

The MOMIS approach for Information Integration Practical Session

www.dbgroup.unimo.it

Dipartimento di Ingegneria "Enzo Ferrari"

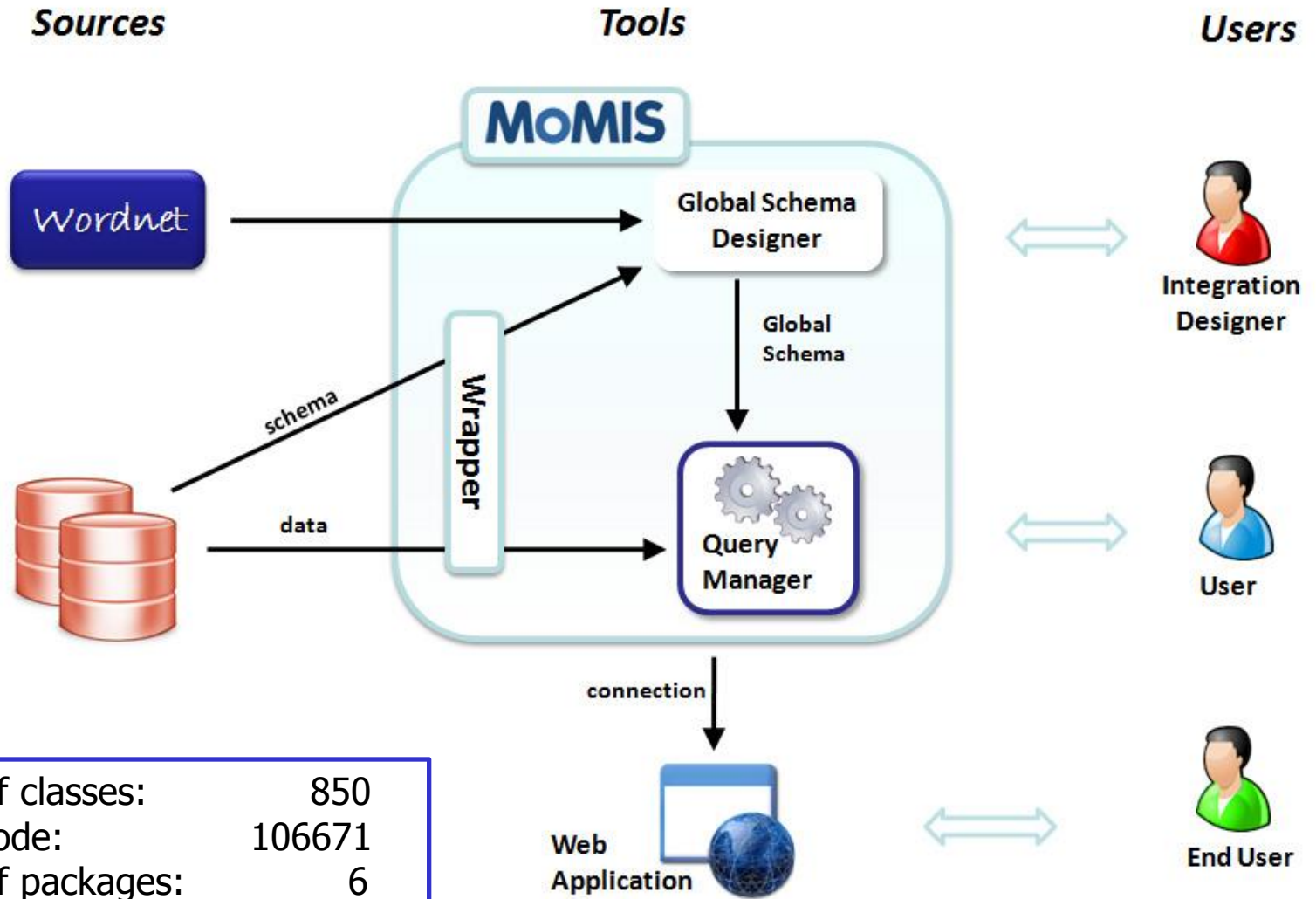
Università di Modena e Reggio Emilia, via Vignolese 905, 41100 Modena

www.datariver.it

Spin-off presso Università di Modena e Reggio Emilia



DataRiver
open source data integration



| | |
|-------------------------|--------|
| Number of classes: | 850 |
| Lines of code: | 106671 |
| Number of packages: | 6 |
| Number of sub-packages: | 86 |

1. Global Schema and Mapping generation

→ the Global Schema is generated and its mappings with the data source are defined

2. Querying the Integrated Data

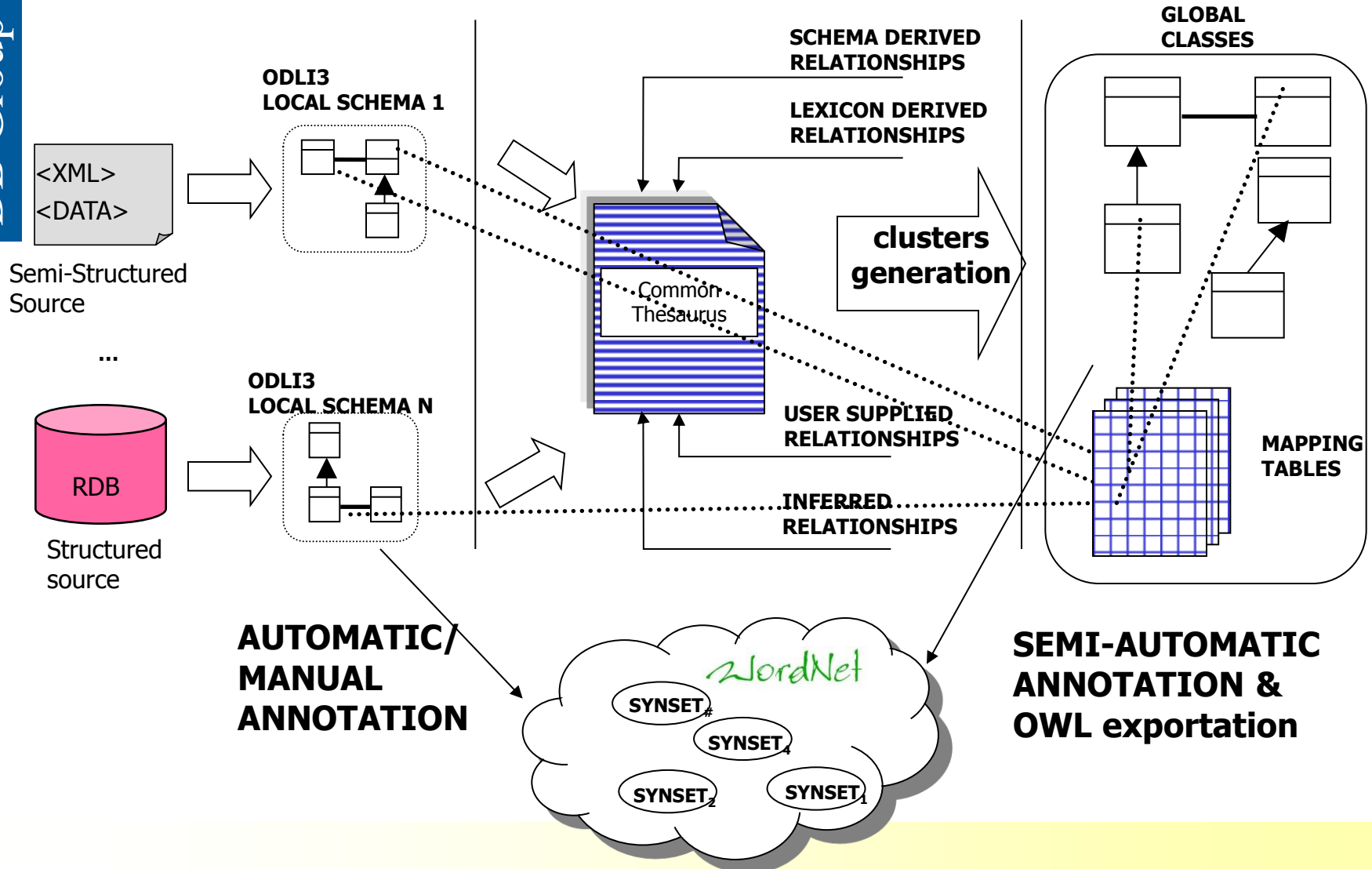
→ the integrated data can be queried by users and software applications

Overview of the GS generation process

COMMON THESAURUS GENERATION

WRAPPING

GS GENERATION



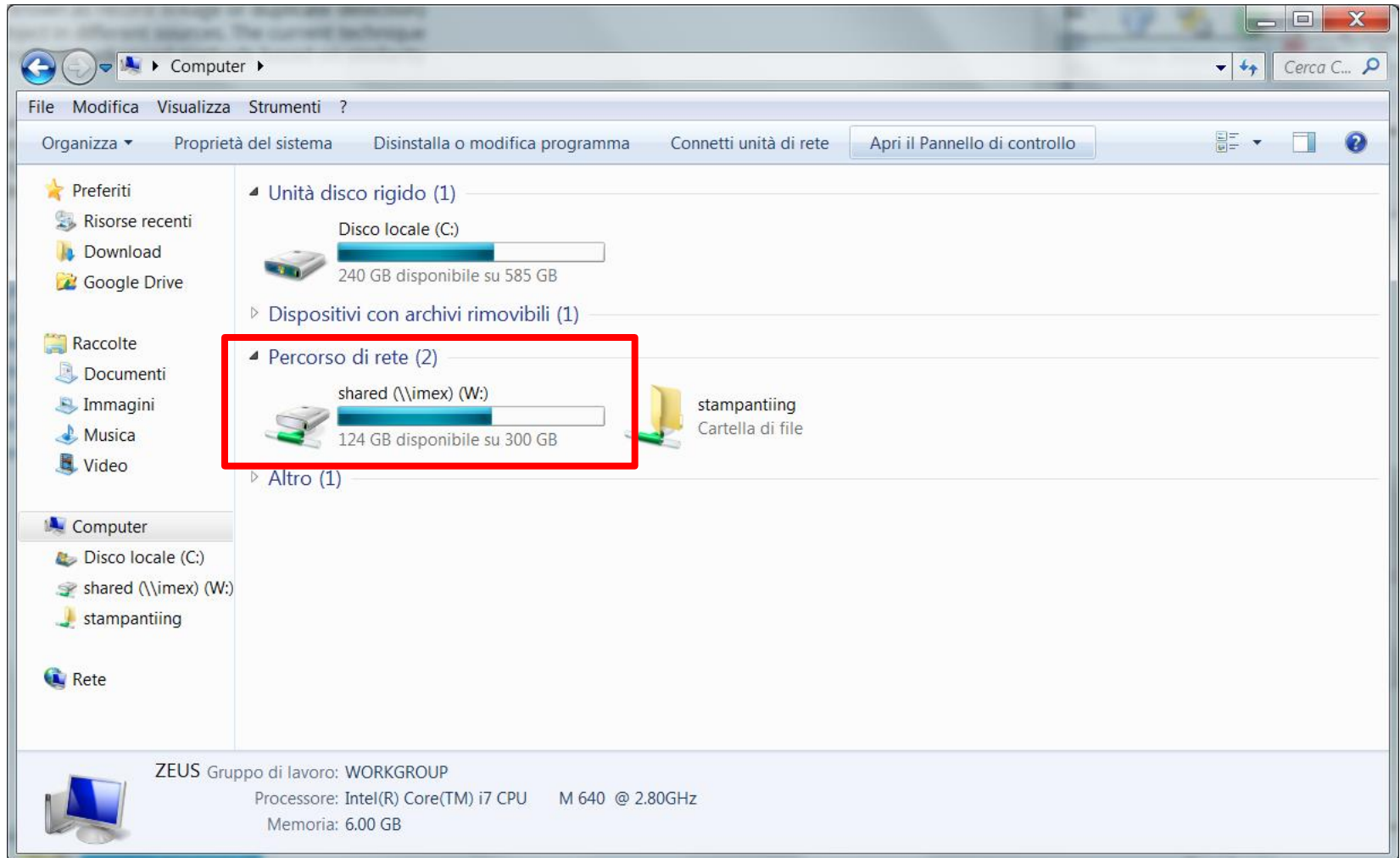
How we can use MOMIS in order to build a global view over different sources

- **Set up the environment** - Download MOMIS
- **Test 1** - Test the integration process over three sources:
 - 2 medical sources containing information about the number of melanoma cases per age
 - 1 source pertaining to vital statistics
- **Test 2** - Use an existing project in order to see the final integration result and how we can pose meaningful query over the integrated view

