DataRiver

open source data integration

RETE ALTA TECNOLOGIA
EMILIA-ROMAGNA
HIGH TECHNOLOGY NETWORK

RegioneEmiliaRomagna

ASTER

Spin-Off dell'Università degli studi Modena e Reggio Emilia fondata nel Giugno 2009:

DBGroup
DATABASE GROUP

professori e ricercatori del gruppo di ricerca **DBGroup** (www.dbgroup.unimo.it)

QUIX

professionisti di **QUIX s.r.l.** (www.quix.it) con esperienza pluriennale nel campo del ICT, attiva nel campo dell'Information Technology dal 1990

DataRiver Team

Il team di **DataRiver** è composto da professori, ricercatori e professionisti nel campo del ICT.
Il team di impresa è in grado di integrare:

- **componente scientifica** competenze riconosciute a livello internazionale nei campi della Data Integration, Semantic Web e Business Intelligence

- **componente imprenditoriale** con esperienza decennale nel settore ICT, forti competenze tecniche e grandissima conoscenza del mercato sia privato che della pubblica amministrazione

Soci:
Mirko Orsini, Ph.D.
Prof. Sonia Bergamaschi
Prof. Domenico Beneventano
Università degli studi di Modena e Reggio Emilia (Prof. Luigi Rovati)
Alberto Corni, Ph.D.
Laura Po, Ph.D.
Serena Sorrentino, Ph.D.

DataRiver open source data integration

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

DataRiver Team

I ruoli funzionali all'interno dell'impresa sono i seguenti:

- **Mirko Orsini**: Presidente e Direttore esecutivo, responsabile dei Servizi di consulenza, dello Sviluppo Software e delle Risorse umane
- **Sonia Bergamaschi**: Vice Presidente e responsabile del Comitato Scientifico
- **Domenico Beneventano**: Consigliere di amministrazione e membro del Comitato Scientifico, responsabile del Sistema Qualità
- **Silvano Pancaldi** (Injenia Srl): responsabile dell'Area Commerciale e Marketing.
- **Laura Po**: membro del Comitato Scientifico e servizi di consulenza
- **Serena Sorrentino**: membro del Comitato Scientifico servizi di consulenza
- **Alberto Corni**: responsabile Architettura Hardware e Sofware e sviluppo software
- **Entela Kazazi**: responsabile dell'Unità Operativa Data Integration
- **Enrico Calanchi**: responsabile dell'Unità Operativa Clinical Data Management

DataRiver open source data integration

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Prodotti e servizi

→ **Data Integration**, **Semantic Web** e **Business Intelligence**:

MoMIS

the Olive Tree Project

→ **Clinical Data Management**: Fondazione Italiana Linfomi, Centro Oncologico Modenese, Associazione "Angela Serra", Benghazi Cancer Registry, Olive Tree Project

ASSOCIAZIONE ANGELA SERRA PER LA RICERCA SUL CANCRO

GRUPPO ITALIANO STUDIO LINFOMI — GISL

FONDAZIONE ITALIANA LINFOMI

REGISTRO TUMORI DELLA PROVINCIA DI MODENA

BENGHAZI CANCER REGISTRY

DataRiver open source data integration

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Premi e riconoscimenti

- **Premio "Intraprendere a Modena 2009" categoria ICT** pari a 3.000 €. Premio speciale **Lapam ICT** pari a 1.000 €

PORTALE DEGLI INCUBATORI UNIVERSITARI
E DELLE BUSINESS PLAN COMPETITION ACCADEMICHE ITALIANE

- partecipazione alla fase finale del **Premio Nazionale per l'Innovazione 2009**

- **2° Premio per il settore ICT** manifestazione **InnovaDay 2011**

Rete Alta Tecnologia Emilia-Romagna

- **Accreditamento Istituzionale come Struttura di Ricerca e Innovazione** della Regione Emilia Romagna. La Giunta Regionale ha deliberato l'Accreditamento definitivo di DATARIVER Srl per gli ambiti ricerca industriale e trasferimento tecnologico (Bollettino Ufficiale Telematico della Regione n. 30 del 15/02/2012)

- Partnership con il laboratorio **SOFTECH-ICT** (www.softech.unimore.it) della Rete Alta Tecnologia della regione Emilia Romagna

softech-ict

DataRiver
open source data integration

UNIVERSITÀ DEGLI STUDI
DI MODENA E REGGIO EMILIA

Finanziamenti

- Finanziata dal Bando della Regione Emilia Romagna per l'attuazione della Attività I.2.1 del POR FESR 2007-2013 "**Sostegno allo start-up di nuove imprese innovative**" per il progetto MOMIS, anno 2011 (44.000 €)
- Finanziata dal bando "**F.I.T. Start-Up**" del Ministero per lo Sviluppo Economico per il progetto **DataRiver Data Integrator**, della durata di 3 anni (684.000 €)
- Bando "**Dai distretti produttivi ai distretti tecnologici 2**" - ICT/MULTIMEDIA. Progetto "Business Analytics per generare valore dai Big Data" di 1 anno (84.000 €) alla Rete TELOS (DataRiver Srl, Apex Srl ed Injenia Srl)

DataRiver
open source data integration

UNIVERSITÀ DEGLI STUDI
DI MODENA E REGGIO EMILIA

Progetto “Business Analytics per generare valore dai Big Data”

- Progetto di 1 anno iniziato a novembre 2012 (84.000 €) alla Rete TELOS (**DataRiver Srl**, **Apex Srl** ed **Injenia Srl**)
- Finanziato dal bando "Dai distretti produttivi ai distretti tecnologici 2" - ICT/MULTIMEDIA

Business Analytics per generare valore dai Big Data

Attività di ricerca e sviluppo per l’integrazione di sorgenti dati aziendali come ERP, DW, CRM con sorgenti dati esterne:

- Big Data
- OpenData
- Dati geografici (Location Intelligence)

DataRiver open source data integration

Progetto “Business Analytics per generare valore dai Big Data”

1. Analisi dei sistemi di Data Integration, Big Data, Open Data (**DataRiver**)
2. Analisi dei sistemi di Business Analytics (**Apex**)
3. ETL semantici per la BI (**DataRiver**)
4. Studio ed applicazione di tecniche di Location Intelligence (**Injenia**)
5. Best practice per la definizione dei Business Analitycs (**Apex**)

DataRiver open source data integration

DEMO CENTER SIPE
CENTRO PER L'INNOVAZIONE E IL TRASFERIMENTO TECNOLOGICO

CENTURIA-RIT CERR CNA INNOVAZIONE CRIT DEMOCENTER-SIPE LARCOICOS MUSP REGGIO EMILIA INNOVAZIONE T3LAB
DISTRETTI 2
dai distretti produttivi ai distretti tecnologici - 2

