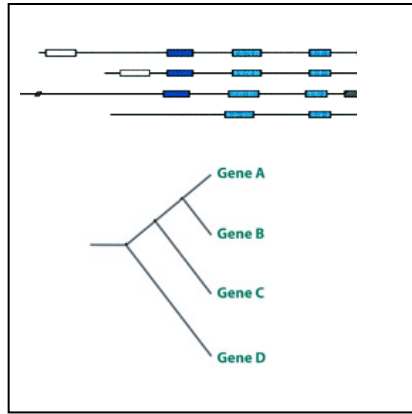


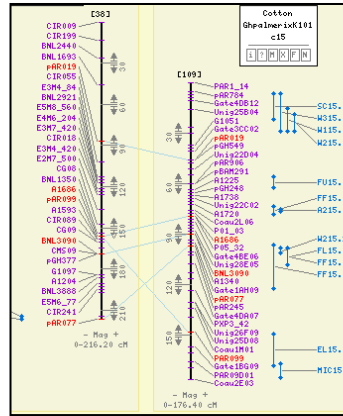
Ontology-based data integration projects at IBC

Data and resources

Contexte



Comparative genomics



Genetic and physical maps, QTL



Genetic resources



Markers

Genepiasm	mMacR01	mMacR03	mMacR07	mMacR08	mMacR13	mMacR150	mMacR164	mMac									
400	-99	-99	122	124	-99	-99	251	267	280	286	257	261	155	164	401	407	238
401	254	254	122	124	-99	-99	261	267	280	286	257	257	194	164	-99	-99	258
402	258	266	122	127	158	160	261	261	286	288	257	257	194	194	313	407	200
403	250	254	120	124	170	170	261	264	286	288	257	259	161	161	405	405	268
408	254	258	122	124	158	170	261	267	280	286	257	261	164	164	401	401	238
409	254	255	122	124	158	170	261	267	280	286	257	261	155	164	401	407	238
410	258	296	122	127	158	160	261	261	286	288	257	257	194	194	313	407	200
412	-99	-99	122	122	-99	-99	261	261	286	286	257	257	-99	-99	-99	-99	234
413	256	298	122	127	158	170	261	261	270	270	257	264	182	182	313	313	238
Missing Data	22.2%	22.2%	0.0%	0.0%	33.3%	33.3%	0.0%	0.0%	0.0%	0.0%	0.0%	11.1%	11.1%	22.2%	22.2%	0.0%	
Monomorphic	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	

Genotyping studies

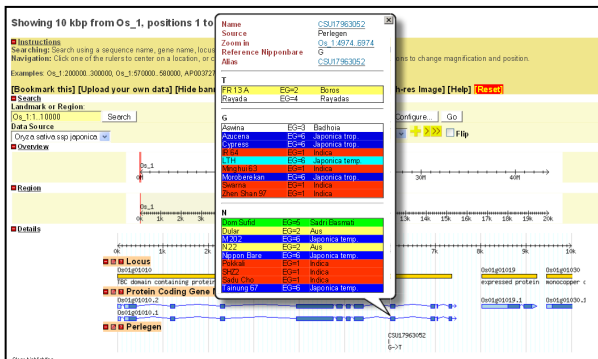
Individuals

Phenotypes



Geographic data

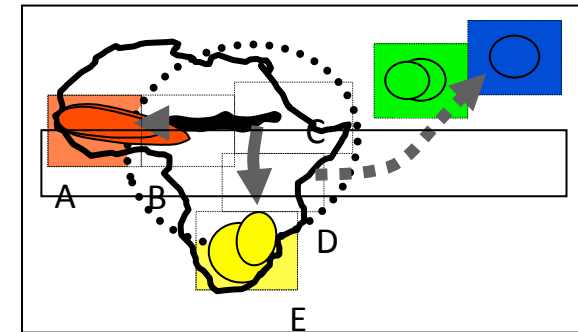
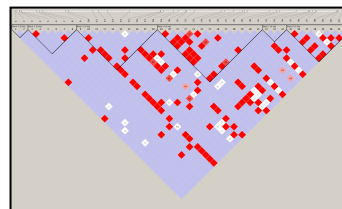
Genomic annotations

