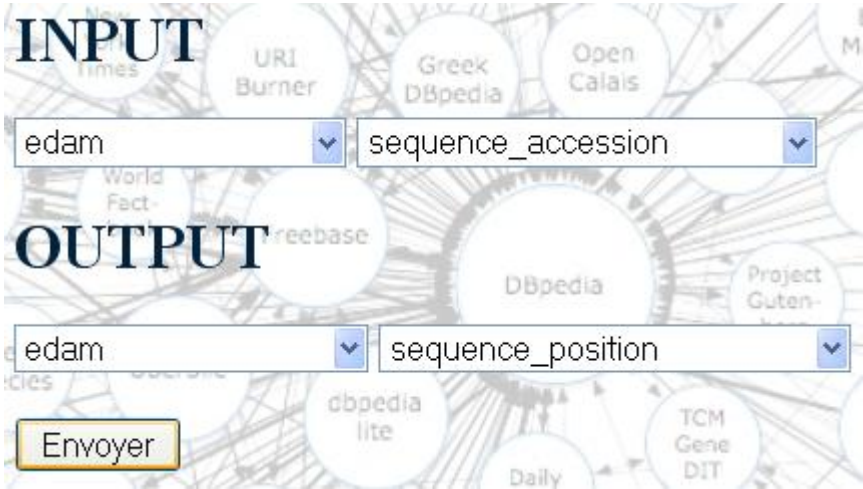
Conclusion

- Détournement de l'utilisation de D2RQ
- Ajout de métadonnées aux vues D2RQ
- Création automatique de requêtes
- Utilisation dans BioSemantic

http://southgreen.cirad.fr/?q=content/BioSemantic

From Julien Wollbrett





From Julien Wollbrett

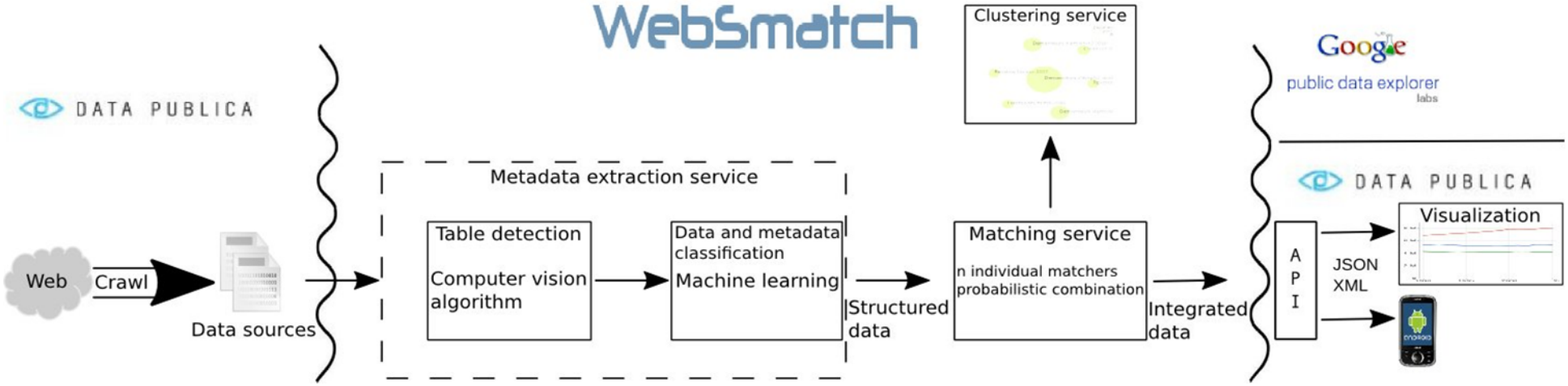
Travail dans le cadre d'IBC

- Prise en compte des instances lors de la création des vues D2RQ
  - Pondérer les tables en fonction du nombre d'instances
  - Automatiser le mapping entre ontologies et schéma de base de données relationnelles
  - Remonter au niveau du schéma des termes ontologiques fortement utilisés dans la base de données

From Julien Wollbrett

Automatiser le mapping entre ontologies et schéma de base de données relationnelles

Institut de biologie computationnelle



Patrick Valduriez

Mise en place de collaborations

Taverna

Crop Ontology