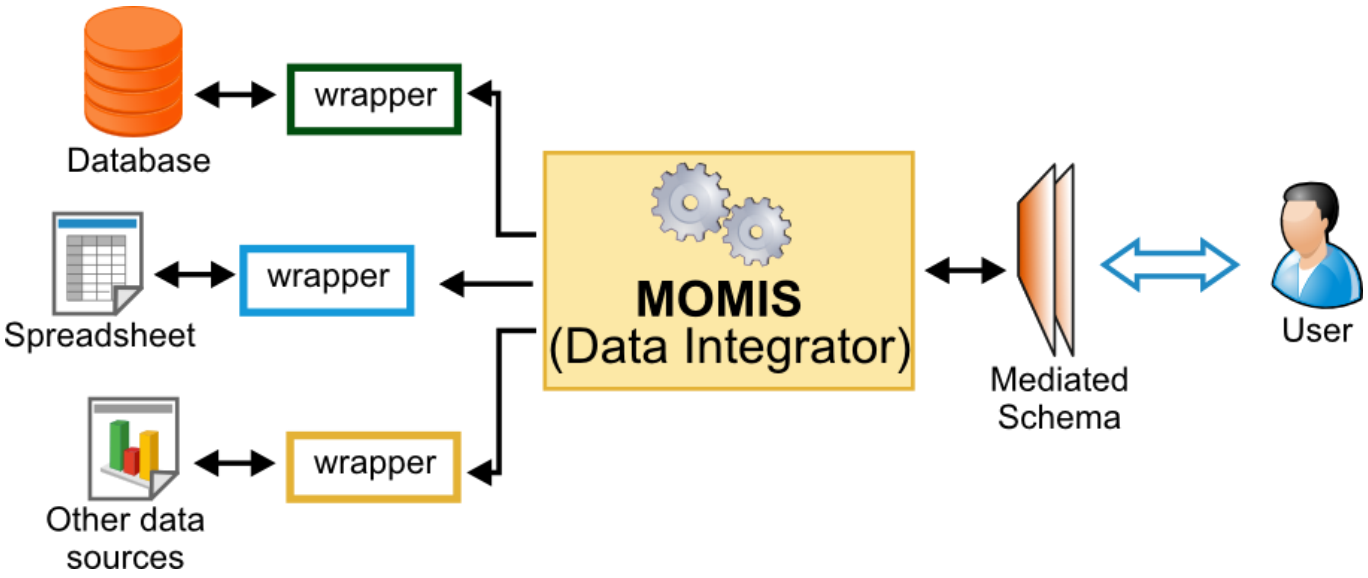
RegioneEmiliaRomagna

“L’integrazione dati è una soluzione tecnologica che ha come obiettivo la costruzione di una base di conoscenza condivisa ed integrata.”

MOMIS
Mediator environment for Multiple Information Sources

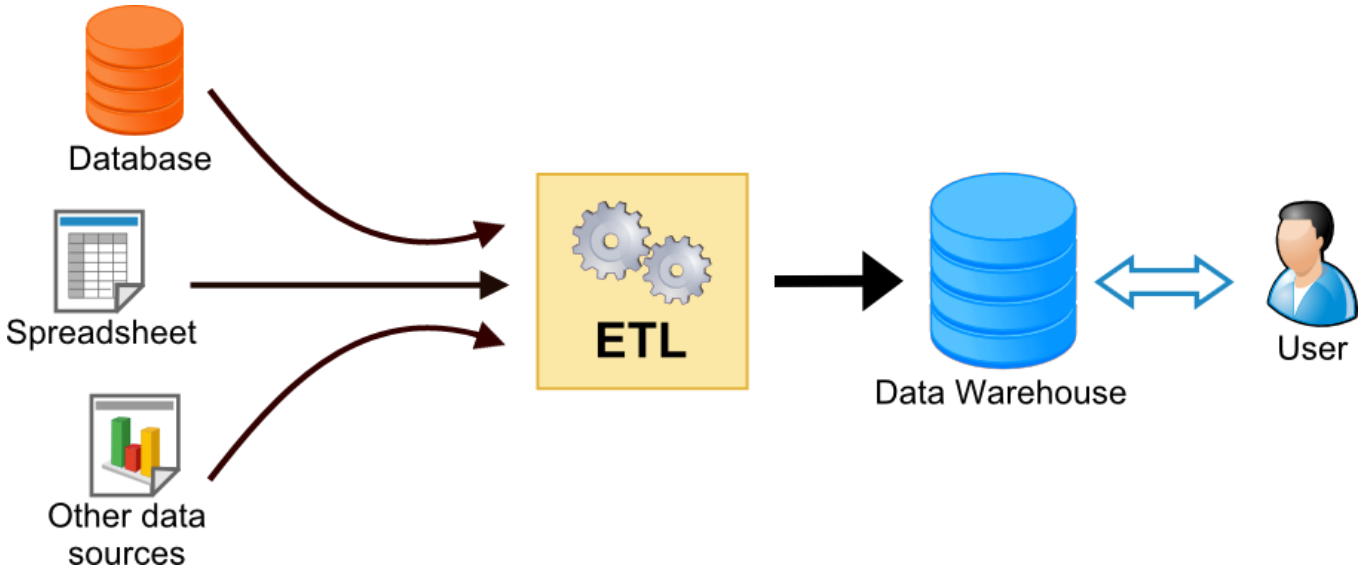
MOMIS (Mediator envirOnment for Multiple Information Sources) è un framework che permette l'estrazione e l'integrazione di sorgenti dati **distribuite** ed **eterogenee** (strutturate e semi-strutturate) in modo semi-automatico, sviluppato dal DBGroup

Integrazione Dati Virtuale



Uno ***Schema Mediato*** fornisce una vista virtuale ed integrata delle sorgenti dati locali coinvolte nell'integrazione. Non viene creata un copia centralizzata dei dati contenuti nelle sorgenti dati, la query posta dall'utente sullo schema mediato viene trasformata in un insieme di query sulle sorgenti locali.

Data Warehouse & Integrazione Dati



I dati contenuti nelle diverse sorgenti vengono ***estratti***, ***trasformati*** e ***caricati*** in un Data Warehouse, sul quale gli utenti possono eseguire le query.

Integrazione Dati Virtuale VS Data Warehouse

	Integrazione Dati Virtuale	Data Warehouse
Dati sempre aggiornati	☺	☹
Integrazione incrementale delle sorgenti dati	☺	☹
Autonomia & Sicurezza	☺	☹
Costo di esecuzione delle query	☹	☺

Strumenti di Estrazione, Trasformazione e Mapping (ETM)

Strumenti avanzati di Estrazione, Trasformazione e Mapping (ETM) vengono forniti dai sistemi per l'integrazione dati

	Virtual Data Integration	Data Warehouse
Scalabilità / Costo del processo di integrazione	☺	☹

DataRiver open source data integration

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

MOMIS
Mediator environment for Multiple Information Sources

MOMIS è un sistema innovativo nel mercato della **Data Integration**. Le **caratteristiche principali** sono:

- Facilità nell'interrogazione dei **Sistemi legacy distribuiti**: non devono essere apportate delle modifiche ai sistemi esistenti (sola lettura)
- **Scalabilità**: per progetti di integrazione complessi (più di 3 sorgenti dati)
- **Dati sempre aggiornati**: Approccio Virtuale, non è necessaria una copia centralizzata dei dati
- **Integrazione di sorgenti dati eterogenee**: sono supportati diversi tipi di sorgente dati (strutturate e semi-strutturate)
- L'**Autonomia** e la **Sicurezza** delle sorgenti dati viene preservata
- **Riduzione** del **costo** del processo di integrazione: sfruttando la **semantica** delle sorgenti dati vengono scoperti i **mapping** tra gli elementi degli schemi delle sorgenti locali

MoMIS

Mediator environment for Multiple Information Sources

Il sistema **MOMIS** è stato utilizzato dal DBGroup in diversi progetti di ricerca. **Portali Web verticali**, **motori di ricerca semantici** e **soluzioni di integrazione dati** sono stati sviluppati in diversi settori: Turismo, Tessile, Meccanico, Logistico, Agro-alimentare, Medico.

WISDOM

sewasie

Semantic Webs and Agents in Integrated Economies

NeP4B

STASIS

Software for Ambient Semantic Interoperable Services

CROSS

LISEA

STIL

Cerealab

siteia

Olive Tree Project

Nel progetto **Olive Tree**, il sistema **MOMIS** è stato utilizzato per l'integrazione di dati clinici di pazienti provenienti da dieci diversi paesi dell'area del Mediterraneo.