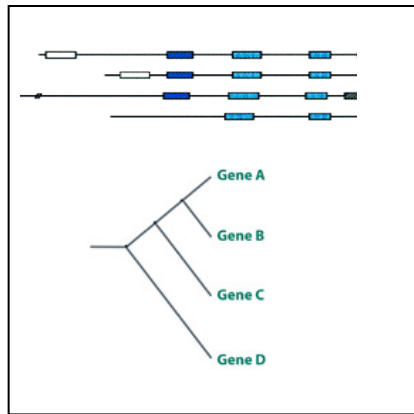


From M. Ruiz

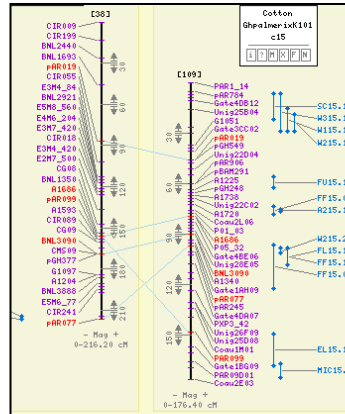
Functional annotations

Analysis Workflows





Comparative genomics

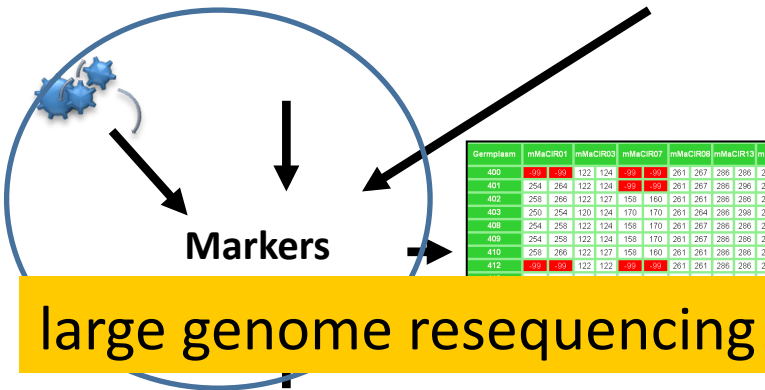


Genetic and physical maps, QTL



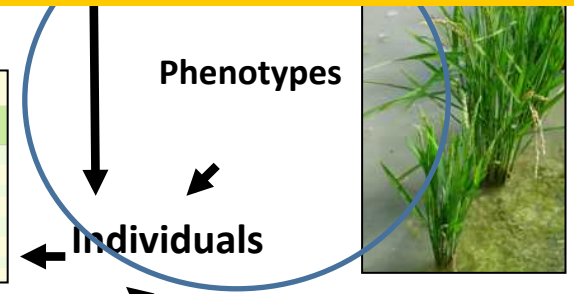
Genetic resources

high-throughput phenotyping



Genoplasm	mMacR01	mMacR03	mMacR07	mMacR08	mMacR13	mMacR150	mMacR152	mMacR154	mMac
400	-0.9	-0.9	1.22	1.24	-0.9	-0.9	2.51	2.67	2.86
401	2.54	2.54	1.22	1.24	-0.9	-0.9	2.51	2.67	2.86
402	2.58	2.68	1.22	1.27	1.58	1.60	2.61	2.61	2.86
403	2.50	2.54	1.20	1.24	1.70	1.70	2.61	2.64	2.86
408	2.54	2.58	1.22	1.24	1.58	1.70	2.61	2.67	2.86
409	2.54	2.55	1.22	1.24	1.58	1.70	2.61	2.67	2.86
410	2.58	2.96	1.22	1.27	1.58	1.60	2.61	2.61	2.86
412	-0.9	-0.9	1.22	1.22	-0.9	-0.9	2.51	2.67	2.86

Genotyping studies

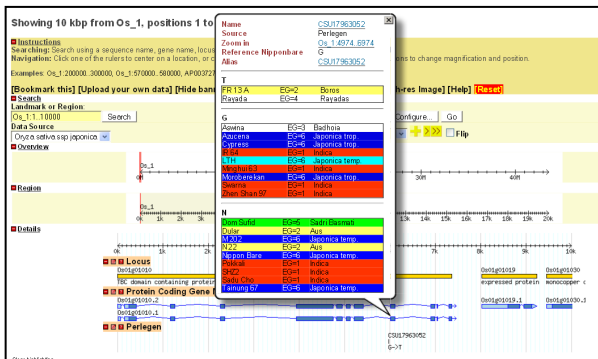


Phenotypes

Individuals

Geographic data

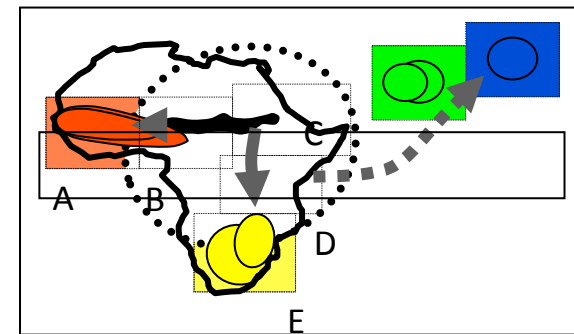
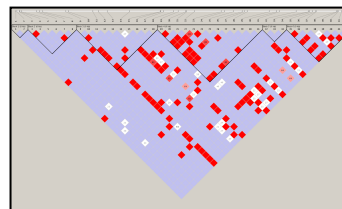
Genomic annotations



From M. Ruiz

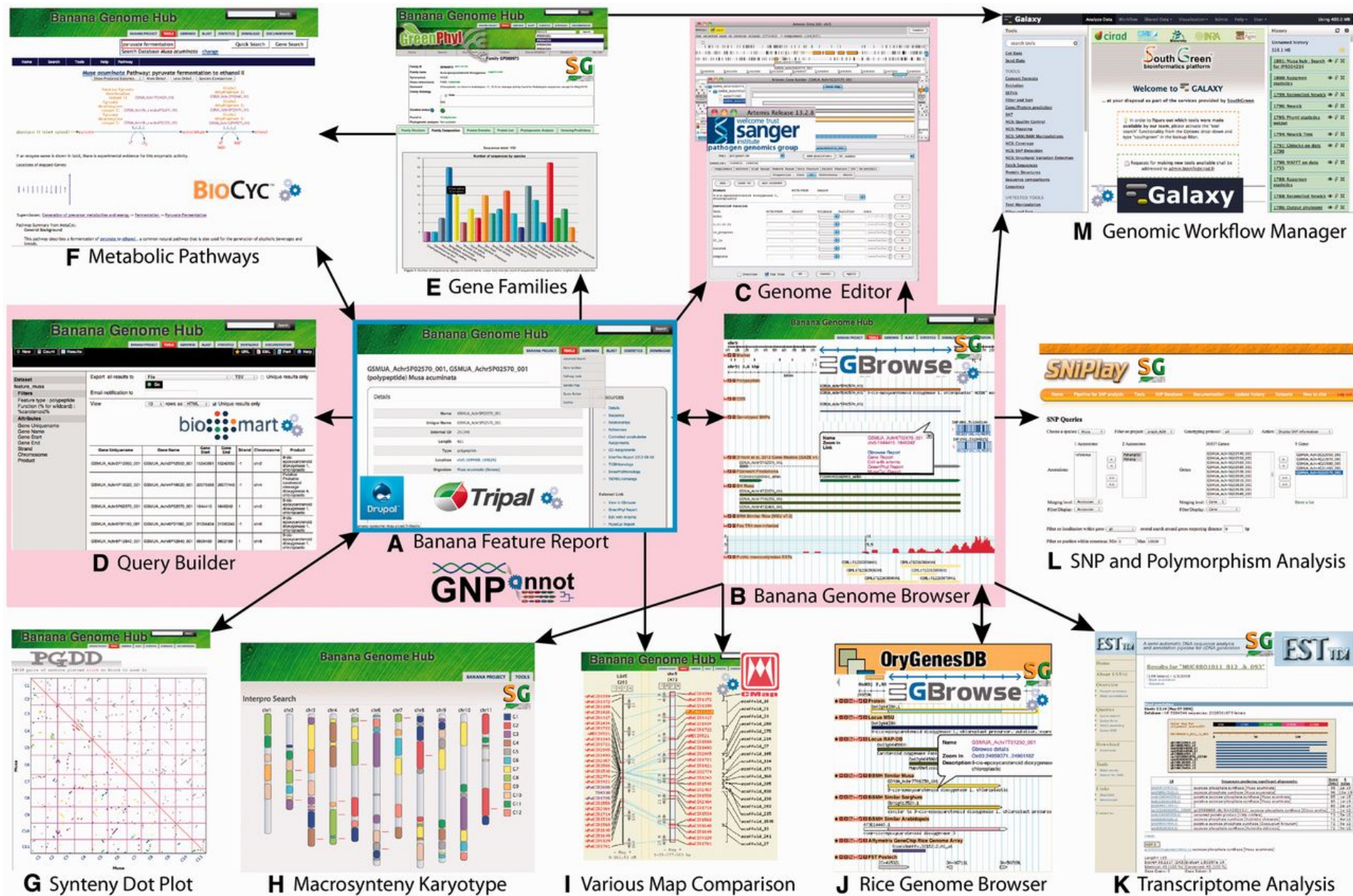
Functional annotations

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 workflows**
 - ESTtik
 - SNIPlay
 - Galaxy