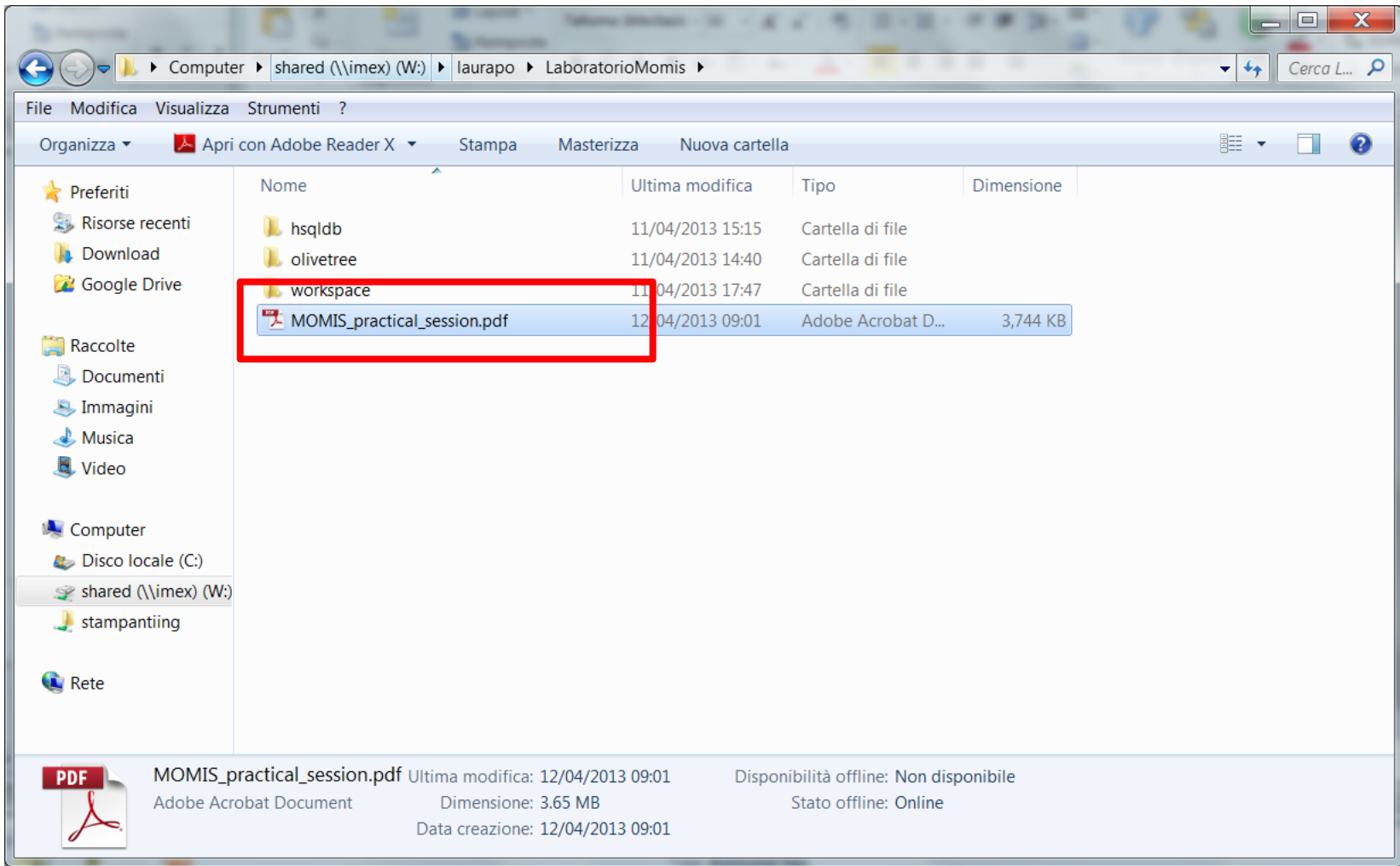
SET UP THE ENVIRONMENT

1. Download this presentation
2. Connect to www.datariver.it
3. Create an account and download
momis_1-2_win32_x86_jre
4. Run MOMIS

Connect to the IMEX server



- Open the folder
W:\laurapo\LaboratorioMomis\
 - Download the presentation
MOMIS_practical_session.pdf



If you do not find the network resource IMEX, you can reach the server by using the connection command

- `net use W: \\imex\shared /user:imex\reader primo99`

www.datariver.it

Google

ENGLISH

Data River
open source data management

SPIN OFF
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

COMPANY DATA INTEGRATION CLINICAL DATA MANAGEMENT CONTACT US

**GAIN A CLEAR AND INTEGRATED VIEW
OF YOUR ORGANIZATION'S DATA**
Data Integration + Semantics

Kpi 1 Kpi 2
Kpi 3 Kpi 4



Data Integration

Data River is specialized in data integration.

We design solutions for:

- Data integration
- Information Management



Clinical Data Management

Collaboration with qualified research centers.

We realize management information systems for:

- Clinical Studies

Go to the download page

www.datariver.it

Google

ENGLISH

DataRiver
open source data management

SPIN OFF
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

COMPANY

DATA INTEGRATION

CLINICAL DATA MANAGEMENT

CONTACT US

**GAIN A CLEAR AND
OF YOUR ORGAN
Data Integrati**

**ED VIEW
DATA**

SOLUTIONS
PROJECTS
TRAINING
MOMIS

HOW IT WORKS
TUTORIALS
DOWNLOAD

Kpi 1
Kpi 2
Kpi 3
Kpi 4

Data Integration
Data River is specialized in data integration.
We design solutions for:
- Data integration

Clinical Data Management
Collaboration with qualified research centers.
We realize management information systems for:
- Clinical Studies

www.datariver.it/data-integration/momis/download/management

MOMIS

[SOLUTIONS](#)[PROJECTS](#)[TRAINING](#)[MOMIS](#)[HOW IT WORKS](#)[TUTORIALS](#)[DOWNLOAD](#)

DOWNLOAD

Login

[Lost Password](#)

Register

Please register or login to have access to download area

Username E-mail

Password Verify password



CAPTCHA Code

DOWNLOAD
MOMIS 

MOMIS is a Free Software released under the GNU General Public License (v. 2.0).

MOMIS

SOLUTIONS

PROJECTS

TRAINING

MOMIS

HOW IT WORKS

TUTORIALS

DOWNLOAD

DOWNLOAD

Login

[Lost Password](#)

Register

Please register or login to have access to download area

Username E-mail

Password Verify password



CAPTCHA Code

DOWNLOAD
MOMIS 

MOMIS is a Free Software released under the GNU General Public License (v. 2.0).

Select the momis_1-2_win32_x86_jre version

DOWNLOAD

MOMIS is a Free Software released under the GNU General Public License (version 2).

DataRiver releases the 1.2 version of the [MOMIS Data Integration System](#). The new Open Source version of the system is now available for download. The system enables users and applications to execute queries on a set of data sources, automatically fuses together the partial results coming from the sources, and returns a single, unified, answer. The supported data sources are: MySQL, Microsoft SQL Server, Oracle, DB2, PostgreSQL, JDBC data sources, JDBC-ODBC data sources, Microsoft Excel files, CSV files, data sources available via Web Service.

The main new features of the release 1.2 are:



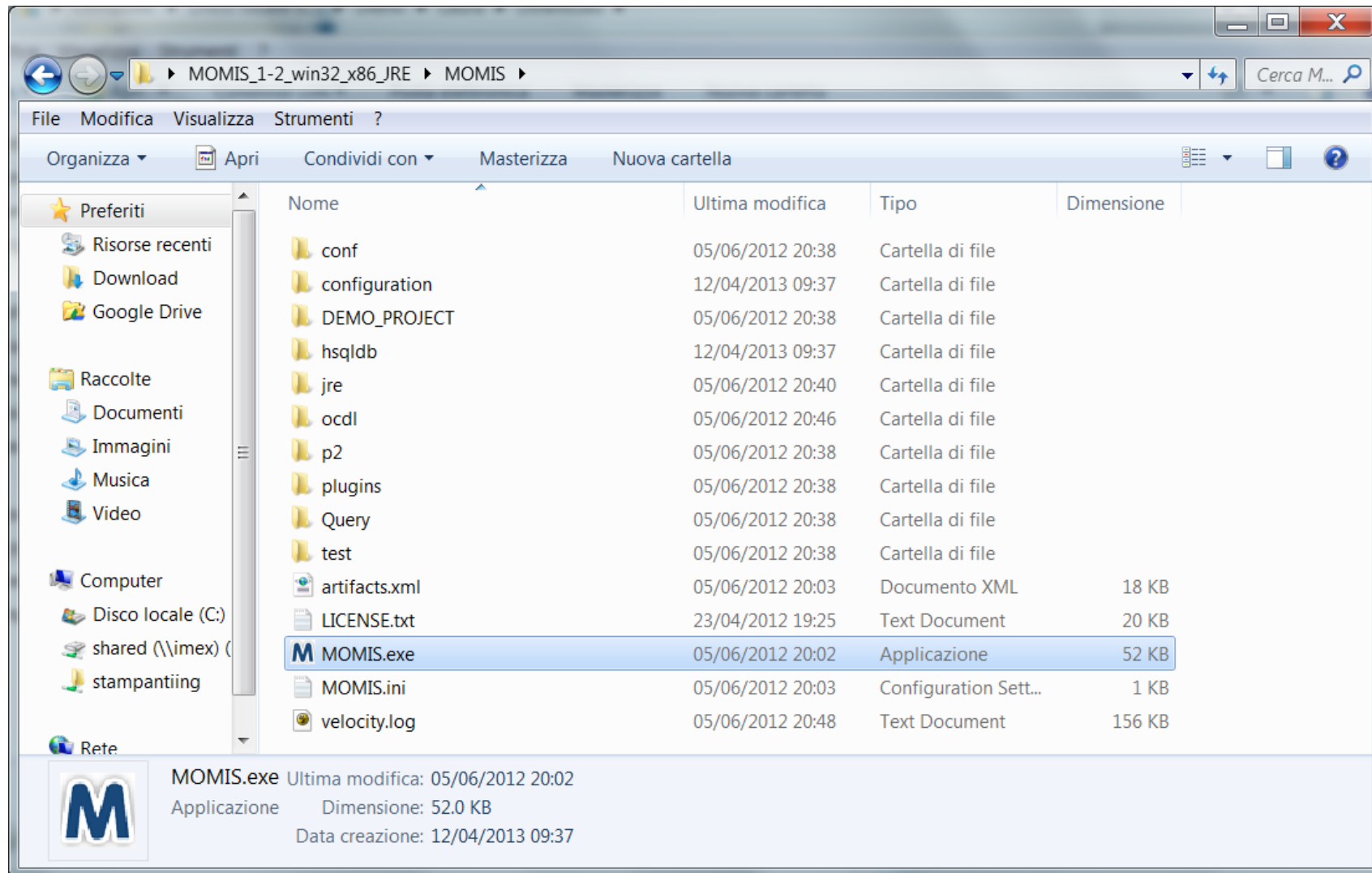
MOMIS is a Free Software released under the GNU General Public License (v. 2.0).

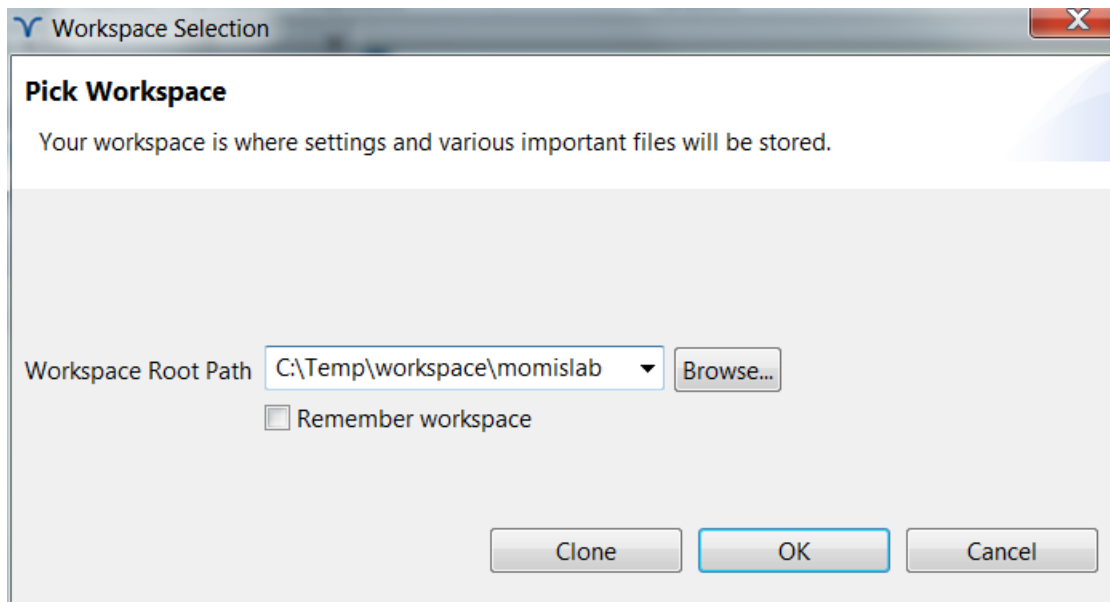
MOMIS requires the Java Runtime Environment (JRE) 1.6 or above. If you already have a JRE (or JDK) installed on your computer, you can download the MOMIS version without the JRE included. If you don't have a JRE installed on your computer, download the MOMIS version with the JRE included.