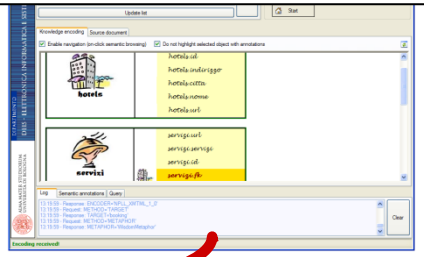
DataRiver

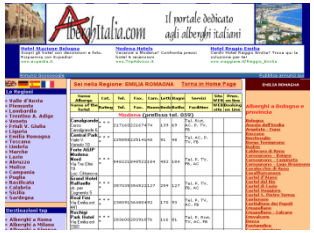
open source data integration

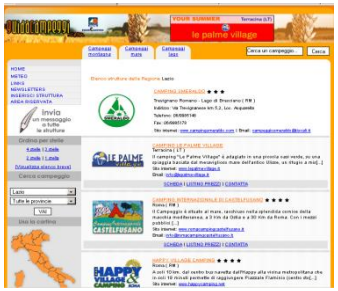
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

MOMIS: Dominio del Turismo

- Nel progetto **WISDOM** (**W**eb **I**ntelligent **S**earch based on **DOM**ain ontologies) (www.dbgroup.unimo.it/wisdom) il sistema **MOMIS** è stato utilizzato per l'integrazione di diversi siti web sul turismo e per lo sviluppo di un Portale Web Verticale sul Turismo.



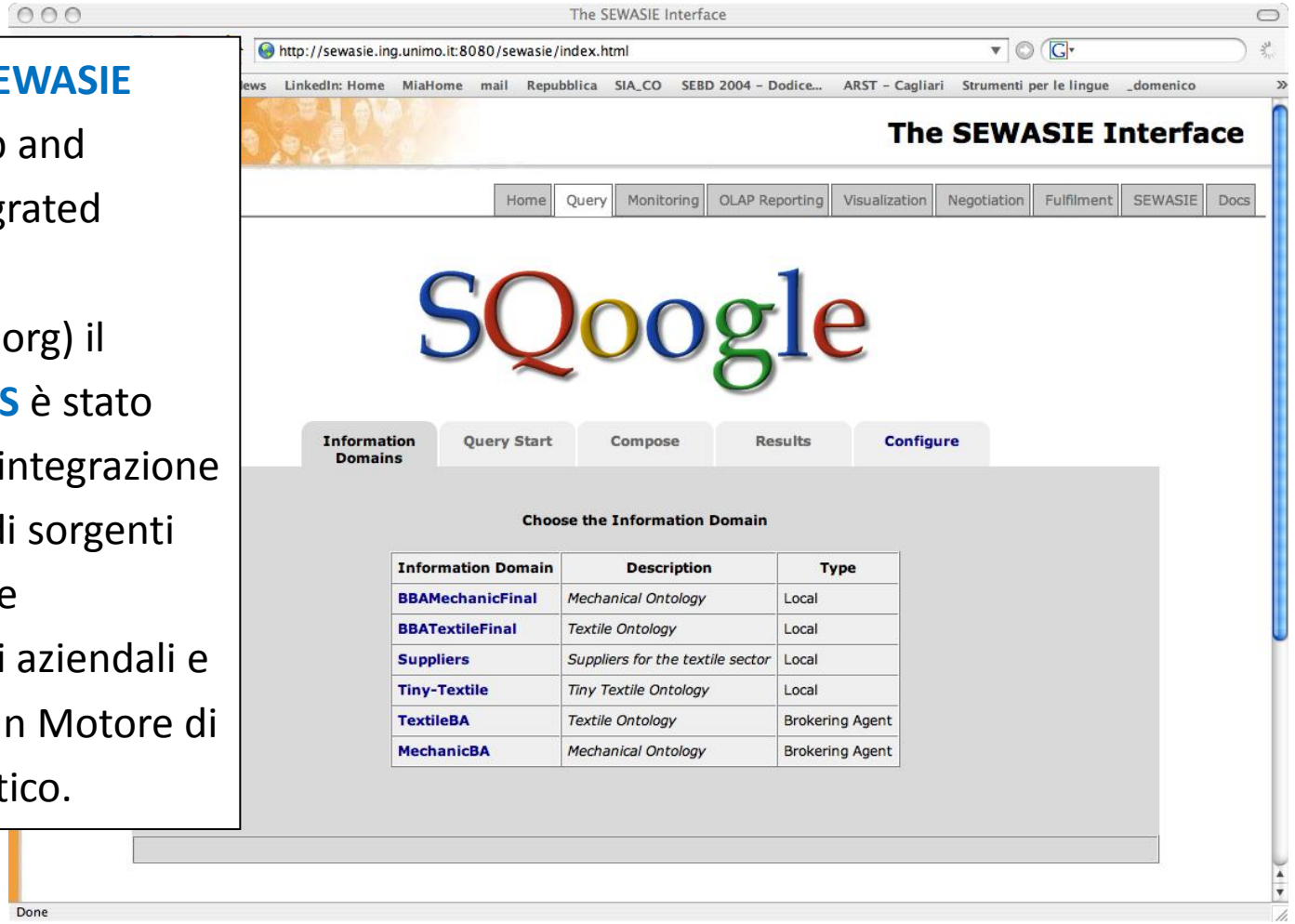




...