The banana genome hub



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

Banana Genome Hub

[HOME](#) | [GENOME DETAILS](#) | [TOOLS](#) | [GBROWSE](#) | [BLAST](#) | [DOWNLOAD](#) | [DOCUMENTATION](#)

File - Help -

Banana: 3.061 kbp from chr1:9,269,057..9,272,117

[Browser](#) | [Select Tracks](#) | [Snapshots](#) | [Custom Tracks](#) | [Preferences](#)

Search
 Landmark or Region :

Examples : [GSMUA_Achr1P12150_001](#), [GSMUA_Achr5P02570_001](#).
 Data Source :

Scroll/Zoom: << < > >> Show 3.061 kbp

Overview

Region

Details

chr1: 3.061 kbp
 1 kbp |-----|
 9270k |-----| 9271k |-----| 9272k

- ★ Non coding tRNAs
- ★ Protein Coding Gene Model (Manually curated Old version)
GSMUA_Achr1T12150_001
- ★ Polypeptide (Manually curated Old version)
GSMUA_Achr1P12150_001
GSMUA_Achr1T12150_001" TAK14, putative, expressed" At5g39020" complete
- ★ Non coding miRNA
- ★ Polypeptide
GSMUA_Achr1P12150_001
GSMUA_Achr1T12150_001" TAK14, putative, expressed" At5g39020" complete
- ★ Protein Coding Gene Model

- 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

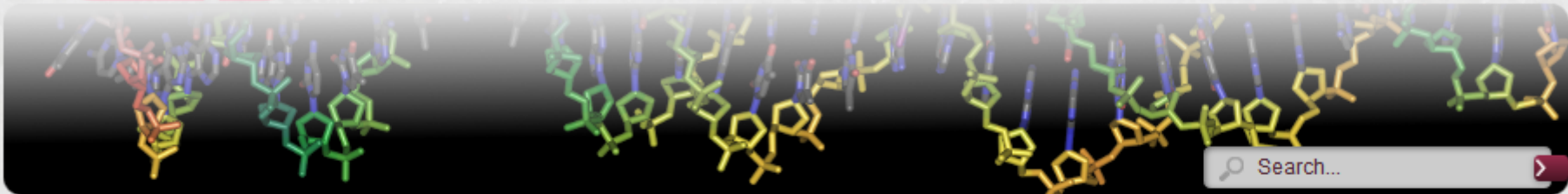
The Computational Biology Institute (IBC)





Computational Biology Institute

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



Home

Workpackages

Labs & Partners

Publications

Platforms

Open Positions

Events

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](#)

[Subscribe 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.

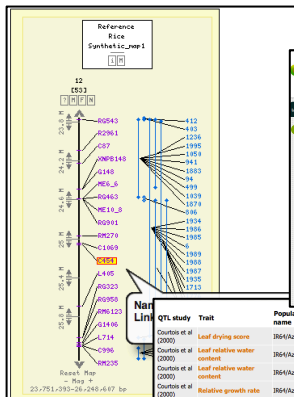
Nous allons lier les données « multi-échelle » : environnements, phénotypes, ressources génétiques, données de génotypage, QTL et séquence génomique.

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

The Rice South Green Hub

OryGenesDB

OryGenesDB interface showing search filters and genomic tracks. The interface includes a search bar with the keyword "WRKY", filter options for flanking sequence tags, regions, orientations, and outputs. A list of genes is displayed with their genomic coordinates and associated features.



TropGeneDB interface showing QTL analysis results. The interface displays a list of QTLs with their names, traits, and associated genomic regions.

QTL study	Trait	Population name	Population type	Population size	QTL Conditions	Phenotypic R ² (%)	Sampling method	Map	Chromosome	Left locus	Right locus	QTL start position (bp)	QTL stop position (bp)
Quilan et al (2000)	Leaf drying score	IR64/Azusa	DH	56	Field, aerobic; aerobic conditions; short and severe water stress starting 39 days; dry season	6.00	Per plot	SYNTHETIC_MAP12	SDH1	RG463	19620461	24366292	
Quilan et al (2000)	Leaf relative water content	IR64/Azusa	DH	56	Field, aerobic; aerobic conditions; short and severe water stress starting 39 days; dry season	18.50	31 dabs	SYNTHETIC_MAP12	RG274	RG216	15948335	24366992	
Quilan et al (2000)	Leaf relative water content	IR64/Azusa	DH	56	Field, aerobic; aerobic conditions; short and severe water stress starting 39 days; dry season	7.80	31 dabs	SYNTHETIC_MAP12	SDH1	RG463	19620461	24366292	
Quilan et al (2000)	Relative growth rate	IR64/Azusa	DH	56	Field, aerobic; aerobic conditions; well watered treatment; dry season	10.70	10 plants; 24 d	SYNTHETIC_MAP12	AF6	RG467	17988580	19626925	
Quilan et al (2003)	Maximum root length	IAC65G39	RLS F7	125	Greenhouse; PVC pipes; aerobic conditions; well watered treatment	14.00	1 plant; 45 d	SYNTHETIC_MAP12	RG381	RG659	2488270	27486769	
Quilan et al (2003)	Root dry weight in the 0-30cm layer	IAC65G39	RLS F7	125	Greenhouse; PVC pipes; aerobic conditions; well watered treatment	12.20	1 plant; 45 d	SYNTHETIC_MAP12	RG301	RH235	24599302	26109094	

TropGeneDB

GreenPhyIDB

GreenPhyIDB interface showing gene ontology and phylogenetic analysis. The interface displays a list of genes with their ontologies and a bar chart showing the number of sequences by species.

Oriza Tag Line interface showing mutant information. The interface displays a table with columns for Trait, Phenotypic class, and Organ, and a section for Available observations.

Trait	Referenced or designated mutants	Generation/Stage
Heading date	delayed flowering (abbr. del)	T1 adults
panicle		T0 mature plant
adults		T0 mature plant
Late flowering; with or without erect and dark green leaf and/or flag leaf remain erect.		T0 mature plant
Late flowering (1 plant), (Photos 1, 2).		T0 mature plant
1 mutant(s) over 7 plants => %		T0 mature plant

Oriza Tag Line interface showing passport data. The interface displays a table with columns for Line ID, Ontology ID, and About, and a section for Available observations.