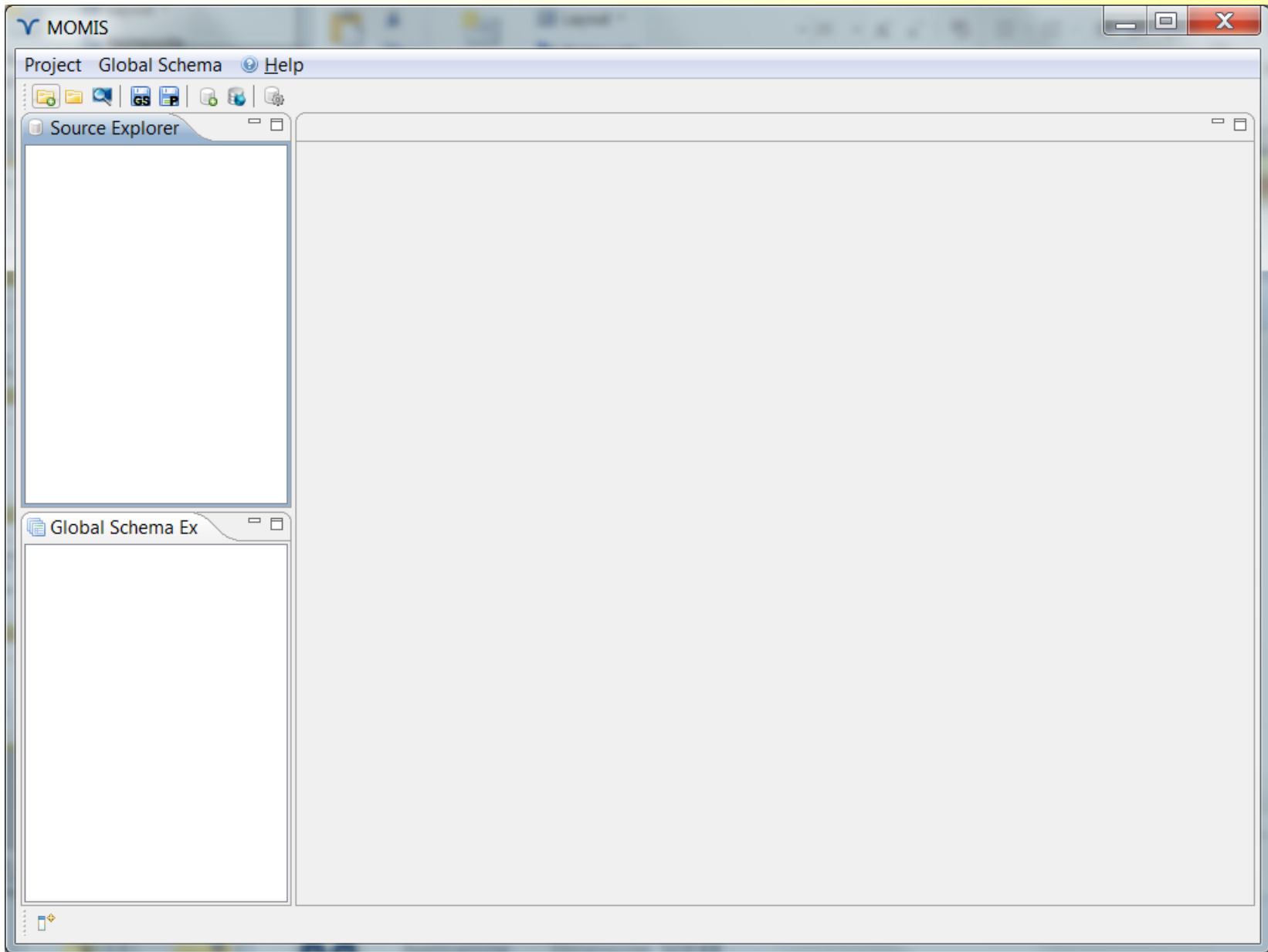
- ✓ MOMIS
 - ✓ momis source code 1.2 Version:1.2.0
 - ✓ momis_1-2_macosx_cocoa_x86_64 Version:1.2.0
 - ✓ momis_1-2_macosx_cocoa_x86 Version:1.2.0
 - ✓ momis_1-2_win32_x86_64 Version:1.2.0
 - ✓ momis_1-2_win32_x86_jre Version:1.2.0
 - ✓ momis_1-2_win32_x86 Version:1.2.0
 - ✓ momis_1-2_linux_gtk_x86_64 Version:1.2.0
 - ✓ momis_1-2_linux_gtk_x86 Version:1.2.0

- ✓ TUTORIALS
- ✓ WORDNET

- Extract the MOMIS archive on the desktop
- Execute the file MOMIS.exe that is located in
MOMIS_1-2_win32_x86_JRE\MOMIS\
MOMIS_1-2_win32_x86_JRE\MOMIS\
- Define the workspace
C:\Temp\workspace\momislab







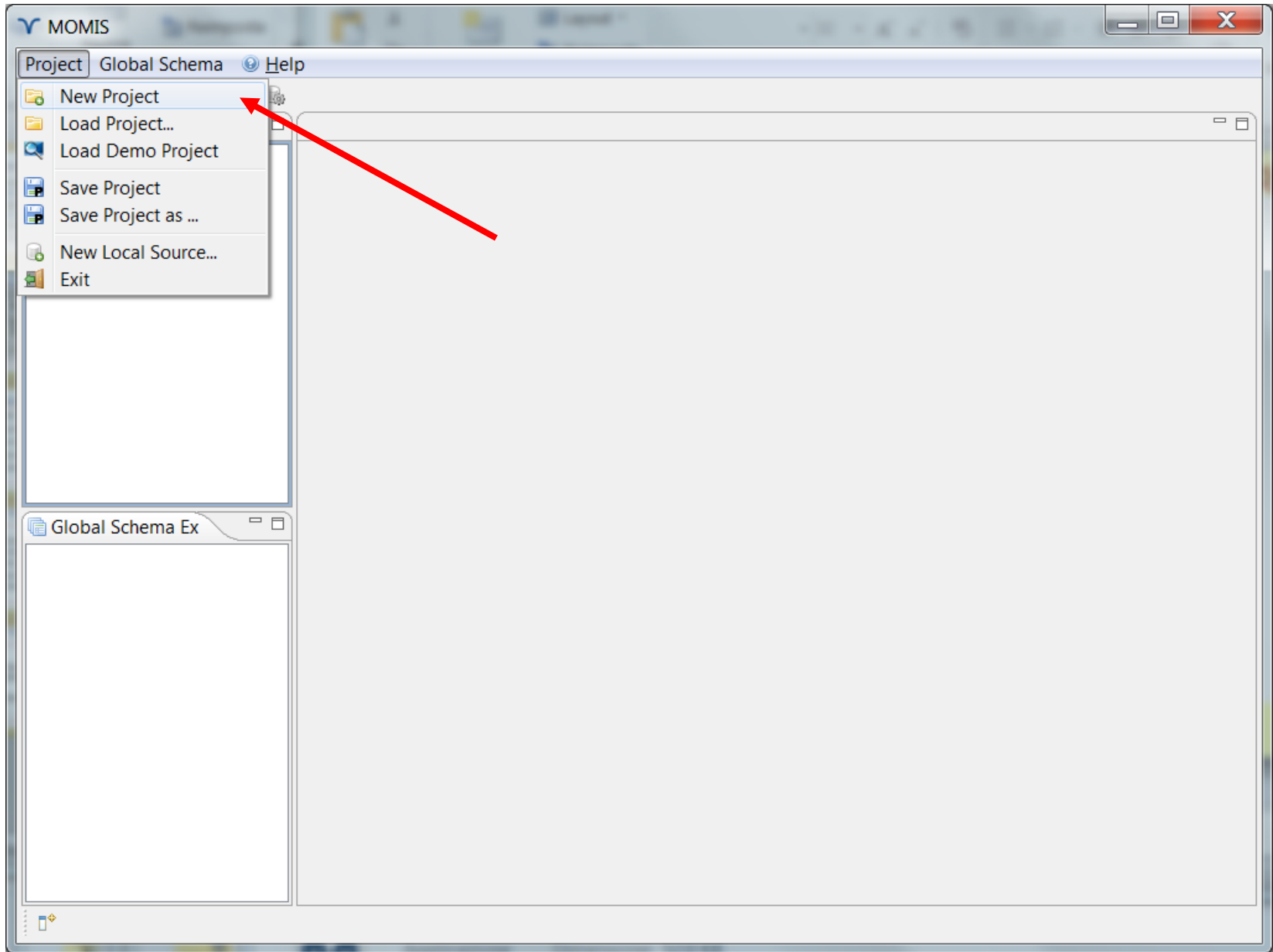
TEST 1

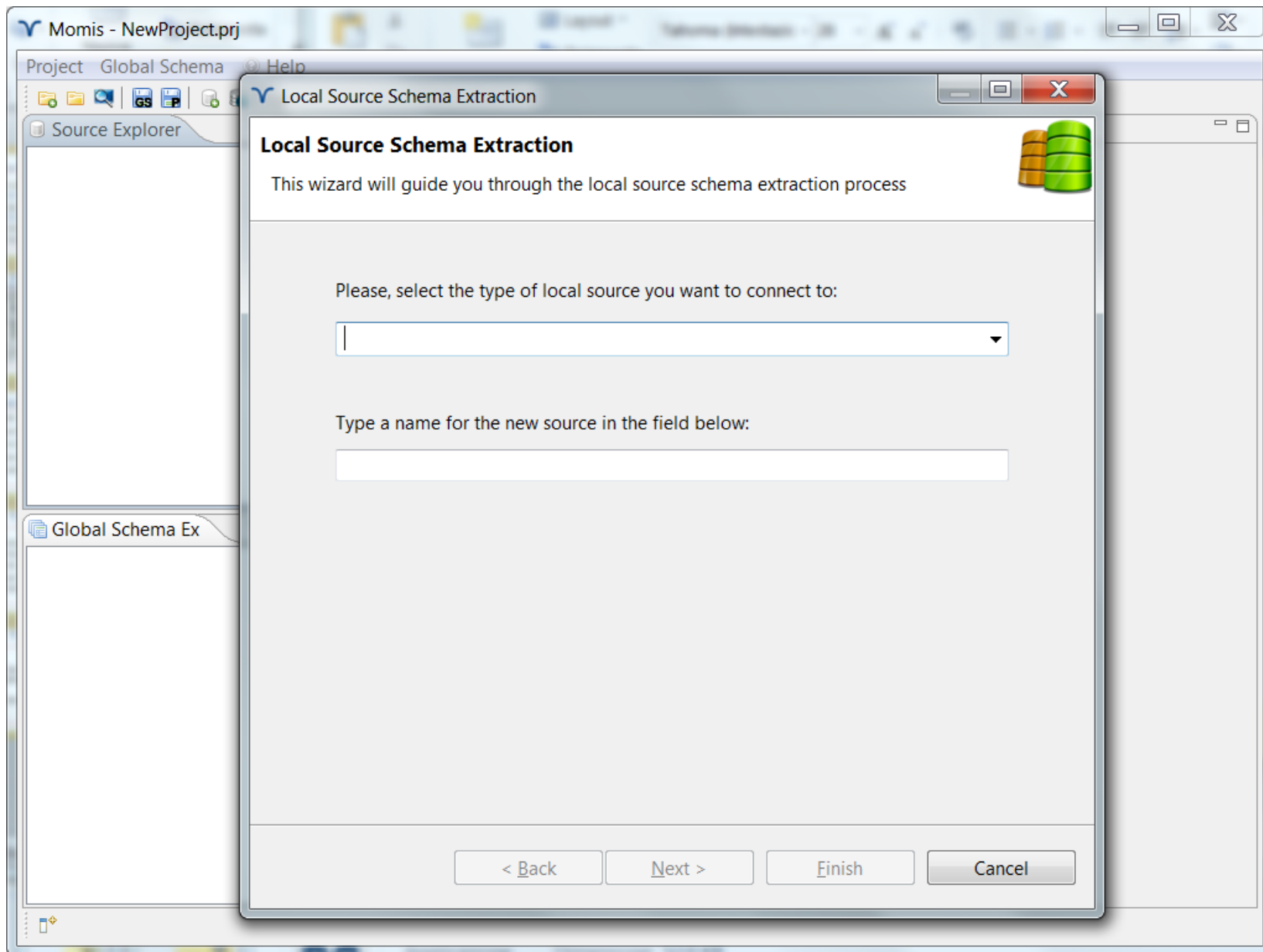
1. Download the workspace
2. Download the sources
3. Run MOMIS
4. Start a new integration project

- Close MOMIS
- Access the following folder on IMEX:
`W:\laurapo\LaboratorioMomis`
- Copy the contents of
`W:\laurapo\LaboratorioMomis\workspace`
- in your local workspace
`C:\Temp\workspace\momislab`

- Access the following folder on IMEX:
W: \laurapo\LaboratorioMomis
- Copy the folder `olivetree` in your local folder
MOMIS_1-2_win32_x86_JRE\MOMIS\test\sourcesDb\
- Run MOMIS

Create a new project






Set up the connection to an Excel file

- Set up the parameters
 - **type:** Microsoft Excel file
 - **name:** Population
 - **file:**
C:\Temp\workspace\momislab\csv\Population0liveTree.xlsx
- Select the table Population0liveTree

Local Source Schema Extraction

This wizard will guide you through the local source schema extraction process



Please, select the type of source

Microsoft Office Excel

Type a name for the new source

Population

< Back

Local Source Schema Extraction

Please click 'Next' to get the source schema

Microsoft Office Excel

Absolute file path

C:\Temp\workspace\momislab\csv\Population

Advanced Parameters

Column Name Line Number

0

Skip Empty Columns

false

< Back

Next >

Local Source Schema Extraction

Click 'Finish' to add the schema of the new source with the selected table(s) and attribute(s) in the project

| Tables/Views and Attributes | |
|-------------------------------------|---------------------|
| <input checked="" type="checkbox"/> | PopulationOliveTree |
| <input checked="" type="checkbox"/> | Age_Group |
| <input checked="" type="checkbox"/> | Year |
| <input checked="" type="checkbox"/> | Sex |
| <input checked="" type="checkbox"/> | City |
| <input checked="" type="checkbox"/> | Population_Register |
| <input checked="" type="checkbox"/> | Population_Mondo |

Select All Select None Data Preview

< Back Next > Finish Cancel

Set up the connection to the first H2 database

- Set up the parameters
 - **type:** Jdbc Source
 - **name:** Corsica
 - **JDBC Driver class:** org.h2.Driver
 - **user:** sa
 - **no password**
 - **Conn String:**
jdbc:h2:test/sourcesDb/olivetree/MelanomaCORSIKA/
db
 - **Upload Driver:** C:\Temp\workspace\momislab\lib\h2-
1.2.125.jar
- Select connect
- Select the table CorsicaMelanomaView