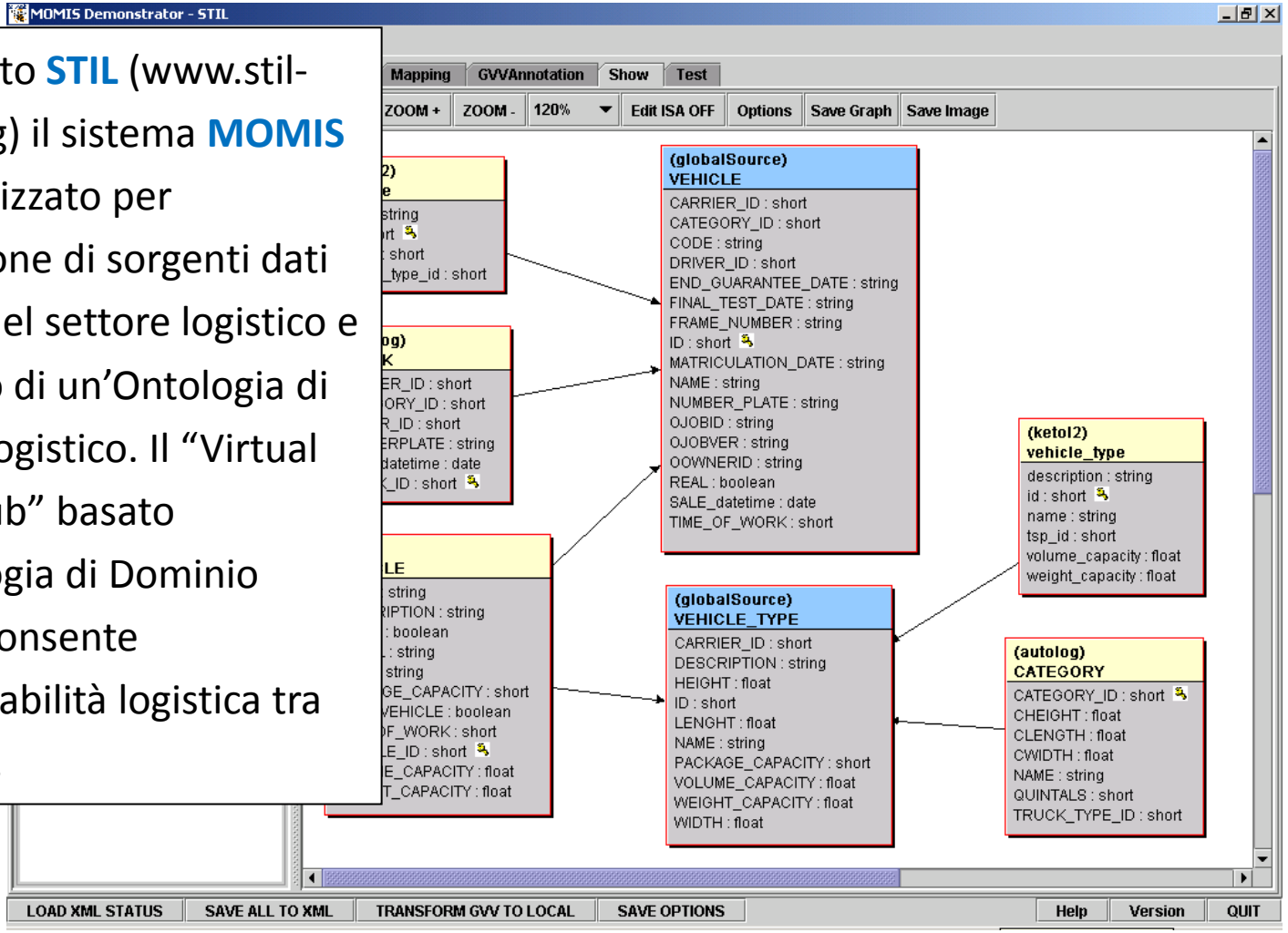
MOMIS: Dominio Tessile e Meccanico

- Nel progetto **SEWASIE** (**SE**mantic **W**eb and **A**gent**S** in **I**ntegrated **E**conomies) (www.sewasie.org) il sistema **MOMIS** è stato utilizzato per l'integrazione di un insieme di sorgenti dati eterogenee contenenti dati aziendali e lo sviluppo di un Motore di Ricerca Semantico.

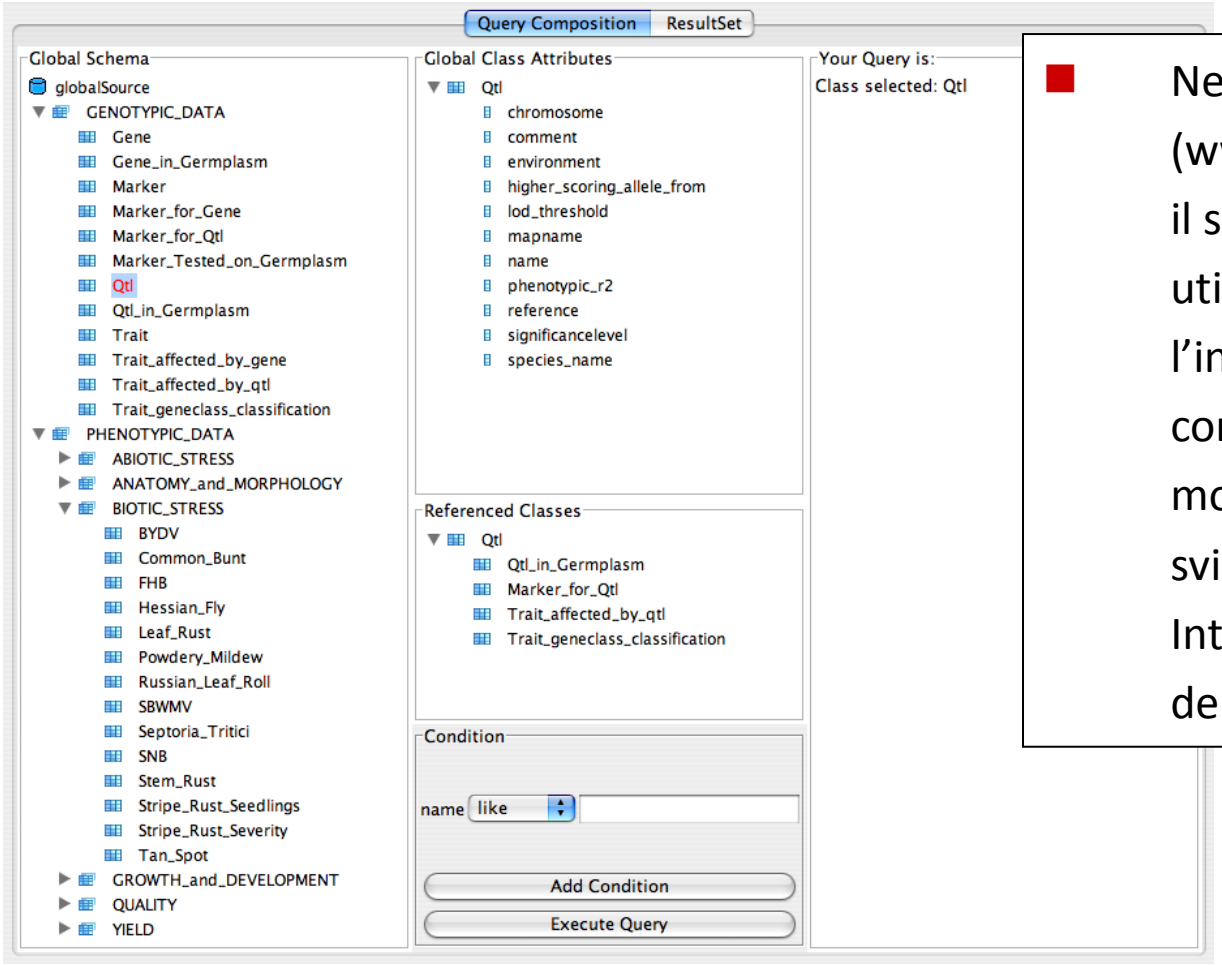


MOMIS: Dominio Logistico

- Nel progetto **STIL** (www.stil-project.org) il sistema **MOMIS** è stato utilizzato per l'integrazione di sorgenti dati aziendali del settore logistico e lo sviluppo di un'Ontologia di Dominio Logistico. Il "Virtual Logistic Hub" basato sull'Ontologia di Dominio Logistico consente l'interoperabilità logistica tra le aziende.