Phenotype class	Trait	Referenced or designated mutants	Generation/Stage	Select
Phenology	Heading date	delayed flowering	T1 adults	<input checked="" type="checkbox"/>
Reporter gene	Organ/Tissue	Expression level	Generation/Stage	Select
GFP	flower: stamens	strong	T0 mature plant	<input type="checkbox"/>
GFP	flower: stamens	strong	T0 mature plant	<input type="checkbox"/>
GFP	flower: guard cells	strong	T0 mature plant	<input type="checkbox"/>

Oriza Tag Line

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

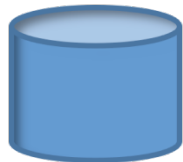
**Types de données (SNPs, InDels, SV) seront disponibles
directement après séquençage pour la communauté**

Les composants



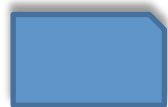
Workflows

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



SSWAP.info

SciFloware

ETAPE 2 : Publication

BioSemantic



Ontologies

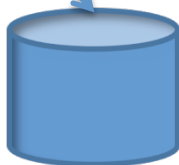


Formats semi-structurés



Ressources distantes

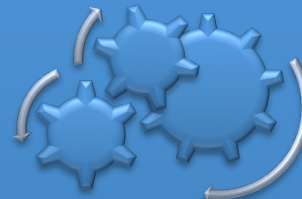
ETAPE 1 : Annotation sémantique



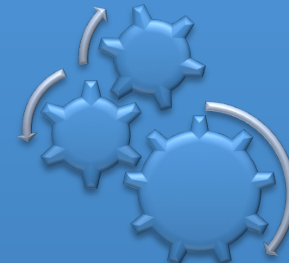
SGBD (locales et distantes)

WebSmatch

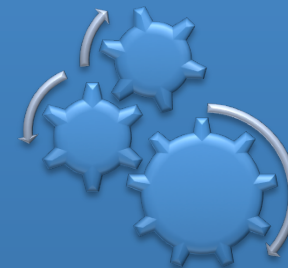
Galaxy



Taverna



ETAPE 3: Utilisation des ressources



Open Alea

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



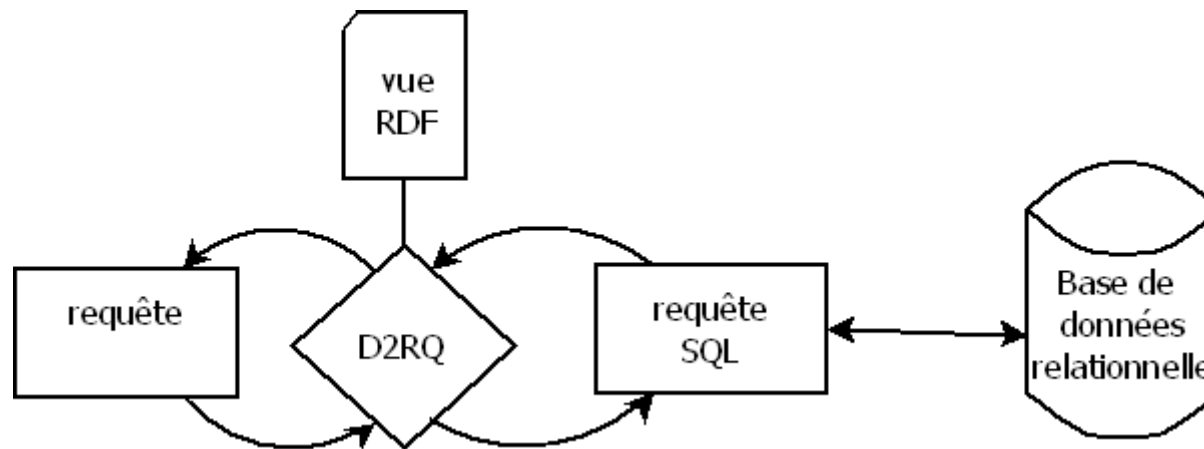
Mise en correspondances entre base de données et ontologies