Local Source Schema Extraction

This wizard will guide you through the local source schema extraction process

Please, select the type of source

JDBC source

Type a name for the new source

Corsica

Local Source Schema Extraction

Please click 'Next' to get the schema of the selected source

Jdbc

JDBC Driver Class Name:

org.h2.Driver

Username:

sa

Connection String:

jdbc:h2:test/sourcesDb/olivetree/MelanomaCORSI

Connect

< Back Next >

Local Source Schema Extraction

Click 'Finish' to add the schema of the new source with the selected table(s) and attribute(s) in the project

| Tables/Views and Attributes | |
|-------------------------------------|-----------------------|
| <input checked="" type="checkbox"/> | CORSICA_MELANOMA_VIEW |
| <input checked="" type="checkbox"/> | ADIA |
| <input checked="" type="checkbox"/> | SESSO |
| <input checked="" type="checkbox"/> | CITTA |
| <input checked="" type="checkbox"/> | AGE_GROUP |
| <input checked="" type="checkbox"/> | N_CASES |

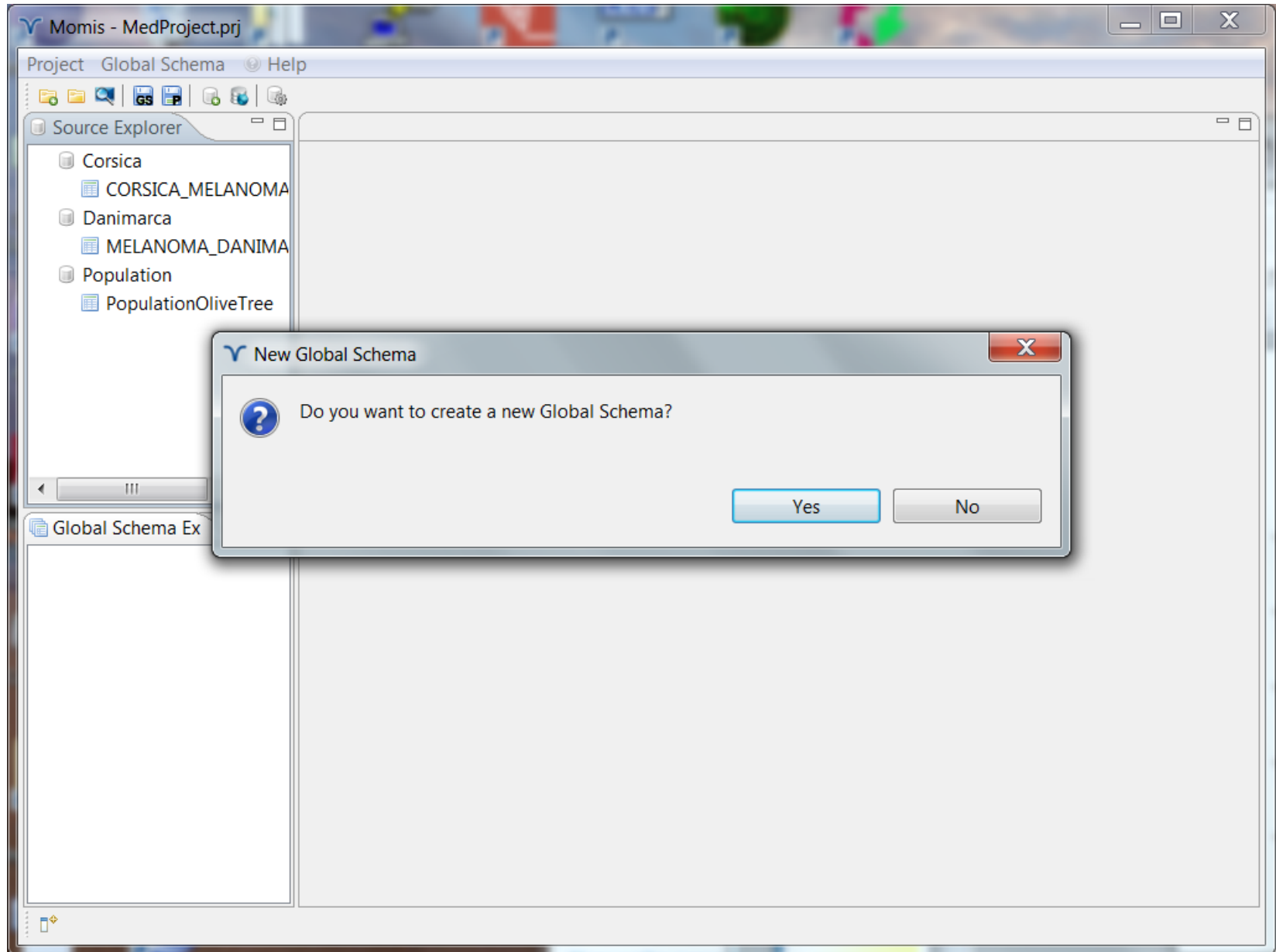
Select All Select None Data Preview

< Back Next > Finish Cancel

Set up the connection to the second H2 database

- Set up the parameters
 - **type:** Jdbc Source
 - **name:** Danimarca
 - **JDBC Driver class:** org.h2.Driver
 - **user:** sa
 - **Conn String:**
jdbc:h2:test/sourcesDb/olivetree/MelanomaD
ANIMARCA/db
- Select connect
- Select the table MelanomaDanimarcaView

Select to create a new GS



Start the integration process following the steps

Momis - MedProject.prj

Project Global Schema Help

Source Explorer

- Corsica
 - CORSICA_MELANOMA
 - Danimarca
- MELANOMA_DANIMA
- Population
- PopulationOliveTree

Global Schema Ex

- GS

Global Schema Designer: Overview

- Local Sources**
In this section it is possible to select the desired sources for the integration
nrniert
[edit section](#)
- Sources Annotation**
In this section it is possible to semantically annotate the selected sources...
[edit section](#)
- Semantic Relationships**
In this section it is possible to visualize and define inter-schema and intra-schema relations which are necessary for the clustering phase
[edit section](#)
- Mapping Refinement**
In this section it is possible to manually refine mappings automatically generated at the end of the integration project
[edit section](#)
- Test Schema**
Execute queries on the global schema

Overview Local Sources Annotation Semantic Relationships Mapping Refinement

Test 1 will go through the different steps of the integration process:

- Local source selection:
- Annotation:
 - Simple
 - Adding new terms and new meanings
- Relationship discovery:
 - Automatic computation
 - Adding new relationships
 - Compute inferred relationships
- Cluster generation and Mapping refinement
- Query the final GS

Step1 - Add all the 3 sources to the GS

The screenshot displays the 'Global Schema Designer: Local Sources' window. On the left, the 'Source Explorer' shows a tree view with 'Corsica' expanded, containing 'CORSICA_MELANOMA'. A right-click context menu is open over 'CORSICA_MELANOMA', with the option 'Add selected source to the Global Schema' highlighted by a red arrow. Below the menu, a text box explains: 'Right click on a source on the Source Explorer view to include it in your global schema! For each selected source you can have more information in the section below!' and a 'Remove Source' button. The bottom section, 'Source Details: ODLI3 Representation', lists fields: Source Name, Type, Server : PortNumber, Username, and Connection String. To the right is a table with columns: Attribute ..., Ty..., Primary..., Foreign Key. The bottom navigation bar includes tabs for Overview, Local Sources, Annotation, Semantic Relationships, and Mapping Refinement.

Project Global Schema Help

Source Explorer

- Corsica
 - CORSICA_MELANOMA
- Danimarca
 - MELAN
- Population
 - PopulationOliveTree

Global Schema Designer: Local Sources

Overview Annotation

Right click on a source on the Source Explorer view to include it in your global schema!

For each selected source you can have more information in the section below!

Remove Source

Source Details: ODLI3 Representation

Source Name _
Type _
Server : PortNumber _
Username _
Connection String _

| Attribute ... | Ty... | Primary... | Foreign Key |
|---------------|-------|------------|-------------|
| | | | |
| | | | |
| | | | |
| | | | |

Overview Local Sources Annotation Semantic Relationships Mapping Refinement

Proceed with the second step: Annotation

Momis - MedProject.prj

Project Global Schema Help

Source Explorer

- Corsica
 - CORSICA_MELANOMA
- Danimarca
 - MELANOMA_DANIMA
- Population
 - PopulationOliveTree

Global Schema Ex

- GS

Global Schema Designer: Local Sources

Overview Overview **Annotation**

Local Sources Selection

Corsica
Danimarca
Population

Right click on a source on the Source Explorer view to include it in your global schema!

For each selected source you can have more information in the section below!

Remove Source

Source Details: ODLI3 Representation

Source Name _
Type _
Server : PortNumber _
Username _
Connection String _

| Attribute ... | Ty... | Primary... | Foreign Key |
|---------------|-------|------------|-------------|
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

Overview Local Sources Annotation Semantic Relationships Mapping Refinement

Step 2 - Annotation

Project Global Schema Help

Source Explorer

- Corsica
 - CORSICA_MELANOMA
- Danimarca
 - MELANOMA_DANIMA
- Population
 - PopulationOliveTree

Global Schema Ex

- GS

Global Schema Designer: Annotation

Sources Overview Semantic Relationships

Sources

- Corsica
 - CORSICA_MELANOMA**
 - ADIA**
 - AGE_GROUP
 - CITTA
 - N_CASES
 - SESSO
- Danimarca
 - MELANOMA_DANIMA
 - AGE_GROUP
 - CITTA
 - DIAG_YEAR
 - N_CASES
 - SESSO
- Population

Automatic Annotation
Delete All Annotations

Filter Tree Nodes

- View Not Annotated Elemer
- View Annotated Elements

Overview Local Sources Annotation Semantic Relationships Mapping Refinement

Annotation about: ADIA

Type: short

List of the selected senses:

| Base F... | Pos | Sens... | Gloss |
|-----------|-----|---------|-------|
| | | | |
| | | | |
| | | | |
| | | | |

Add annotation Remove

Select annotations for existing lemma

Momis - MedProject.prj

Project Global Schema Help

Source Explorer

- Corsica
 - CORSICA_MELANOMA
- Danimarca
 - MELANOMA_DANIMA
- Population
 - PopulationOliveTree

Global Schema Ex

- GS

Global Schema Designer: Annotation

Sources Overview Semantic Relationships

Sources

- Corsica
 - CORSICA_MELANOMA
 - ADIA
 - AGE_GROUP
 - CITTA
 - N_CASES
 - SESSO
- Danimarca
 - MELANOMA_DANIMA
 - AGE_GROUP
 - CITTA
 - DIAG_YEAR
 - N_CASES
 - SESSO
- Population

Automatic Annotation
Delete All Annotations

Filter Tree Nodes

- View Not Annotated Elemer
- View Annotated Elements

Annotation about: ADIA

Type: short

List of the selected senses:

| Base F... | Pos | Sens... | Gloss |
|-----------|-----|---------|-------|
| | | | |
| | | | |
| | | | |
| | | | |

Add annotation Remove

Overview Local Sources Annotation Semantic Relationships Mapping Refinement

Select annotations for existing lemma

Momis - MedProject.prj

Project Global Schema

Source Explorer

- Corsica
 - CORSICA_MELANOMA_VIEW
- Danimarca
 - MELANOMA_DANIMARCA
- Population
 - PopulationOliveTree

Global Schema Ex

- GS

Annotation about: Corsica>CORSICA_MELANOMA_VIEW>AGE_GROUP

Select one base form from the list (or add a new one):

Add base form

Open WordNet Extender

Remove base form

Select one or more senses:

- a group of people having approximately the same age

Close

Automatic Annotation

Delete All Annotations

Add annotation

Remove

Filter Tree Nodes

- View Not Annotated Elements
- View Annotated Elements

Overview Local Sources Annotation Semantic Relationships Mapping Refinement

Change base form for lemma that are not present in WordNet

The screenshot shows a software interface with two overlapping dialog boxes. The background window is titled 'Momis - MedProject.prj' and shows a 'Source Explorer' with a tree view containing folders like 'Corsica', 'Danimarca', and 'Population'. The top dialog box is titled 'Annotation about: Corsica>CORSICA_MELANOMA_VIEW>CITTA'. It contains a text input field with 'citta', a red arrow pointing to it, and a red warning message: 'Attention: the base form does not exist. You can add a new lemma and map it on an existing synset or add a new synset and map on it the new lemma'. Below the input field are buttons for 'Add base form', 'Open WordNet Extender', and 'Remove base form'. The bottom dialog box is also titled 'Annotation about: Corsica>CORSICA_MELANOMA_VIEW>CITTA'. It contains a text input field with 'city', a green message: 'Lemma city exists into WordNet DataBase', and a red arrow pointing to the 'Add base form' button. Below this are buttons for 'Open WordNet Extender' and 'Remove base form'. At the bottom of this dialog is a section 'Select one or more senses:' with a list of three senses, the first of which is checked. A red arrow points to the first sense. A 'Close' button is at the bottom right of the dialog.

or...Add new lemma to WordNet

The screenshot shows the MedProject application interface. A dialog box titled "Annotation about: Corsica>CORSICA_MELANOMA_VIEW>SESSO" is open. The dialog contains the following elements:

- A text input field containing "sesso".
- An "Add base form" button.
- A red text warning: "Attention: the base form does not exist. You can add a new lemma and map it on an existing synset or add a new synset and map on it the new lemma".
- An "Open WordNet Extender" button, which is highlighted with a red arrow.
- A "Remove base form" button.
- A "Select one or more senses:" section with an empty list box.
- A "Close" button at the bottom right.