MOMIS: Dominio Agroalimentare

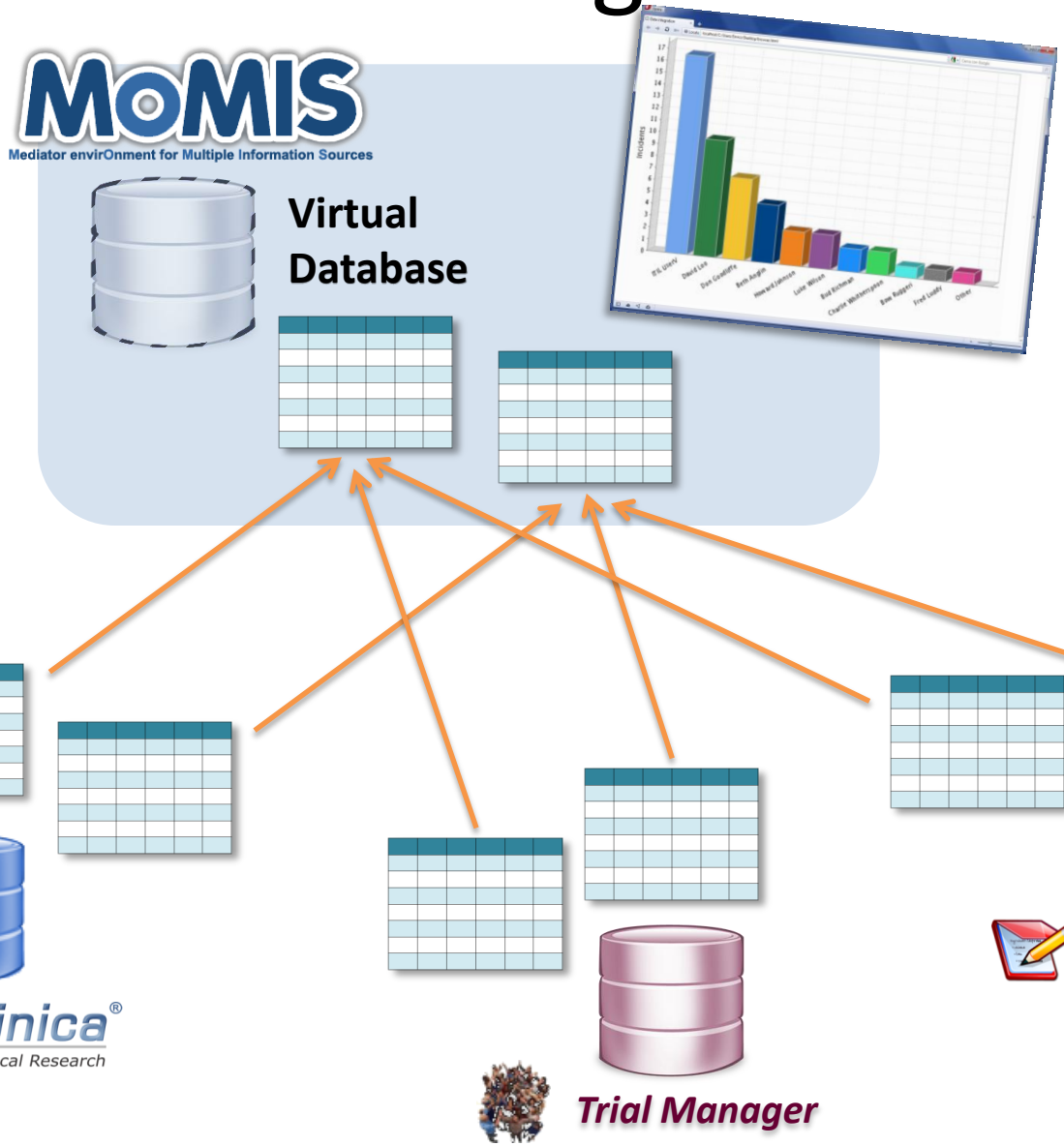


- Nel progetto **CEREALAB** (www.cerealab.unimore.i) il sistema **MOMIS** è stato utilizzato per l'integrazione di sorgenti contenti dati fenotipici e molecolari sui cereali e lo sviluppo di un Database Integrato per i coltivatori dei cereali.