- 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

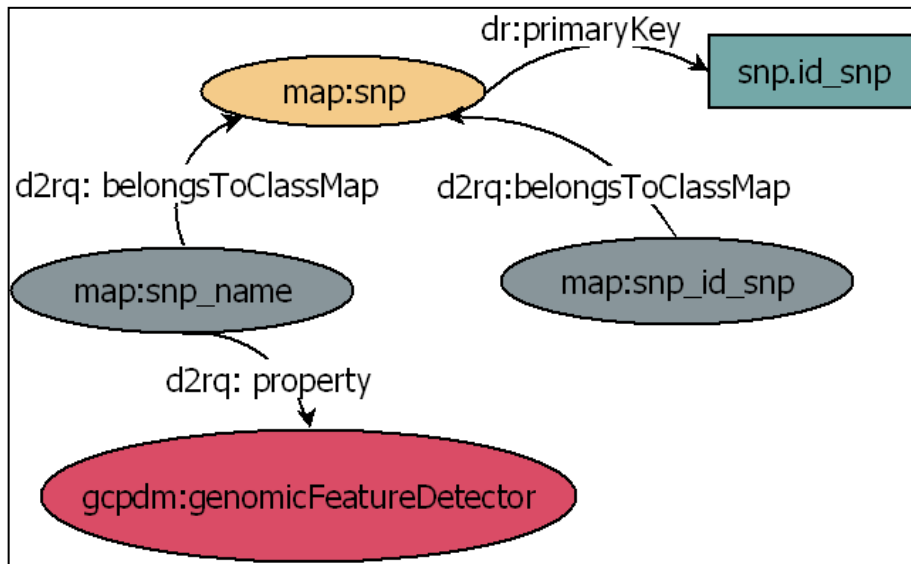
D2RQ

- 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

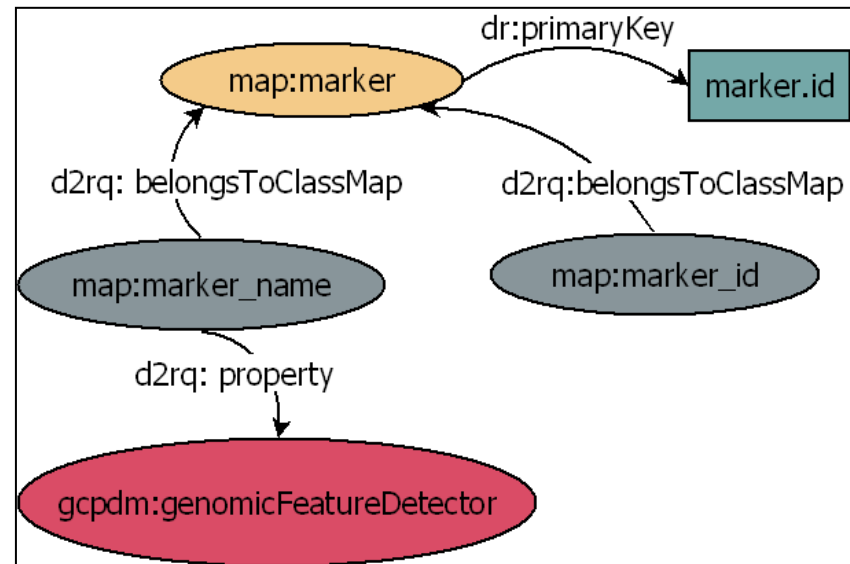


Annotation sémantique de la vue RDF D2RQ

- 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(id_snp, name)`

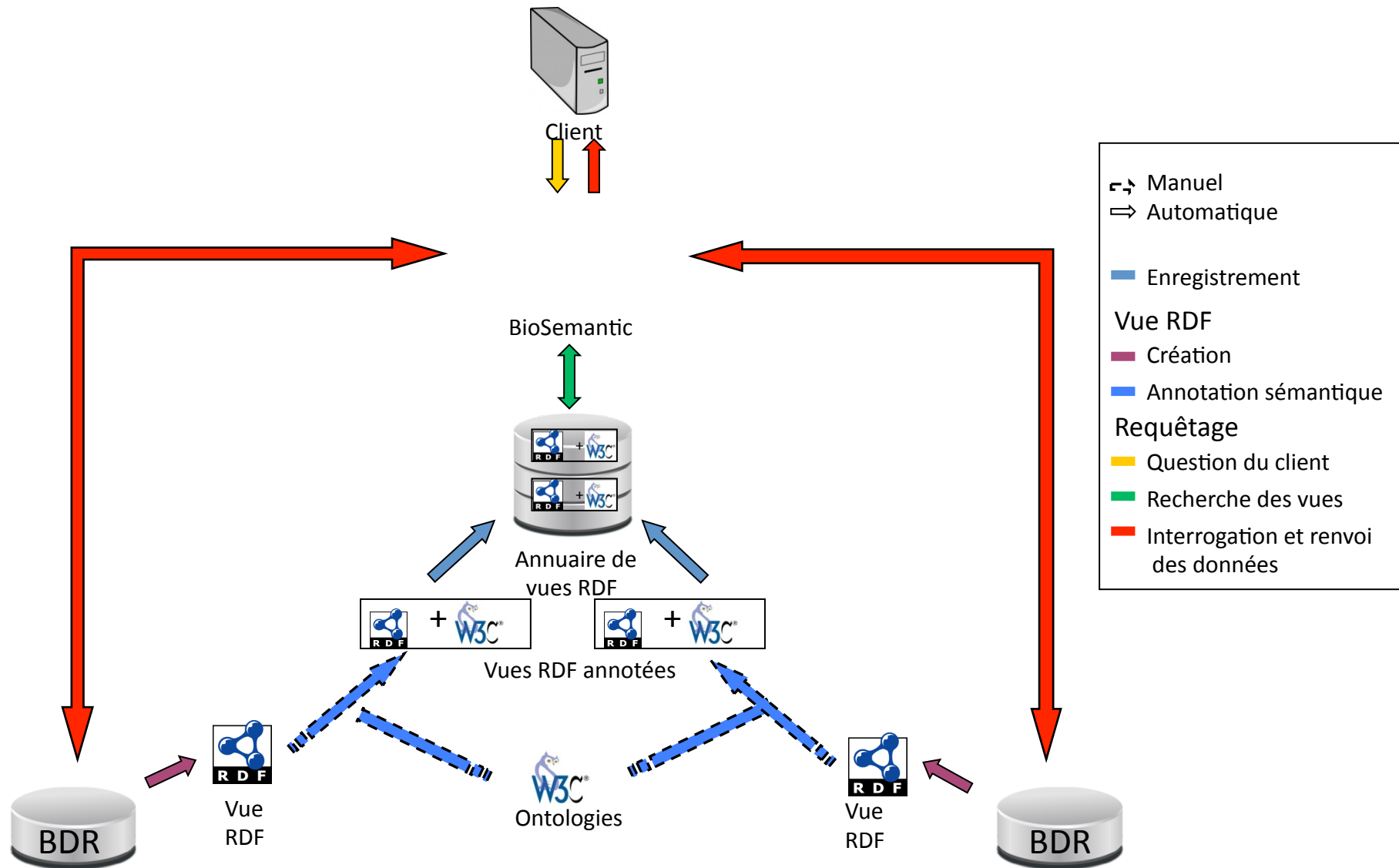


`marker(id, name)`

Plan

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

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

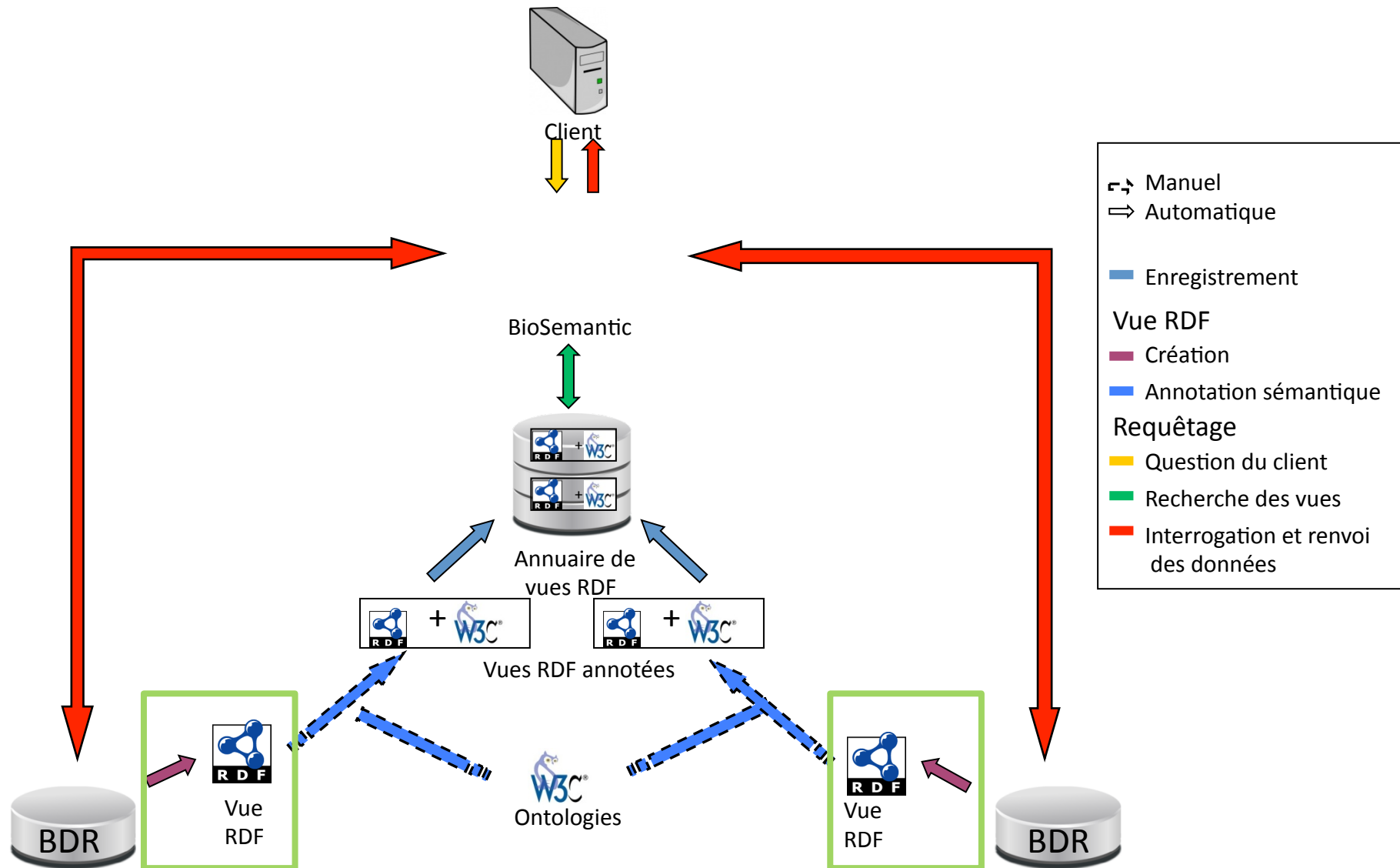
- Le client sélectionne un terme ontologique d'entrée et un terme ontologique de sortie
 - Ex: trouver tous les **germplasms** utilisés dans une **étude** donnée
- Vues RDF des Bases de données relationnelles
- Parcours des vues RDF pour rechercher automatiquement le plus court chemin reliant l'entrée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