The background application window shows a Source Explorer with a tree view containing nodes like Corsica, CORSICA_MELANOMA_VIEW, Danimarca, MELANOMA_DANIMARCA_VIEW, Population, and PopulationOliveTree. The Global Schema Explorer shows a node for GS. The bottom of the application has a toolbar with buttons for "Automatic Annotation", "Delete All Annotations", "Add annotation", and "Remove". There are also checkboxes for "Filter Tree Nodes" (View Not Annotated Elemer, View Annotated Elements) and a tabbed interface with tabs for "Overview", "Local Sources", "Annotation", "Semantic Relationships", and "Mapping Refinement".

WordNet Extender

Lemma and Synsets

Click 'Finish' to map the lemma on the selected synset(s)

Lemma and Syntactic Category

Lemma: Syntactic Category: **noun** verb adjectiv adverb

Lemma "sesso" in the selected syntactic category does not exist into the WordNet DataBase [View Hierarchy Graph](#)

Synsets

Search synsets by inserting similar lemmas or search synsets by gloss keywords, or insert a

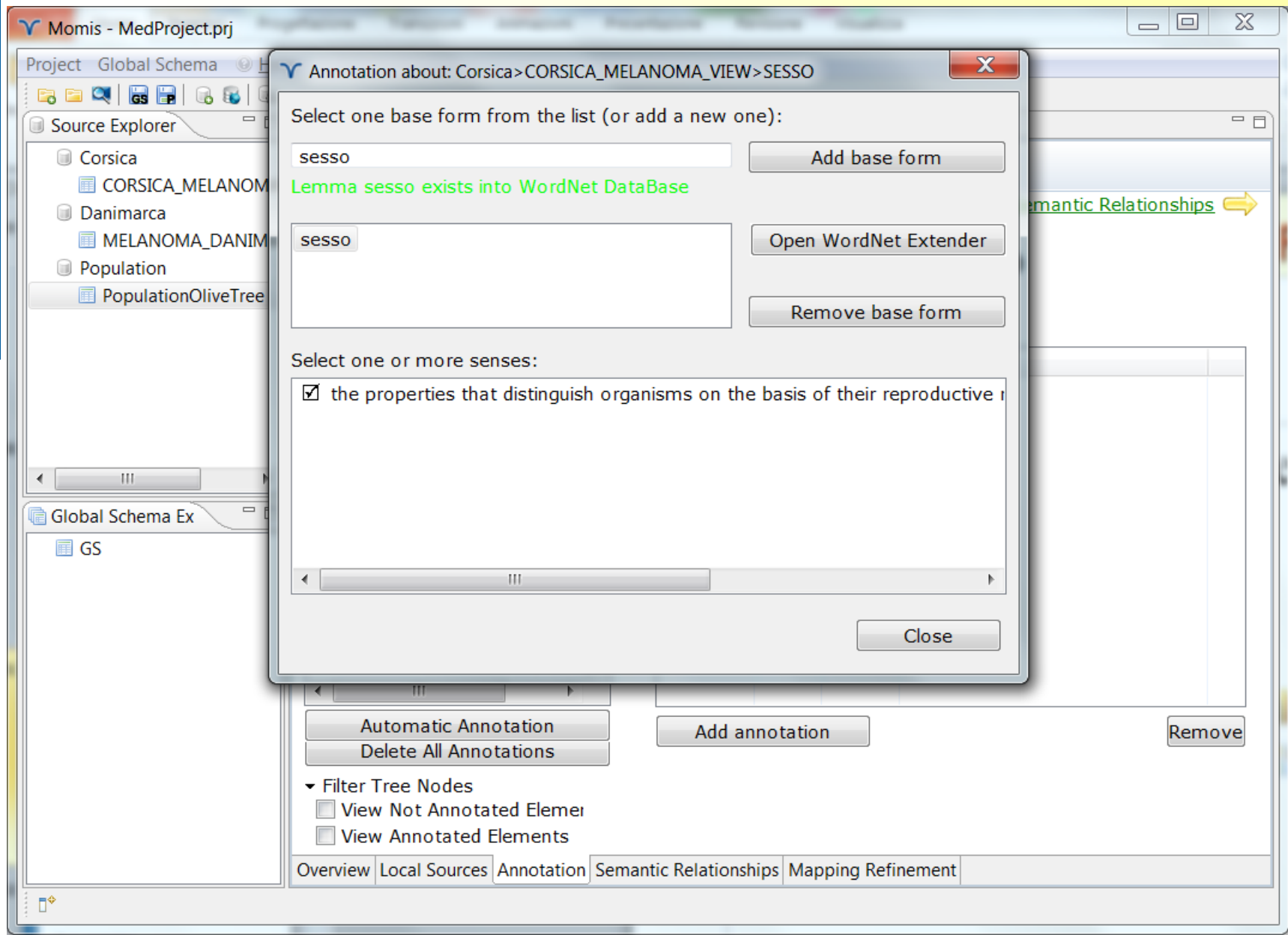
Insert one or more similar lemmas. The search will be Insert one or more gloss keywords. The search will be

Select one or more synsets from below

| Lemma(s) | Gloss | Exte... |
|-------------------------------|--|-----------|
| gender, grammatical_gen... | a grammatical category in inflected ... | wn |
| sex, gender, sexuality | the properties that distinguish orga... | wn |
| gender_agreement | agreement in grammatical gender b... | wn |
| gender_identity | your identity as it is experienced wit... | wn |
| gender_role | the overt expression of attitudes th... | wn |

Found 5 synset(s) Insert a new synset

< Back Next > Finish Cancel



...also new meanings can be specify

WordNet Extender

Lemma and Synsets

Click 'Next' to insert a new synset and map it on the inserted lemma

Lemma and Syntactic Category

Lemma
Insert a new lemma or an existing

Syntactic Category
Select lemma's syntactic category
 noun
 verb adjectiv adverb

Lemma "diagnosis_year" in the selected syntactic category does not exist into the WordNet DataBase

Synsets

Search synsets by inserting similar lemmas or search synsets by gloss keywords, or insert a

Insert one or more similar lemmas. The search will be

Insert one or more gloss keywords. The search will be

Select one or more synsets from below

| Lemma(s) | Gloss | Exte... | |
|----------|-------|---------|--|
| | | | |

Insert a new synset

Define the new meaning

WordNet Extender

New Synset

⊗ Please press " Search Hypernym Synset" to make a search among existing synsets and select a hypernym for the new synset

▼ New Synset

Lemma Name: diagnosis_year **Syntactic Category:**

Synset

Insert the gloss for the new synset

the year of the diagnosis of a disease

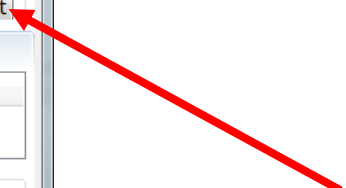
Press "Search Hypernym Synset" to find

▼ Relationships

| Relationship... | Lemma(s) | Gloss | Ext... |
|-----------------|----------|-------|--------|
| | | | |

Relationships that you are going to add into the WordNet Database

< Back Next > Finish Cancel



Select the correct hypernyms

WordNet Extender

Search Hypernym Synset

Search synsets by inserting similar lemmas or search synsets that contain in the gloss the

Select one or more keywords from the list

disease
diagnosis
year

Insert one or more gloss keywords. The search will be

year Search

Insert one or more similar lemmas. The search will be

year Search

Select a synset from below

Found 17 synset(s)

| Lemma(s) | Gloss | Exte |
|-----------------------|---|------|
| yr, year, twelvemonth | a period of time containing 365 (or 3... | wn |
| year | a period of time occupying a regular p... | wn |
| year | the period of time that it takes for a p... | wn |
| class, year | a body of students who graduate tog... | wn |
| year end | the end of a calendar year; "he had t... | wn |

View Hypernym Graph

OK

so that the new meaning is connected in the WordNet hypernym graph

WordNet Extender

New Synset

Click 'Finish' to insert the new synset, the new relationships and map the inserted synset on the inserted lemma

▼ **New Synset**

Lemma Name: `diagnosis_year` **Syntactic Category:**

Synset

Insert the gloss for the new synset

the year of the diagnosis of a disease

Search Hypernym Synset

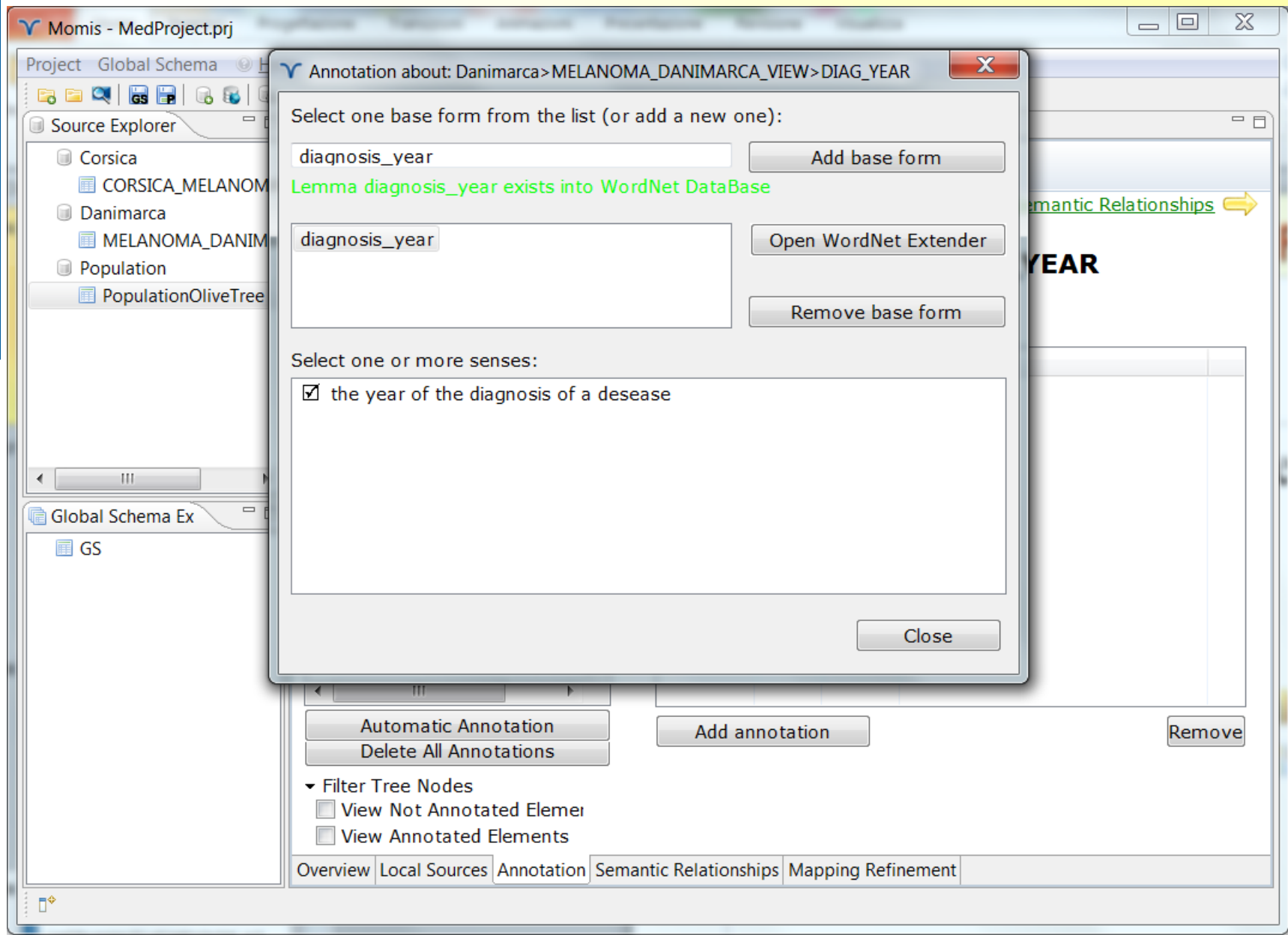
▼ **Relationships**

| Relationship... | Lemma(s) | Gloss | Ext... |
|-----------------|----------|------------------------------|--------|
| is a Hypony... | year | a period of time occupyin... | wn |

Relationships that you are going to add into the WordNet Database

[diagnosis_year] is a Hyponym of [year]
[year] is a Hypernym of [diagnosis_year]

< Back Next > **Finish** Cancel



Once the annotation is finished you can proceed to step 3 – the computation of semantic relationship

Project Global Schema Help

Source Explorer

- Corsica
 - CORSICA_MELANOMA
- Danimarca
 - MELANOMA_DANIMA
- Population
 - PopulationOliveTree

Global Schema Ex

- GS

Global Schema Designer: Semantic Relationships

← Annotation Overview → Mapping

Compute Structural and Lexical Rels. Compute Inferred Rels.

| Producer | Source | Type | Destination |
|----------|--------|------|-------------|
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

Add Delete Delete All

► Filter results by...