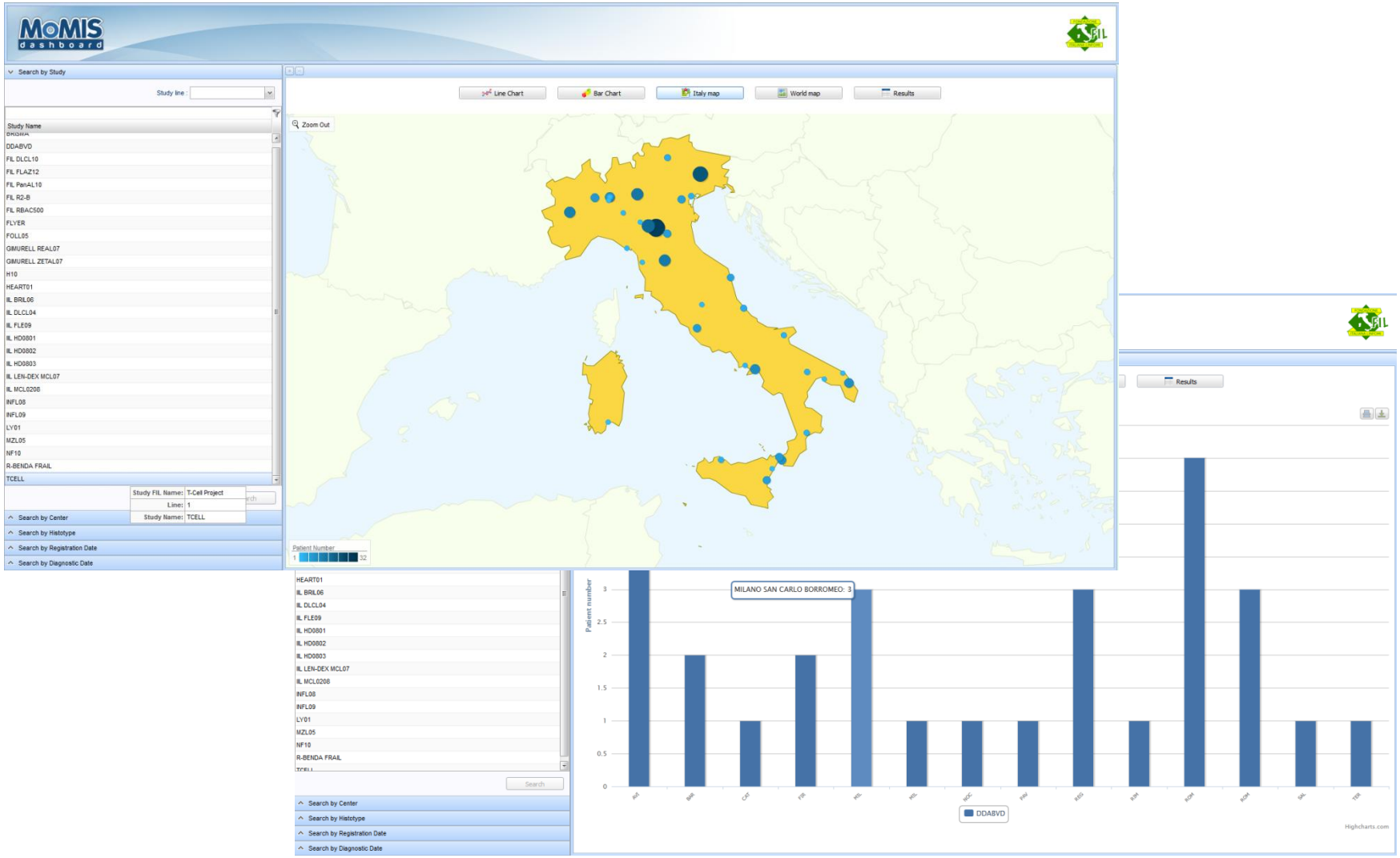
DataRiver open source data integration

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

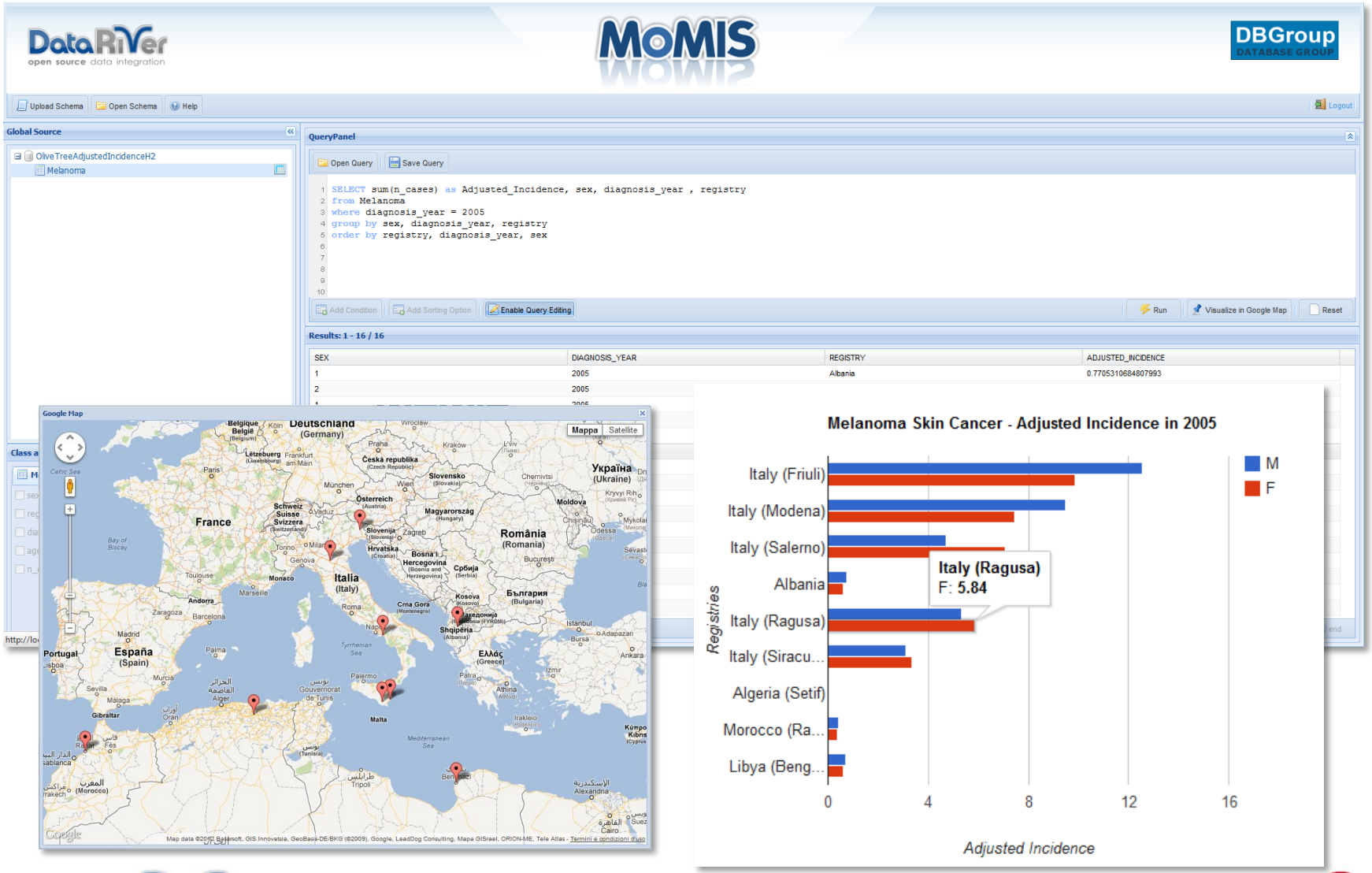
MOMIS: Progetto FIL



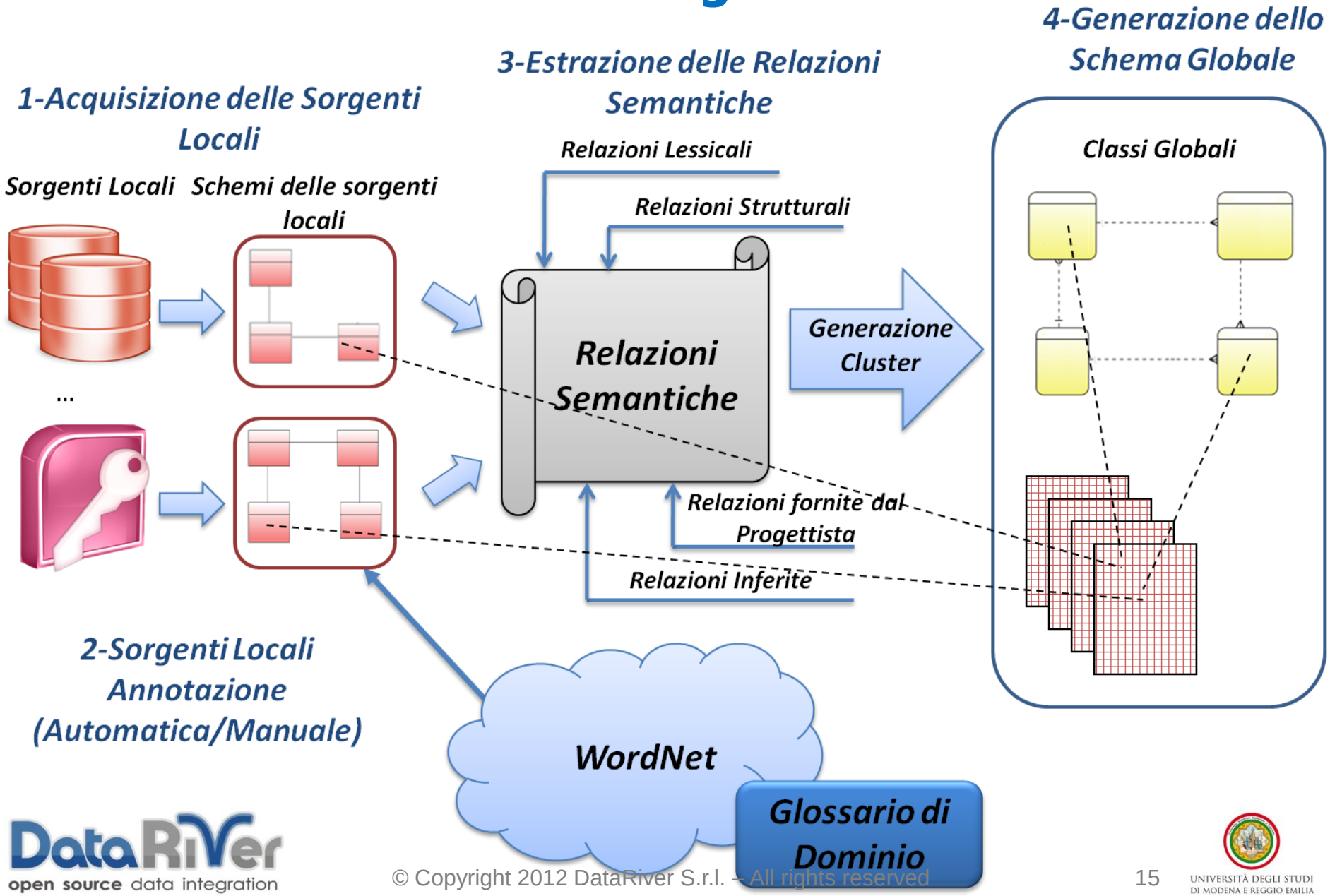
MOMIS: Progetto FIL



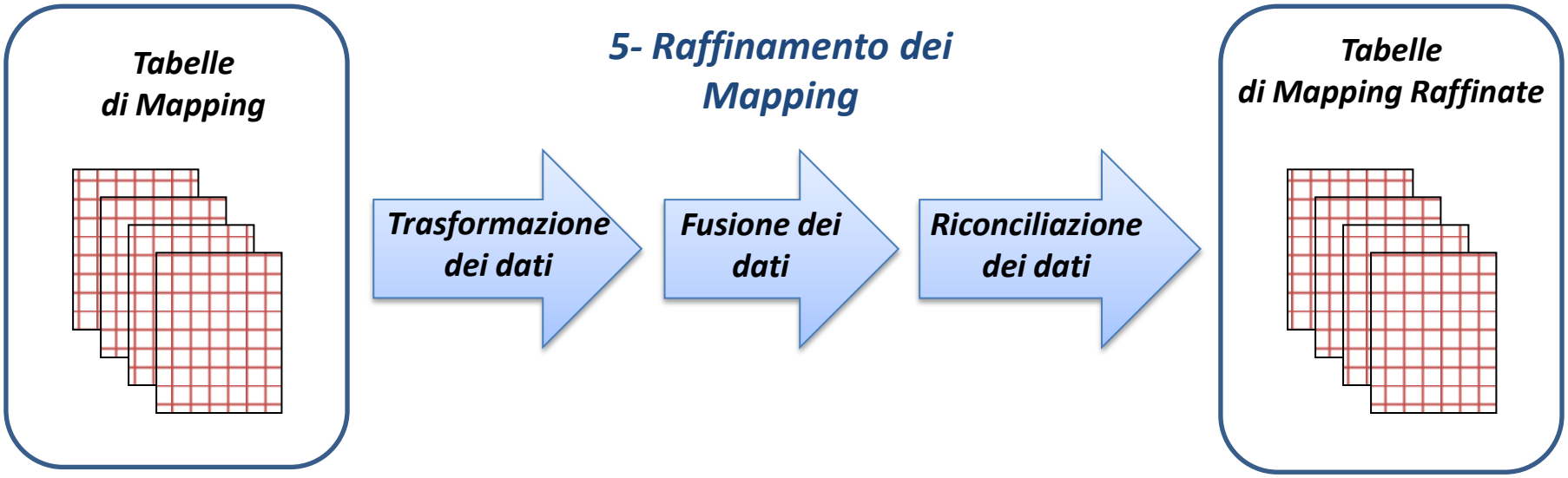
MOMIS: Progetto Olive Tree



Processo di Integrazione Dati



Processo di Integrazione Dati



- *Funzioni di Trasformazione* per trasformare i valori degli attributi
- *Funzioni di Join* per fondere insieme i risultati parziali provenienti dalle sorgenti locali (operatore di default: FOJ)
- *Funzioni di Risoluzione* per risolvere i conflitti tra i dati

Annotazione

Annotazione delle sorgenti locali:

- associare uno o più significati ai nomi delle classi e degli attributi degli schemi delle sorgenti locali, rispetto ad un ontologia lessicale (nel nostro caso viene utilizzato il database lessicale **WordNet**)
- generare relazioni semantiche tra gli elementi (nomi di classi e attributi) degli schemi delle sorgenti locali.