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

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

Contexte spécifique de BioSemantic

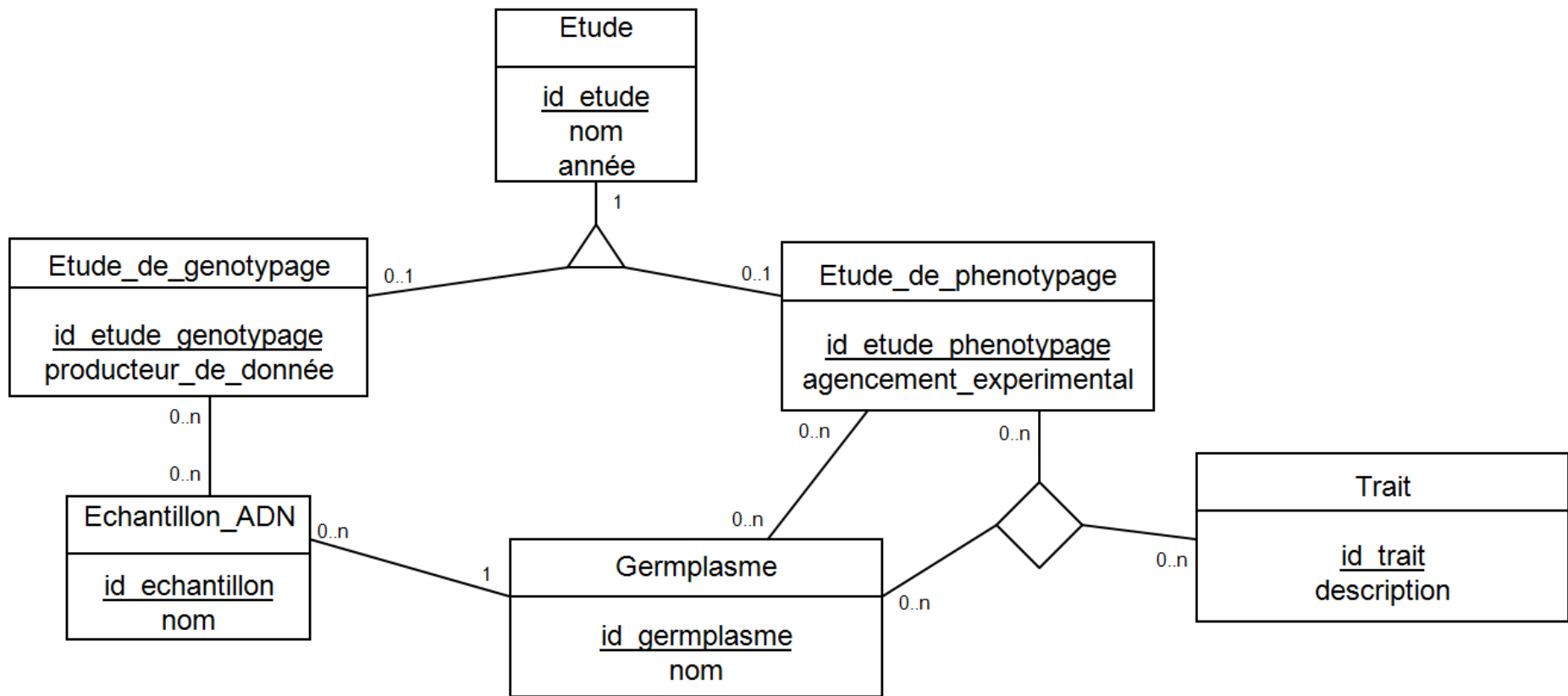
- Approche OBDA
- Recherche de plus court chemin dans un graphe
 - Création automatique de requête
 - Graphe représentant un schéma de base de données relationnelle

Plan

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

Relations concernées par la combinaison de chemin

- Héritage, agrégation, composition



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)