Overview Local Sources Annotation Semantic Relationships Mapping Refinement

You might also add new relationships

Momis - MedProject.prj

Project Global Schema Help

Source Explorer

- Corsica
 - CORSICA_MELANOMA
- Danimarca
 - MELANOMA_DANIMA
- Population
 - PopulationOliveTree

Global Schema Designer: Semantic Relationships

Annotation Overview Mapping

Compute Structural and Lexical Rels. Compute Inferred Rels.

| Producer | Source | Type | Destination |
|----------|--------------------------------|------|-------------------------------------|
| Lexical | Population.PopulationOliveTree | rt | Corsica.CORSICA_MELANOMA_VI... |
| Lexical | Corsica.CORSICA_MELANOMA_VI... | syn | Danimarca.MELANOMA_DANIMAR... |
| Lexical | Population.PopulationOliveTree | rt | Danimarca.MELANOMA_DANIMAR... |
| Lexical | Corsica.CORSICA_MELANOMA_VI... | syn | Danimarca.MELANOMA_DANIMAR... |
| Lexical | Corsica.CORSICA_MELANOMA_VI... | syn | Danimarca.MELANOMA_DANIMAR... |
| Lexical | Corsica.CORSICA_MELANOMA_VI... | syn | Population.PopulationOliveTree.A... |
| Lexical | Danimarca.MELANOMA_DANIMAR... | syn | Population.PopulationOliveTree.A... |
| Lexical | Population.PopulationOliveTree | rt | Population.PopulationOliveTree.A... |
| Lexical | Corsica.CORSICA_MELANOMA_VI... | syn | Population.PopulationOliveTree.City |
| Lexical | Danimarca.MELANOMA_DANIMAR... | syn | Population.PopulationOliveTree.City |
| Lexical | Danimarca.MELANOMA_DANIMAR... | nt | Population.PopulationOliveTree.Year |

Add Delete

Filter results by...

Overview Local Sources Annota

Add Thesaurus Relationship

Please add a new relationship using the following fields:

Source

- GS
 - Corsica
 - CORSICA_MELANOMA_VIE
 - ADIA
 - AGE_GROUP
 - CITTA
 - N_CASES
 - SESSO
 - Danimarca
 - Population

Destination

- GS
 - Corsica
 - Danimarca
 - MELANOMA_DANIMARCA_V
 - AGE_GROUP
 - CITTA
 - DIAG_YEAR
 - N_CASES
 - SESSO
 - Population

Rel Type

SYN

Add Rel Close

in the end, compute the inferred relationships

Momis - MedProject.prj

Project Global Schema Help

Source Explorer

- Corsica
 - CORSICA_MELANOMA
- Danimarca
 - MELANOMA_DANIMA
- Population
 - PopulationOliveTree

Global Schema Ex

- GS

Global Schema Designer: Semantic Relationships

Annotation Overview Mapping

Compute Structural and Lexical Rel. Compute Inferred Rel.

| Producer | Source | Type | Destination |
|----------|--------------------------------|------|-------------------------------------|
| Lexical | Population.PopulationOliveTree | rt | Corsica.CORSICA_MELANOMA_VI... |
| Lexical | Corsica.CORSICA_MELANOMA_VI... | syn | Danimarca.MELANOMA_DANIMAR... |
| Lexical | Population.PopulationOliveTree | rt | Danimarca.MELANOMA_DANIMAR... |
| Lexical | Corsica.CORSICA_MELANOMA_VI... | syn | Danimarca.MELANOMA_DANIMAR... |
| Lexical | Corsica.CORSICA_MELANOMA_VI... | syn | Danimarca.MELANOMA_DANIMAR... |
| Lexical | Corsica.CORSICA_MELANOMA_VI... | syn | Population.PopulationOliveTree.A... |
| Lexical | Danimarca.MELANOMA_DANIMAR... | syn | Population.PopulationOliveTree.A... |
| Lexical | Population.PopulationOliveTree | rt | Population.PopulationOliveTree.A... |
| Lexical | Corsica.CORSICA_MELANOMA_VI... | syn | Population.PopulationOliveTree.City |
| Lexical | Danimarca.MELANOMA_DANIMAR... | syn | Population.PopulationOliveTree.City |
| Lexical | Danimarca.MELANOMA_DANIMAR... | nt | Population.PopulationOliveTree.Year |

Add Delete Delete All

Filter results by...

Overview Local Sources Annotation Semantic Relationships Mapping Refinement

Step 4 – generation of clusters

The screenshot shows the 'Global Schema Designer: Mapping Refinement' window. The interface includes a 'Source Explorer' on the left with a tree view containing 'Corsica', 'Danimarca', and 'Population'. The main area is titled 'Global Schema Designer: Mapping Refinement' and has tabs for 'Semantic...ionships' and 'Overview'. Under 'Clustering Settings', there are seven sliders: 'Relation' (100), 'Relation' (80), 'Relation RT:' (50), 'Affinity Threshold' (50), 'Clustering' (50), 'Naming' (50), and 'Structural' (50). A 'Presets' section on the right has radio buttons for 'Default', 'Preset 1', 'Preset 2', and 'Manual'. Below the sliders are 'Restore' and 'Generate Clusters' buttons. The 'Mapping Refinement' section at the bottom has three empty boxes: 'Global Source', 'Mapped Elements', and 'Unmapped Elements'. A tab bar at the bottom shows 'Overview', 'Local Sources', 'Annotation', 'Semantic Relationships', and 'Mapping Refinement'.

The screenshot displays the 'Global Schema Designer: Mapping Refinement' window. The interface is divided into several sections:

- Source Explorer:** Located on the left, it shows a tree view of sources including Corsica, CORSICA_MELANOMA, Danimarca, MELANOMA_DANIMA, Population, and PopulationOliveTree.
- Global Schema Ex:** Below the Source Explorer, it shows a single source named 'GS'.
- Clustering Settings:** This section contains seven sliders for adjusting parameters: Relation (100), Relation (80), Relation RT: (50), Affinity Threshold (50), Clustering (50), Naming (50), and Structural (50). A 'Presets' panel on the right offers options for Default, Preset 1, Preset 2, and Manual. 'Restore' and 'Generate Clusters' buttons are also present.
- Mapping Refinement:** This section is divided into three panes:
 - Global Source:** A tree view showing 'globalSource' with sub-items 'CORSICA_MELANOMA_VIEW', 'MELANOMA_DANIMARCA_VIEW', and 'PopulationOliveTree'.
 - Mapped Elements:** A list of elements mapped from local sources, including Corsica, CORSICA_MELANOMA_VIEW, Danimarca, MELANOMA_DANIMARCA_VIE, Population, and PopulationOliveTree.
 - Unmapped Elements:** An empty list for elements not yet mapped.

At the bottom, a tabbed interface shows 'Overview', 'Local Sources', 'Annotation', 'Semantic Relationships', and 'Mapping Refinement' (the active tab).

MOMIS - MedProject.prj

Project Global Schema Help

Source Explorer

- Corsica
 - CORSICA_MELANOMA
- Danimarca
 - MELANOMA_DANIMA
- Population
 - PopulationOliveTree

Global Schema Ex

- GS

Global Schema Designer: Overview

Local Sources

In this section it is possible to select the desired sources for the integration project

[edit section](#)

Sources Annotation

In this section it is possible to semantically annotate the selected sources...

[edit section](#)

Semantic Relationships

In this section it is possible to visualize and define inter-schema and intra-schema relations which are necessary for the clustering phase

[edit section](#)

Mapping Refinement

In this section it is possible to manually refine mappings automatically generated at the end of the integration project

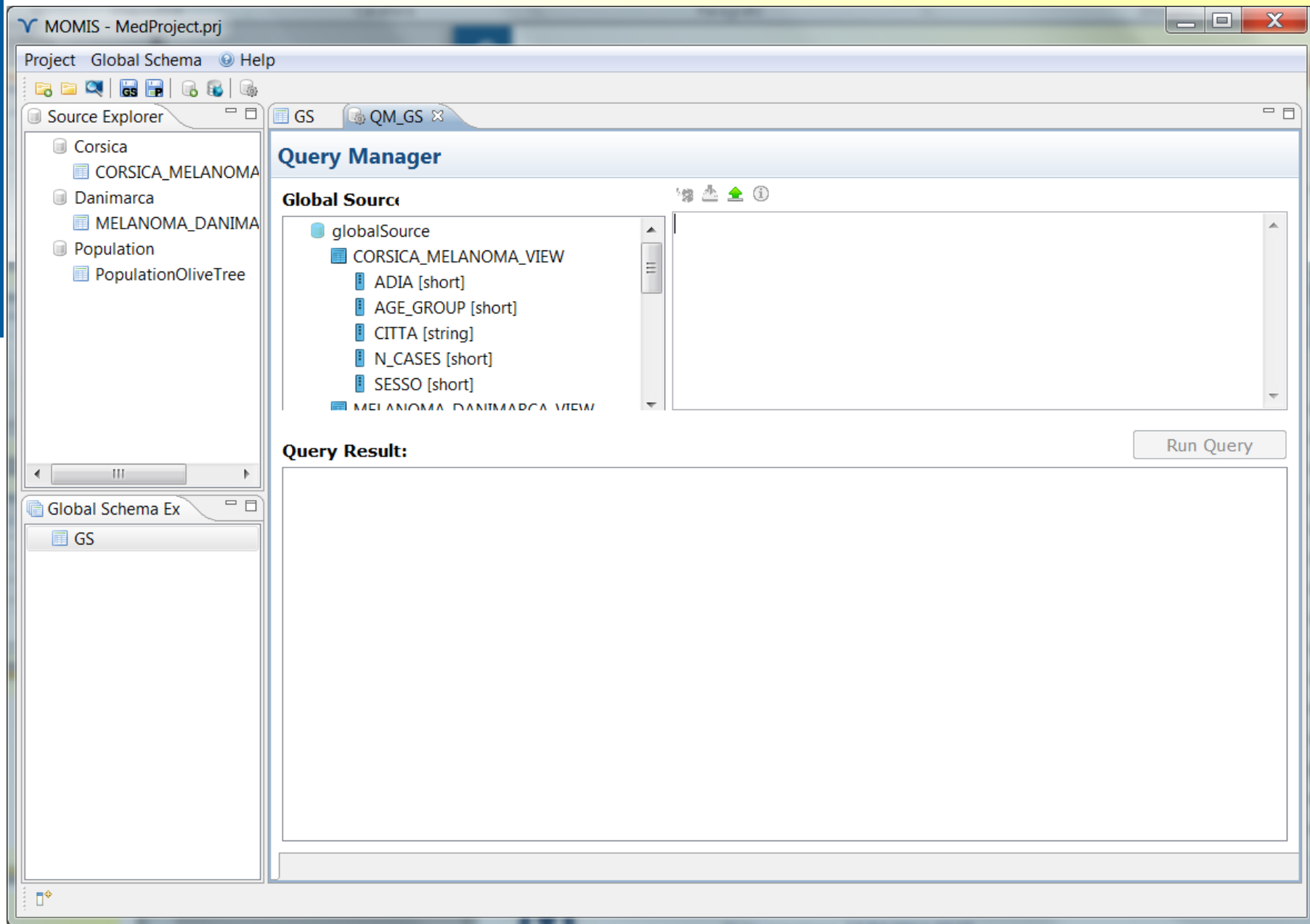
[edit section](#)

Test Schema

Execute queries on the global schema

[Launch Query Manager](#)