  - *Annotazione Manuale*: il progettista manualmente seleziona il significato corretto per ogni elemento dello schema
  - *Annotazione Automatica*: applicazione degli algoritmi di Word Sense Disambiguation (WSD)

Annotazione

Company(Name, Section, Address, Phone)

Enterprise(Company Name, Department, Revenue)

Annotazione: Significati di WordNet + Glossario di Dominio

	Name	Comapny Name	Department	Section	Phone	Revenue
a language unit by which a person or a thing is known	○					
the name by which a corporation is identified		○				
a self-contained part of a larger composition			○			
a specialized division of a large organization			○	○		
the number is used in calling a particular telephone					○	
electronic equipment that converts sound into electrical signals that can be transmitted over distances					○	
the entire amount of income before any deductions are made						○

Database Lessicale (WordNet + Glossario di Dominio) → Annotazione Automatica → Annotazione Manuale → Generazione delle Relazioni Semantiche

DataRiver open source data integration

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Annotazione Automatica: WordNet + Glossario di Dominio

	Name	Comapny Name	Department	Section	Phone	Revenue
a language unit by which a person or a thing is known	✕					
the name by which a corporation is identified		✕				
a self-contained part of a larger composition			✕			
a specialized division of a large organization			○	✕		
the number is used in calling a particular telephone					✕	
electronic equipment that converts sound into electrical signals that can be transmitted over distances					○	
the entire amount of income before any deductions are made						✕

DataRiver open source data integration

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Annotazione Manuale: WordNet + Glossario di Dominio

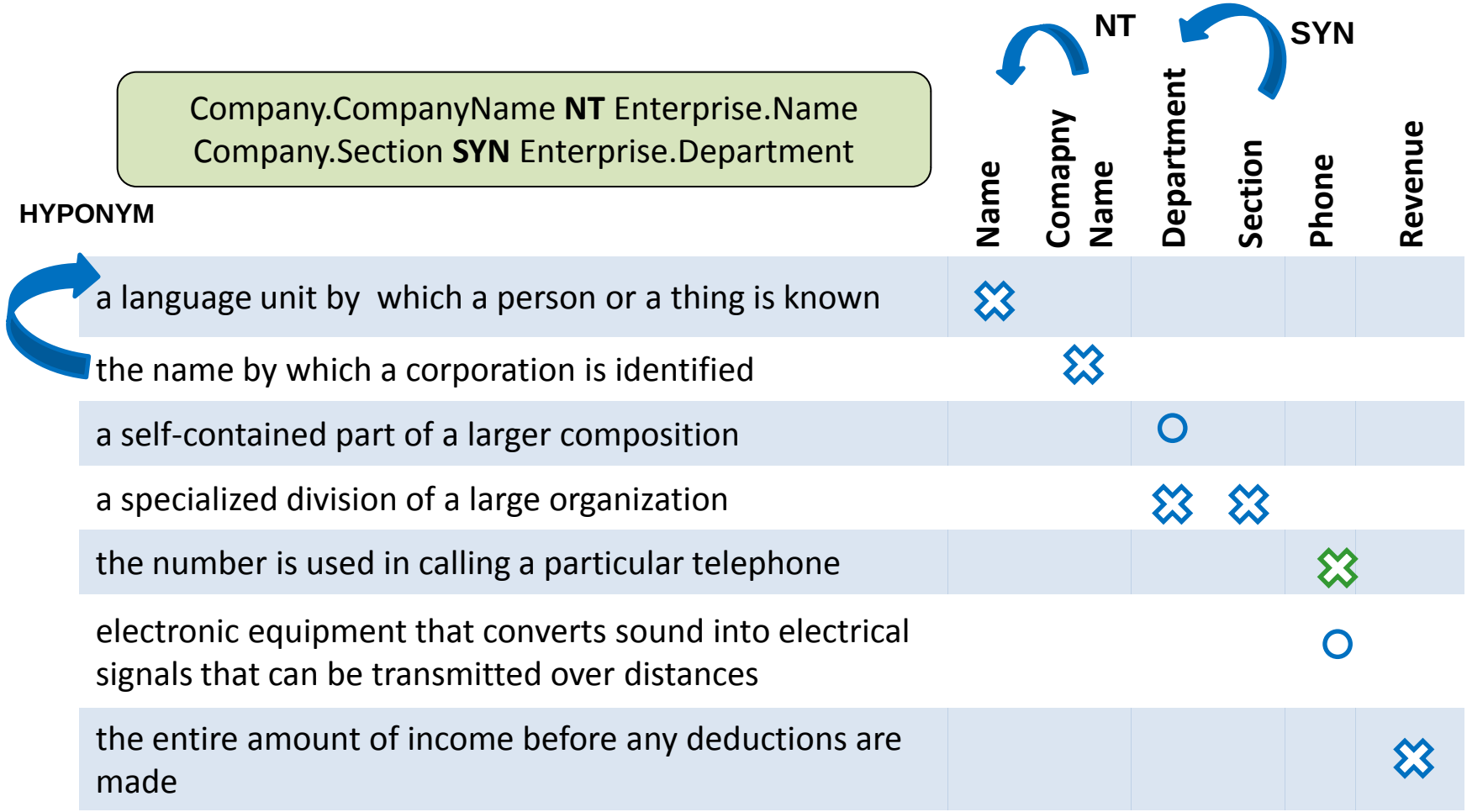
	Name	Comapny Name	Department	Section	Phone	Revenue
a language unit by which a person or a thing is known	✕					
the name by which a corporation is identified		✕				
a self-contained part of a larger composition			○			
a specialized division of a large organization			✕	✕		
the number is used in calling a particular telephone					✕	
electronic equipment that converts sound into electrical signals that can be transmitted over distances					○	
the entire amount of income before any deductions are made						✕

DataRiver open source data integration

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Generazione delle Relazioni Semantiche

Company.CompanyName **NT** Enterprise.Name
Company.Section **SYN** Enterprise.Department



Database Lessicale (WordNet + Glossario di Dominio) → Annotazione Automatica → Annotazione Manuale → Generazione delle Relazioni Semantiche