Détection des relations

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

avec

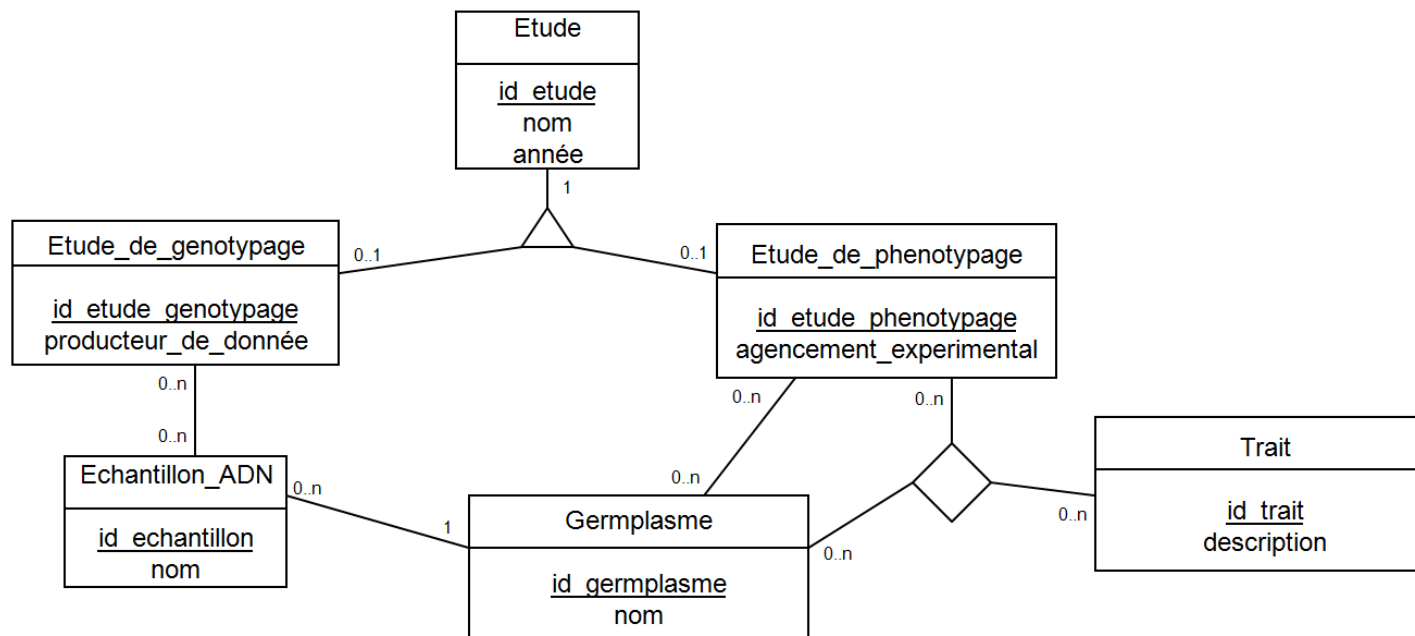
Rel(r) r est une relation

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

Enrichissement de la vue D2RQ

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

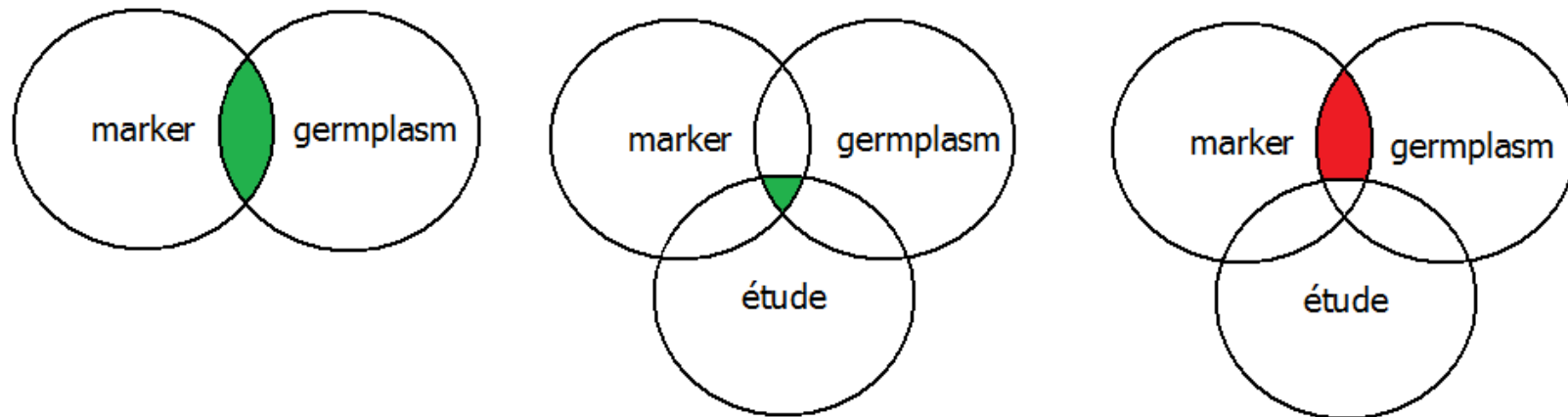


Plan

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

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

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



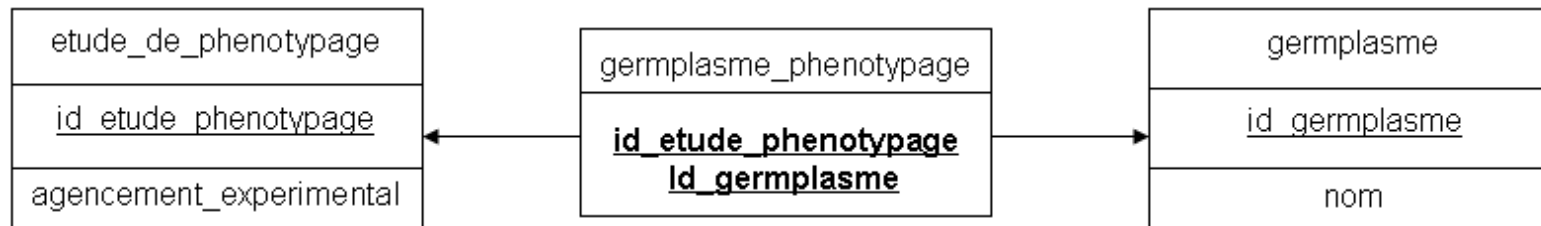
Détection des tables d'association

Algorithme de détection:

pk = clé primaire de *R*

fk = clés étrangères de *R*

$if ((\forall u \in R)(u \in f k \Rightarrow u \in pk)) \{$
 $if ((\forall u \in R)(u \in pk \Rightarrow u \in f k)) \{$
R est une table d'association
 $\}$
 $\}$