Overview Local Sources Annotation Semantic Relationships Mapping Refinement

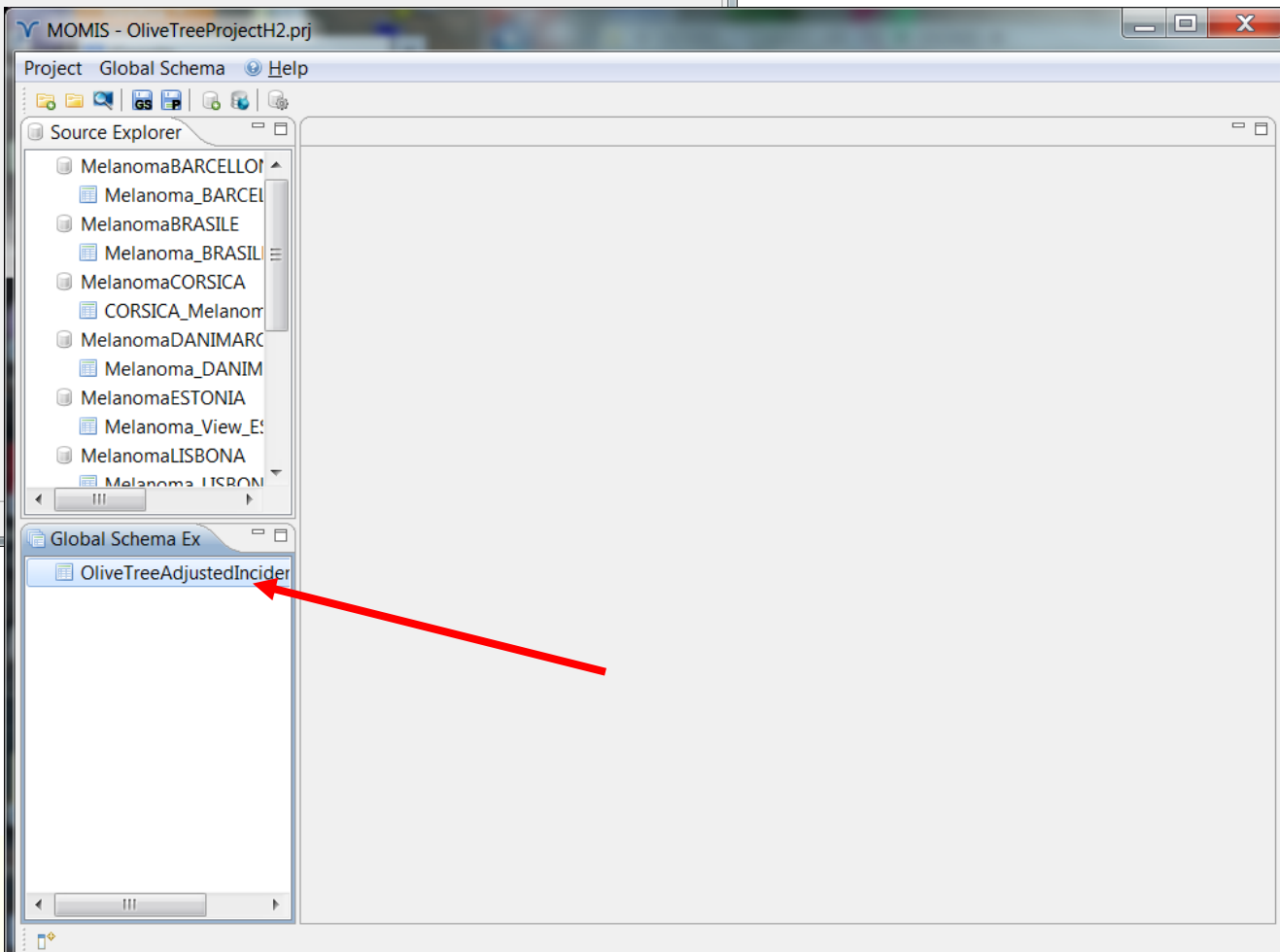
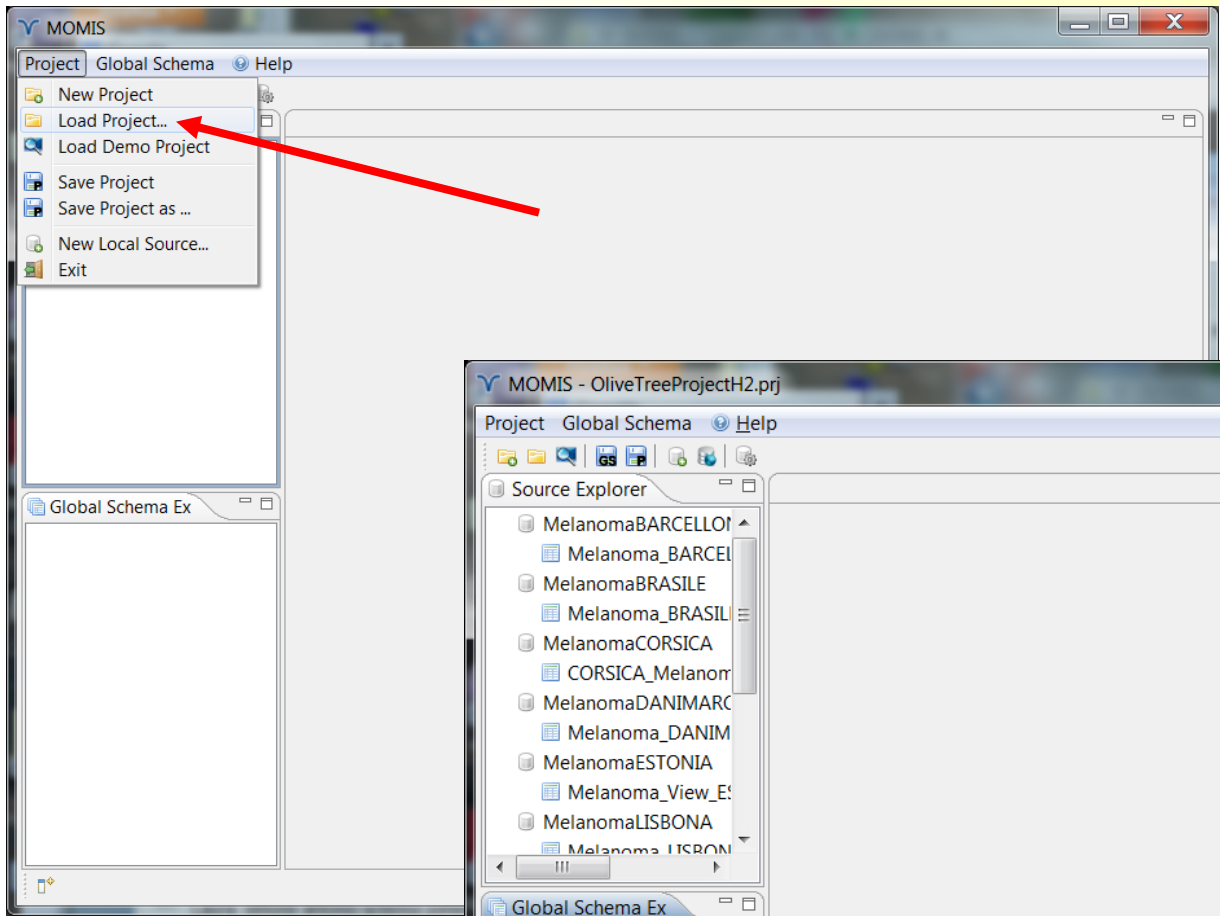


TEST 2

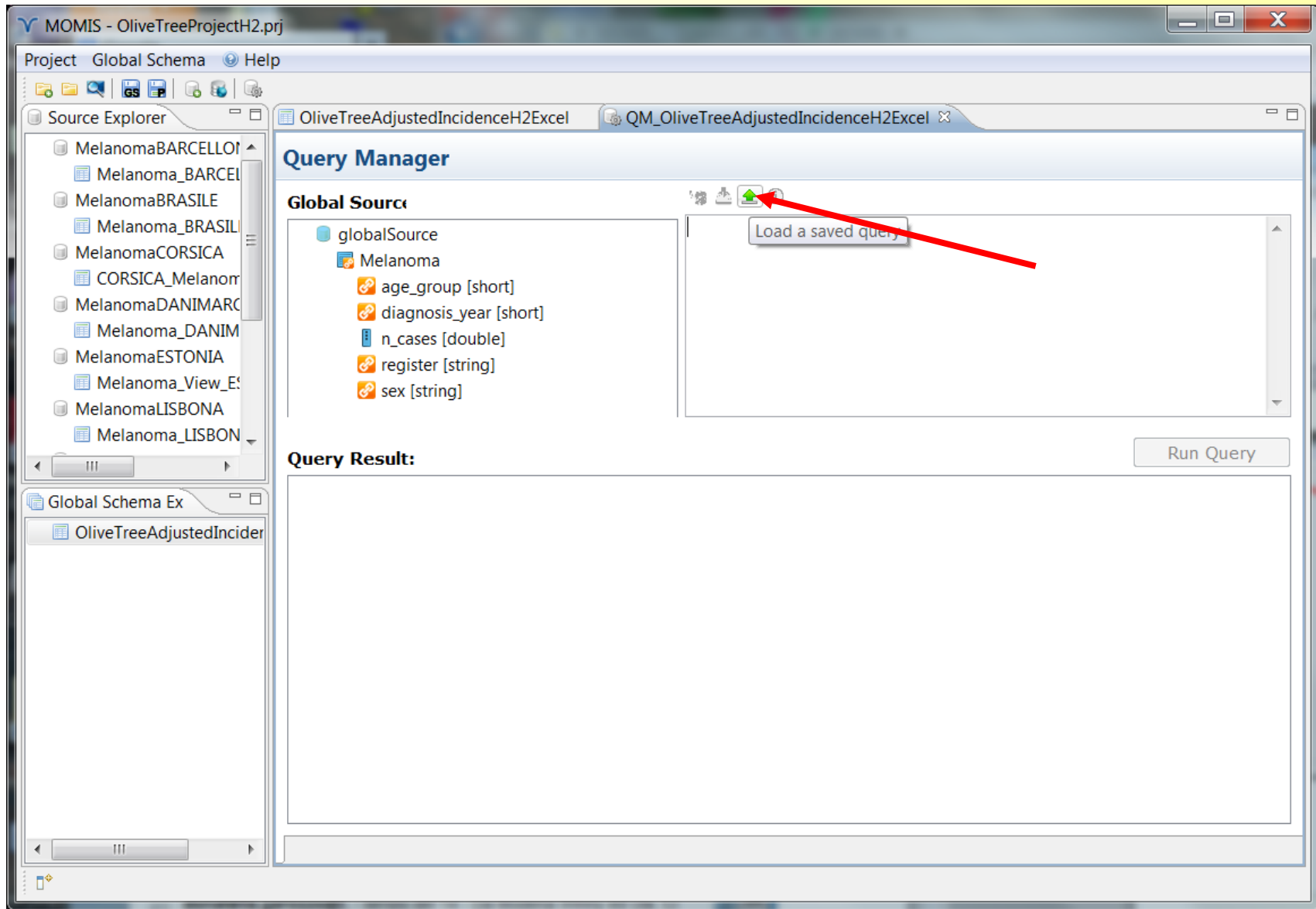
- Close MOMIS
- Access the folder `hsq1` on IMEX and replace the existing one in the Momis folder

```
MOMIS_1-2_win32_x86_JRE\MOMIS\hsq1\
```

- Run MOMIS
- Load the project `OliveTreeProjectH2.prj`
- Explore the annotations and the calculated relationships
- Explore the mapping: join function, resolution function, transformation function



- Launch the Query Manager
- Select and execute some of the queries contained in
`C:\Temp\workspace\momislab\QueryOliveTree`



The query

```
AdjustedIncidenceAllRegistries.oql
```

shows the incidence of occurrence of melanomas for a given geographical area and age group.

MOMIS - OliveTreeProjectH2.prj

Project Global Schema Help

Source Explorer

- MelanomaBARCELLOI
 - Melanoma_BARCELLOI
- MelanomaBRASILE
 - Melanoma_BRASILE
- MelanomaCORSICA
 - CORSICA_Melanoma
- MelanomaDANIMARC
 - Melanoma_DANIMARC
- MelanomaESTONIA
 - Melanoma_View_ESTONIA
- MelanomaLISBONA
 - Melanoma_LISBONA

Global Schema Ex

- OliveTreeAdjustedIncidence

Query Manager

Global Source

- globalSource
 - Melanoma
 - age_group [short]
 - diagnosis_year [short]
 - n_cases [double]
 - register [string]
 - sex [string]

```
SELECT sum(n_cases) as Adjusted_Incidence, sex, diagnosis_year, register
from Melanoma
group by sex, diagnosis_year, register
order by register, diagnosis_year, sex
```

Run Query

Query Result: 44 records

| SEX | DIAGNOSIS_YEAR | REGISTER | ADJUSTED_INCIDENCE |
|-----|----------------|------------|---------------------|
| 1 | 2003 | Barcellona | 2.1451036625576383 |
| 2 | 2003 | Barcellona | 1.980371384888765 |
| 1 | 2004 | Barcellona | 3.0104425863012416 |
| 2 | 2004 | Barcellona | 4.514669767978142 |
| 1 | 2005 | Barcellona | 3.124800243756587 |
| 2 | 2005 | Barcellona | 3.3566675326219886 |
| 1 | 2003 | Brasile | 0.29692047894580265 |
| 2 | 2004 | Brasile | 0.6056476644711939 |
| 1 | 2005 | Brasile | 0.7347276887308196 |
| 2 | 2005 | Brasile | 0.6236927292753149 |
| 1 | 2003 | Corsica | 3.84550207254518 |
| 2 | 2003 | Corsica | 5.508822676822518 |

The formula is expressed in the resolution function

MOMIS - OliveTreeProjectH2.prj

Project Global Schema

Source Explorer

- MelanomaBARCELLONA
- Melanoma_BRACEL
- MelanomaBRASILE
- Melanoma_BRASIL
- MelanomaCORSICA
- CORSICA_Melanom
- MelanomaDANIMARCA
- Melanoma_DANIM
- MelanomaESTONIA
- Melanoma_View_ES
- MelanomaLISBONA
- Melanoma_LISBON

Global Schema Ex

- OliveTreeAdjustedIncid

Resolution Function about: n_cases

Function editor

```
(intToDouble(coalesce({MelanomaBRASILE.Melanoma_BRASILE_View.n_cases},{MelanomaLISBONA.Melanoma_LISBONA_View.n_cases},{MelanomaESTONIA.Melanoma_View_ESTONIA.n_cases},{MelanomaLONDRA.Melanoma_View.n_cases},{MelanomaBARCELLONA.Melanoma_BARCELLONA_View.n_cases},{MelanomaSCOZIA.Melanoma_SCOZIA_View.n_cases},{MelanomaDANIMARCA.Melanoma_DANIMARCA_View.n_cases},{MelanomaVALENCIA.Melanoma_VALENCIA_View.n_cases})))
```

Function

Category: all

- Coalesce
- IF function
- String Concatenation

Transformation functions/Local Attril

- {MelanomaLONDRA.Melanoma_View.n_cases}
- {MelanomaCORSICA.CORSICA_View.n_cases}
- {MelanomaBRASILE.Melanoma_View.n_cases}
- {Population.PopulationOliveTreeAdjustedIncidence}
- {MelanomaSCOZIA.Melanoma_View.n_cases}
- {MelanomaDANIMARCA.Melanoma_View.n_cases}
- {MelanomaESTONIA.Melanoma_View.n_cases}
- {MelanomaLISBONA.Melanoma_View.n_cases}
- {MelanomaVALENCIA.Melanoma_View.n_cases}
- {MelanomaBARCELLONA.Melanoma_View.n_cases}

Help

coalesce({f(GA.L1)},{f(GA.L2)})

The coalesce function returns the value of the first of its input parameters that is not NULL.