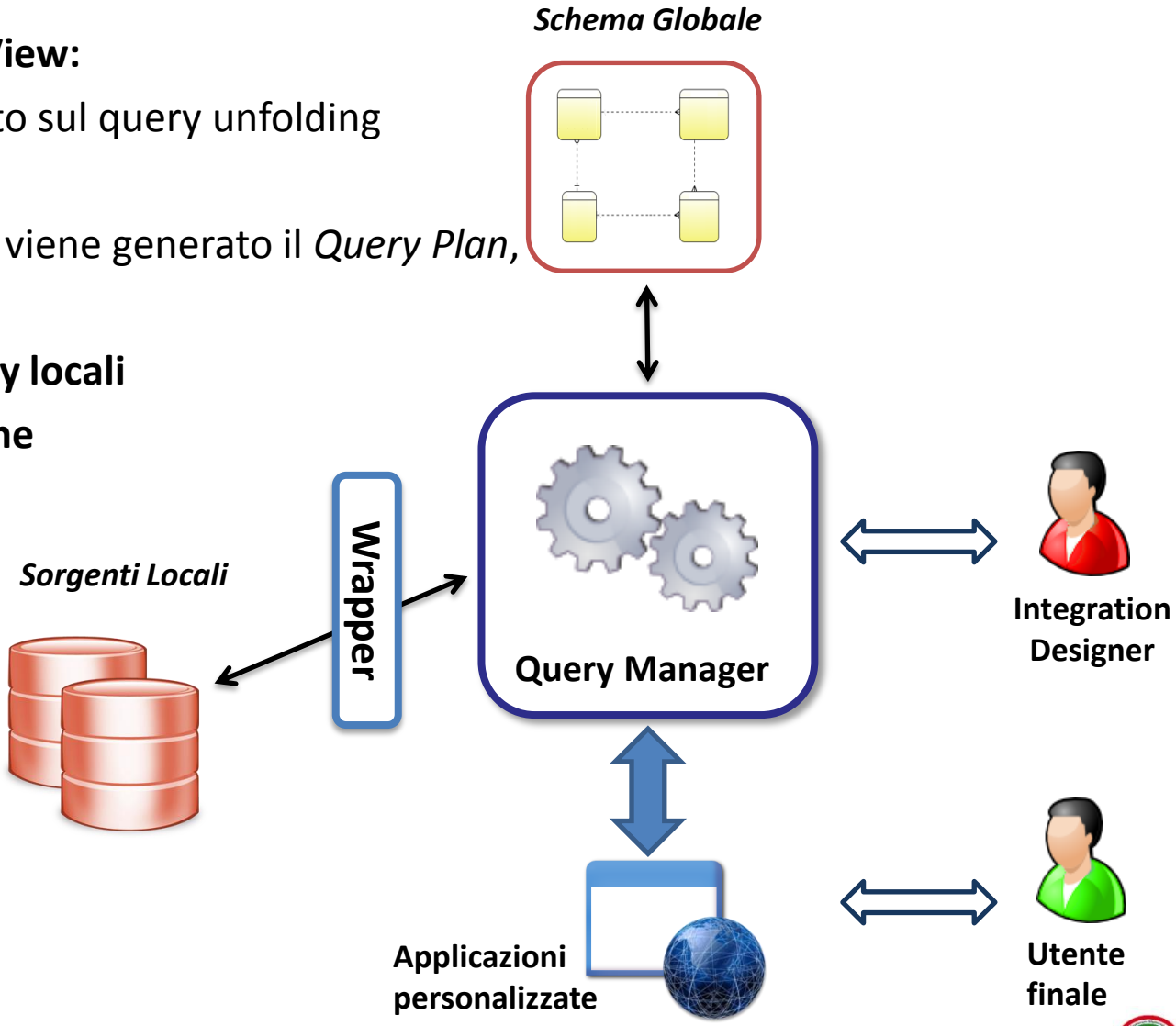
Interrogazione dello Schema Globale

Approccio Global-As-View:

query processing basato sul query unfolding

Per ogni query globale viene generato il *Query Plan*, costituito da:

- **un insieme di query locali**
- **una query di fusione**
- **una query finale**



Sfide e Soluzioni per la Data Integration

Le sfide del processo di integrazione

In che modo MOMIS 1.2 affronta queste sfide?

- Capire qual è lo Schema Globale più appropriato per il problema di integrazione in questione
  - Possibilità di creare diversi schemi globali e confrontarli durante il processo di integrazione
- Superare la conoscenza parziale delle sorgenti dati e capire il dominio di applicazione
  - una suite di strumenti per l'annotazione semantica delle sorgenti dati rispetto ad una risorsa lessicale e/o al glossario di dominio
- Minimizzare i costi del processo di integrazione
  - una interfaccia grafica che facilita il processo di integrazione
  - generazione semi-automatica dei mapping e dello Schema Globale
- Esaminare il risultato del processo di integrazione in ogni fase
  - Un set di strumenti di esplorazione e preview consentono al progettista di visualizzare in anteprima il risultato dell'integrazione durante ogni fase.

La Release 1.2 di MOMIS

- La release 1.2 del sistema **Open Source MOMIS** è disponibile per il **download** sul sito **www.datariver.it**
- MOMIS è distribuito sotto la licenza GNU General Public License (GPLv2)
- Un **manuale** ed un insieme di **video tutorial** che dimostrano come è possibile integrare velocemente le sorgenti dati con MOMIS sono disponibili sul sito **www.datariver.it**
- Incoraggiamo sia gli sviluppatori che i ricercatori a scaricare la versione 1.2 del software e a contribuire alle versioni future del sistema MOMIS.
- Utenti registrati : 131
- Download di MOMIS 1.1: 168 (da Aprile 2011), video tutorial: 928, YouTube: 664

DataRiver open source data integration

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

MOMIS 2.0 : roadmap

- **Annotazione Automatica** per ***velocizzare*** il ***processo di integrazione***:
  - Combinazione di diversi **metodi di annotazione**, anche **probabilistici**
  - Gestione ed espansione delle **abbreviazioni** e degli **acronimi**
  - **Nomi Composti** (composti da più parole)

- **Provenance** : Provenance (o **lineage**) per determinare la **provenienza** dei dati e **come** quest'ultimi siano stati derivati. La **provenance** viene utilizzata a fini statistici, e per effettuare il **data cleaning** (pulizia dei dati delle sorgenti locali)

- **Object Identification**: Per identificare istanze diverse dello stesso oggetto del mondo reale, nelle diverse sorgenti locali (detto anche **record linkage** o **duplicate detection**)
  In MOMIS 1.2 : corrispondenza esatta
  In MOMIS 2.0 : metodi probabilistici basati su misure di similarità

- **Ambiente di collaborazione**: per incentivare e rendere possibile la collaborazione tra gli integration designer

DataRiver open source data integration

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Prodotti e Servizi

DataRiver si occupa di progettare e sviluppare soluzioni per la Data Integration risolvendo problematiche di incongruenza, eterogeneità e pulizia dei dati, tramite tecniche provenienti dalla ricerca nel campo del Semantic Web.

Le soluzioni offerte da **DataRiver** consentono di creare valore derivante dai dati integrati, e **di migliorare i processi decisionali**, **produttivi** e **previsionali** ottimizzandone costi e tempi.

DataRiver fornisce soluzioni all'avanguardia e **consulenza specializzata** per problematiche di **Data Integration, Semantic Web, Business Intelligence e Clinical Data Management**

DataRiver ha acquisito una vasta esperienza nella realizzazione di **portali web verticali** e nelle **soluzioni per l'integrazione dei dati** in diversi settori: **Turistico**, **Tessile**, **Meccanico**, **Logistico**, **Agro-alimentare**, **Medico**

DataRiver ha sviluppato una profonda competenza nell'ambito del **Clinical Data Management.** I servizi attualmente forniti riguardano lo sviluppo di **Sistemi Gestionali per Studi Clinici** e di **Sistemi Gestionali per Registri Tumori.**

Il Team di DataRiver

Il team di DataRiver è composto da professori, ricercatori e professionisti del campo dell' ICT.

Staff:

Mirko Orsini, Ph.D.
Ing. Entela Kazazi
Ing. Enrico Calanchi
Ing. Sara Quattrini
Ing. Giovanni Simonini
Ing. Fabio Benedetti

Collaboratori:

Silvano Pancaldi (Injenia Srl)

Soci:

Mirko Orsini, Ph.D.
Prof. Sonia Bergamaschi
Prof. Domenico Beneventano
Alberto Corni, Ph.D.
Laura Po, Ph.D.
Serena Sorrentino, Ph.D.

DataRiver
open source data integration

Mirko Orsini
DataRiver Srl
Via Vignolese, 905
Facoltà di Ingegneria “Enzo Ferrari”
Università di Modena e Reggio Emilia

Web: **www.datariver.it**
Email: **mirko.orsini@datariver.it**