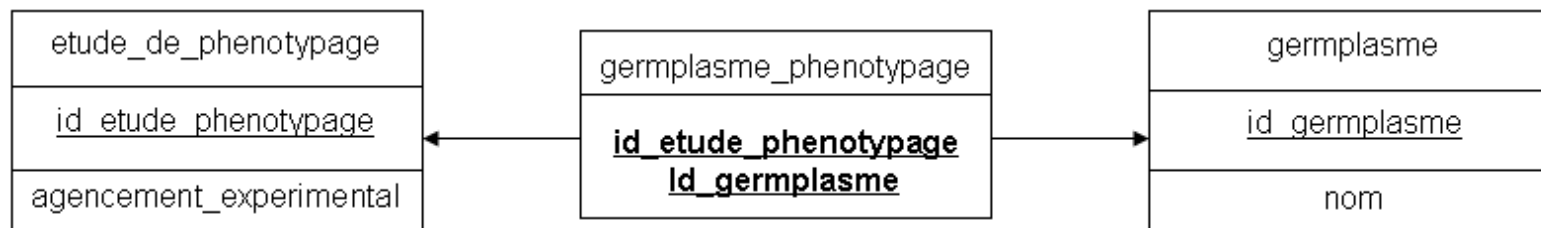
Enrichissement de la vue D2RQ

- 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

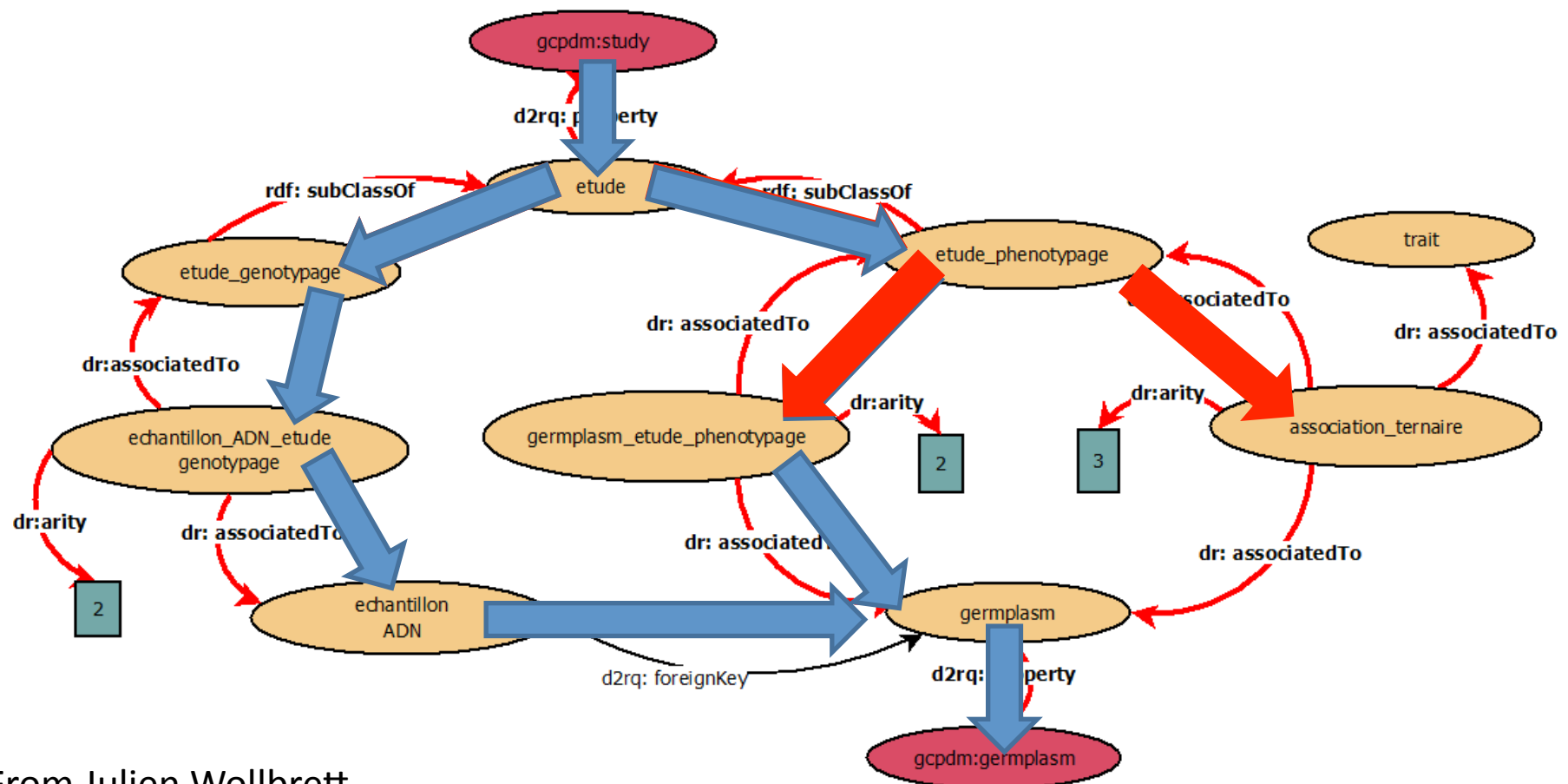


Plan

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

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

Requête souhaitée: trouver tous les germplasm 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

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>

BIOSEMANTIC

Automatically creating Semantic Web Services for Biological Relational Databases

Introduction

Actions

- Create a RDF view
- Upload a RDF view
- View/edit already existing RDF view
- Create semantic Web Services
- View already existing Semantic Web Services
- Return a D2RQ RDF View BioSemantic compatible

About BioSemantic

- Help
- Contacts
- Publications

jdbcURL
jdbc:mysql://servername/databasename

user name
DBpedia

password

driver class
mysql

create RDF View

INPUT

edam sequence_accession

OUTPUT

edam sequence_position

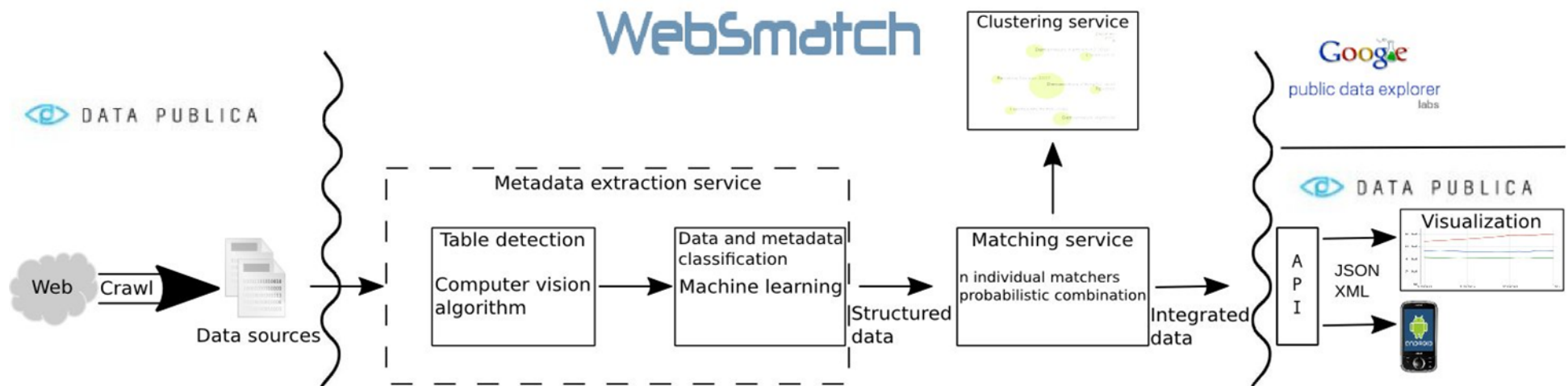
Envoyer

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

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



Mise en place de collaborations



Crop Ontology