Save Close

n_cases [double]

- register [string]
- sex [string]

MelanomaDANIMARCA

- MelanomaESTONIA
- MelanomaLISBONA
- MelanomaLONDRA
- MelanomaSCOZIA
- MelanomaVALENCIA

Overview Local Sources Annotation Semantic Relationships Mapping Refinement

Local Sources

Elements

ADDITIONAL MATERIAL

- Set of intensional and extensional relationships expressing intra-schema and inter-schema knowledge

- **Intensional Relationships**

between class and attribute names (T)

- < T_i SYN T_j > *Synonymy*
 - < T_i NT T_j > *(Narrower Term - NT)*
 - < T_i RT T_j > *(Related Term - RT)*

- **Extensional Relationships** - between classes (C)

the instances of C1 are ...

- < C1 SYN_{Ext} C2 > : ... the same instances of C2
 - < C1 NT_{Ext} C2 > : ... a subset of the instances of C2
 - < C1 DIS_{Ext} C2 > : ... disjoint from the instances of C2

- **Common Thesaurus generation:**

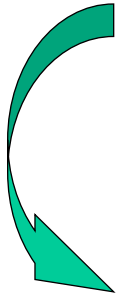
- (1) *schema derived relationships*
- (2) *lexicon derived relationships*
- (3) *designer supplied relationships*
- (4) *inferred relationships (exploiting ODB-Tools capabilities)*

Lexicon-derived Relationships

- Extracted from the WordNet lexical ontology
- In WordNet:
 - Word forms are organized in synonym set (*synset*)
 - Semantic relationships between *synset* (meanings)
 - Hyponymy (Hypernymy)
 - Meronymy
 - Correlation (between *synsets* having the same Hypernym)
- Relationships between class and attribute names are obtained using the WordNet semantic relationships as follows:
 - Synonymy \Rightarrow SYN
 - Hyponymy \Rightarrow NT
 - Meronymy and Correlation \Rightarrow RT

Annotation and Lexicon-derived Relationships

**Hyponymy
(is a kind of)**



(Narrower Term)
NT



| | Word form | | |
|---|------------------|---------------|--------------------|
| Meaning (synset) | <i>Book</i> | <i>Volume</i> | <i>Publication</i> |
| a written work or composition that has been published (printed on pages bound together) | ✗ | | |
| physical objects consisting of a number of pages bound together; "he used a large book as a doorstep" | ✗ | ✗ | |
| the amount of 3-dimensional space occupied by an object | | ● | |
| a copy of a printed work offered for distribution | | | ✗ |

Lexicon derived relationships

| | | |
|------|-----|-------------|
| Book | SYN | Volume |
| Book | NT | Publication |

- **WordNet Editor**
- If a class or attribute name has no correspondent in WordNet, the designer may add a new meaning and proper relationships to the existing meanings.
- The designer may add a new meaning (for an existing word-form or for a new one) by:
 - writing the gloss explicitly, or
 - using an existing synset chosen among a list of candidates obtained by an explicit search (using one or more keywords) or by exploiting similarity search techniques.
- The designer may add relationships for the new synset
 - Related synsets are obtained by an explicit search (using one or more keywords) or by exploiting similarity search techniques.

Common Thesaurus Generation: Other rules

■ **Schema-derived relationships**

- RT relationships derived from foreign keys in a relational schema
- NT relationships from inheritance in a object-oriented schema
- NT relationships from couples IDs and IDREFs in XML data files
- ...

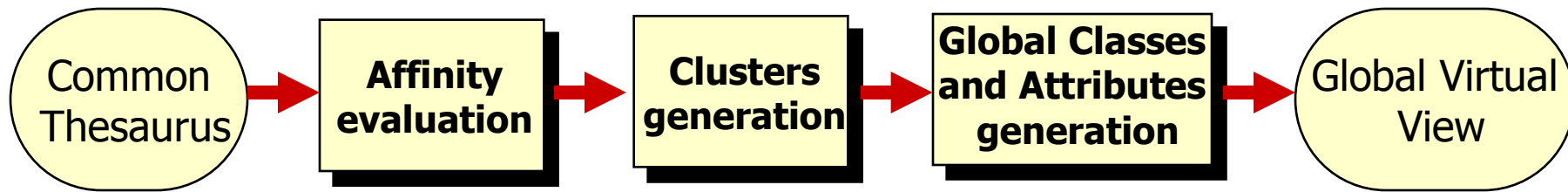
■ **Inferred relationships**

- Exploiting Description Logics techniques (by using ODB-Tools) a new set of relationships are inferred

■ **Designer supplied relationships**

- The designer can add/delete relationships to the Common Thesaurus

Global Virtual View and Mapping Table Generation



■ **GS generation :**

A global class $C=(\mathbf{L},\mathbf{GA})$ is generated for each cluster :

- **L** are the local classes of the cluster
- **GA** are the global attributes of C
 - Union of the local attributes
 - Fusion of "similar attributes" (by using the Common Thesaurus)

■ **MT generation :**

For each global class $C=(\mathbf{L},\mathbf{GA})$, a *Mapping Table* (MT) is generated, to represent the mappings between global and local attributes

- MT is a table **GAXL** : An element $MT[GA][L]$ represents the attributes of the local class L mapped into the global attribute GA .

GS and MT generation : example

- Cluster **Company={prontocomune.Azienda,fibre2fashion.Company,usawear.Company}**

- Mapping Table of C

| | prontocomune. Azienda | fibre2fashion. Company | usawear. Company |
|-------------|--------------------------|---------------------------|---------------------|
| Name | Nome | Name | CompanyName |
| Address | Indirizzo | Address | Address |
| Description | | AboutUs | Description |
| Category | Categoria | Category | |
| Phone | Telefono | Tel | Phone |

- MT generation :**
 Since " AboutUs SYN Description" is in CT, these local attributes are "fused" into to the same global attribute "Description"
- GS annotation :**
 - the name and the meaning of the class Company correspond to the name and the meaning of fibre2fashion.Company (the most general class)

- Global-As-View (**GAV**) approach:
the GS is expressed in terms of the local schemata
- **Global-as-View (GAV) mappings:**
for each global class C we define a **view** V_C over the local classes of C.
- The integration designer, supported by the Ontology Builder graphical interface, can implicitly **define** V_C by the Mapping Table refinement:
 - 1. Data Transformation** : *converting data from local source data formats into a global schema format (Conversion Functions)*
 - 2. Data Fusion** : *fusing records representing the same real-world object into a single, consistent, and clean record:*
 - 1. Object Identification**
 - 2. Data Reconciliation**

Data Transformation: THALIA Benchmark

- THALIA: **T**est **H**arness for the **A**ssessment of **L**egacy information **I**ntegration **A**pproaches

public available testbed and benchmark for information integration systems

provides over 40 downloadable sources representing University course catalog from computer science around the world

systematic classification of the different types of syntactic and semantic heterogeneities described by the twelve queries provided

- MOMIS Data Transformation can deal with all the twelve queries of the THALIA benchmark by using a simple combination of declarative translation functions and without the overhead of new code.

THALIA's Query 2 example

Q2: 'Find all database courses that meet at 1:30pm on any given day'

Complex Mappings: Mapping between the Time attribute of Carnegie Mellon University and the Times attribute of University of Massachusetts.

**Course
Mapping
Table**

| Course | Course (cmu) | Course (umb) |
|-------------|--------------|---------------------------|
| CourseTitle | CourseTitle | TitleCredits |
| Time | Time | MDTF[Time] [umb.Times] |

MDTF[Time] [umb.Times] =

```

CASE WHEN ISNUMERIC(SUBSTRING(Times, 1, 2)) = 1
  THEN CASE WHEN CAST(SUBSTRING(Times, 1, 2) AS int) > 12
    THEN CAST(CAST(SUBSTRING(Times, 1, 2) AS integer) - 12 AS nvarchar(2))
    ELSE SUBSTRING(Times, 1, 2)
    END
  + SUBSTRING(Times, 3, 4) +
    CASE WHEN CAST(SUBSTRING(Times, 7, 2) AS int) > 12
  THEN CAST(CAST(SUBSTRING(Times, 7, 2) AS integer) - 12 AS nvarchar(3))
    ELSE SUBSTRING(Times, 7, 2)
    END
  + SUBSTRING(Times, 9, 3)
END AS Time

```

Mapping Refinement: Data Conflicts Resolution

- **Data Conflicts** : the same attribute from one or more sources do not agree on its value
 - 1) **Uncertainty** : it is a conflict between a not-null value and one or more null values that describe the same attribute of the same object
 - 2) **Contradictions** : it is a conflict between two or more different not-null values that describe the same attribute of the same object.

- **Example:** data contradictions on the Phone attribute

L1

| Name | Address | Phone |
|-----------|---------------------|---------------|
| RAMOTEX | ...Mirpur-1216Dh | +390828015393 |
| CASTORAMA | ...Casalecchio (BO) | +390516113011 |

L2

| Name | Address | Phone |
|--------------|-------------------|---------------|
| RAMOTEX | ...Mirpur-1216Dh | 880-5-801466 |
| Koramsa Corp | ...Guatemala City | +502 439 6868 |

- What operator for Data Fusion ?
- **Full Join Merge** Operator
 - **Full Join** : to include into the result *all tuples of all local sources*
 - Computed on the basis of the Object Identification/Join Conditions
 - **Merge** : to perform data reconciliations
 - Application of Resolution functions (including all the results)
- In MOMIS the **Full Join Merge** is the *default* operator, i.e., is *implicitly defined* by using the Ontology Builder graphical interface (see next slide)
- The designer can change this default operator to other join operators (inner join, left/right join)

From the Mapping Table to the Full Join Merge

Mapping Table of the Global Class Hotel = {resort, hotel}

Resolution Functions

Join Conditions

Object identifier

SUM

Resolution Functions

AVG

| | resort | hotel |
|-------|--------|------------|
| Name | name | name |
| Room | rooms | hotelrooms |
| Price | amount | price |
| Star | star | |
| Wifi | | wifi |

```
Select Name,
  AVG(L1.amount, L2.price) as Price,
  SUM(L1.rooms, L2.hotelrooms) as Room
```

...

```
from resort L1 full join hotel L2
  using (name)
```

**Full
Join
Merge**

**Full
Join**

Data Integration and Data Fusion: an example

Global Class $G = \{L1, L2\}$

| G | L1 | L2 |
|----|----|----|
| ID | ID | ID |
| A | A | |
| B | | B |
| C | C | C |

Data Fusion

G as Full Join Merge of L1 and L2

```
SELECT ID,
       L1.A ASA,
       L2.B AS B,
       AVG (L1.C,L2.C) AS C
FROM   L1 FULL JOIN L2
       USING (ID)
```

result

integration

L1

| ID | A | C |
|----|---|---|
| 1 | 3 | 4 |
| 2 | 3 | 1 |
| 3 | | 3 |
| 4 | 8 | 3 |

L2

| ID | B | C |
|----|---|---|
| 1 | 4 | 2 |
| 2 | | |
| 3 | | 3 |
| 5 | | |

| ID | A | B | C=AVG(L1.C,L2.C) |
|----|---|---|------------------|
| 1 | 3 | 4 | 3 |
| 2 | 3 | | 1 |
| 3 | | | 3 |
| 4 | 8 | | 3 |
| 5 | | | |

- **The querying problem:**
How to answer queries expressed on the GS (**global queries**)?
- In a Virtual Data Integration system, data reside at the data sources then the query processing is based on **Query rewriting** :
to rewrite a global query as an equivalent set of queries expressed on the local schemata data sources (**local queries**).
- **GAV** approach: query rewriting is performed by **unfolding**, i.e. by expanding a global query on C according to the **view** associated to C
 - When the view is defined with an *outer-join merge* operator, the query rewriting performs the fusion (object identification and conflict resolution) of the local answers into the global answer.
- **Query Manager**
 - Distributed Query Processing
 - Query Optimization

Query unfolding: Predicate push down

global constraint on *one-to-one attributes* can be push down on local queries

Global Query

Query 1:
 Select Name, Room
 from Hotel
 where Price = 100 and Stars > 3

| Hotel | resort | hotel |
|-------|--------|------------|
| Name | name | name |
| Room | rooms | hotelrooms |
| Price | amount | price |
| Star | stars | - |
| Wifi | - | wifi |

↓ Local queries

Q1 to local source "resort".

Q1_resort:
 Select name, amount, rooms
 from resort
 where stars > 3

Q1 to local source "hotel".

Q1_hotel:
 Select name, price,
 hotelrooms
 from hotel

Query unfolding: Full Outer Join simplification

The local answers (Q_{Li}) are fused into the global answer on the basis of the Full Outer Join-merge operation:

Q_{L1} full join Q_{L2} on $JC(L1, L2)$

✓ **Full Join simplification:**

- (1) $FOJ = Q_{L1}$ left join Q_{L2} on $JC(L1, L2)$**
if there exists predicate pushed down only on L1
- (2) $FOJ = Q_{L1}$ inner join Q_{L2} on $JC(L1, L2)$**
if there exists a predicate pushed down only on L1 and
a predicate pushed down only on L2.

For Query Q1:

$FOJ_{Q1} =$ **select ***
from Q1_resort **left join** Q1_hotel
on Q1_hotel.name = Q1_resort.name

Query unfolding: Resolution Function

RES_FOJ: Application of the resolution functions to FOJ

```
RES_FOJ_Q1 = select
              COALESCE(Q1_hotel.name, Q1_resort.name) AS Name,
              SUM(Q1_hotel.rooms, Q1_resort.hotelrooms) AS Room,
              AVG(Q1_hotel.amount, Q1_resort.price) AS Price
from FOJ_Q1
```

Query Result: Application of the residual conditions to RES_FOJ

```
Query Result = select
                Name,
                Room
from RES_FOJ_Q1
where Price = 100
```