DB Group @ unimo

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

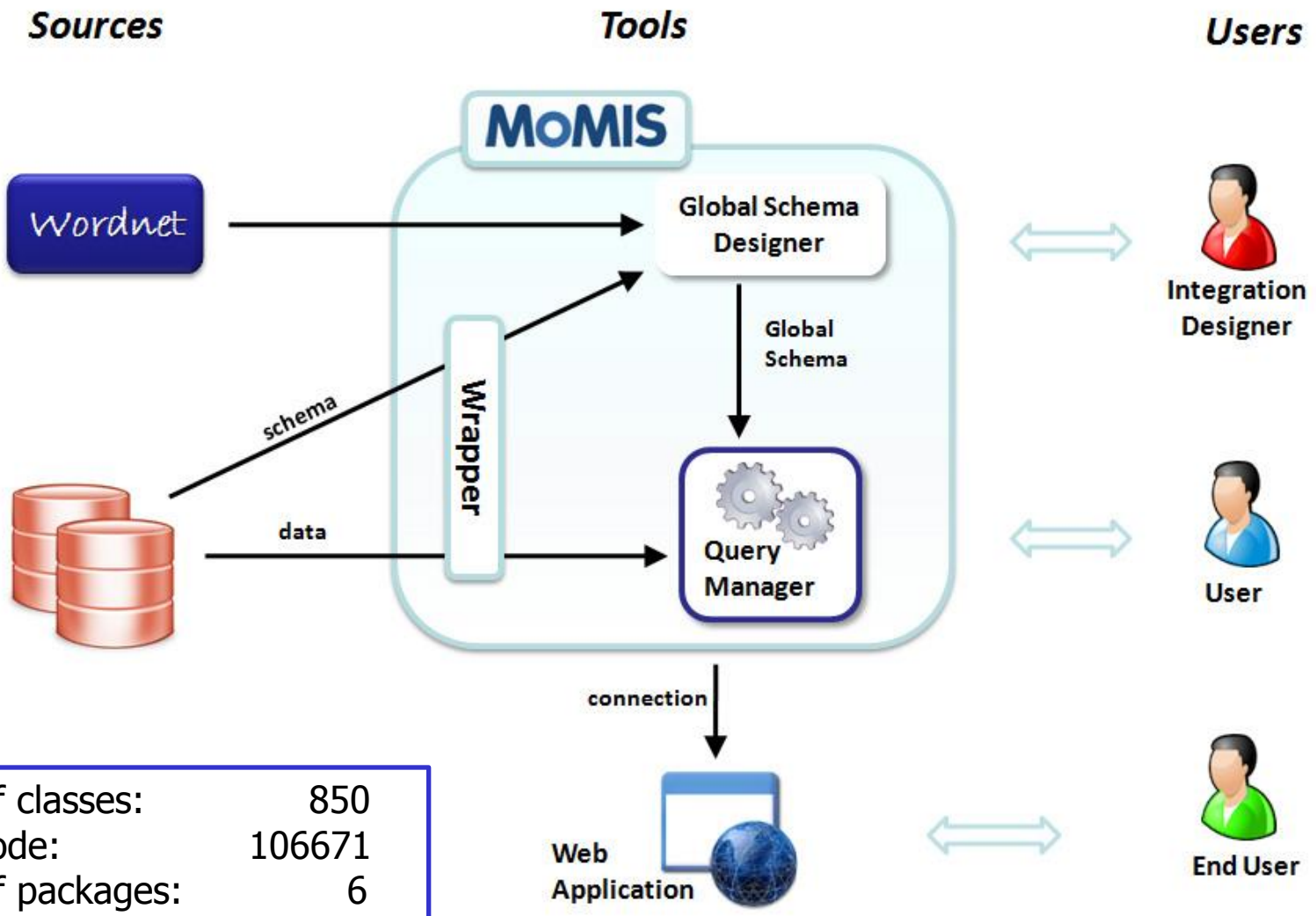
www.datariver.it

Spin-off presso Università di Modena e Reggio Emilia

DataRiver

open source data integration

Software architecture



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

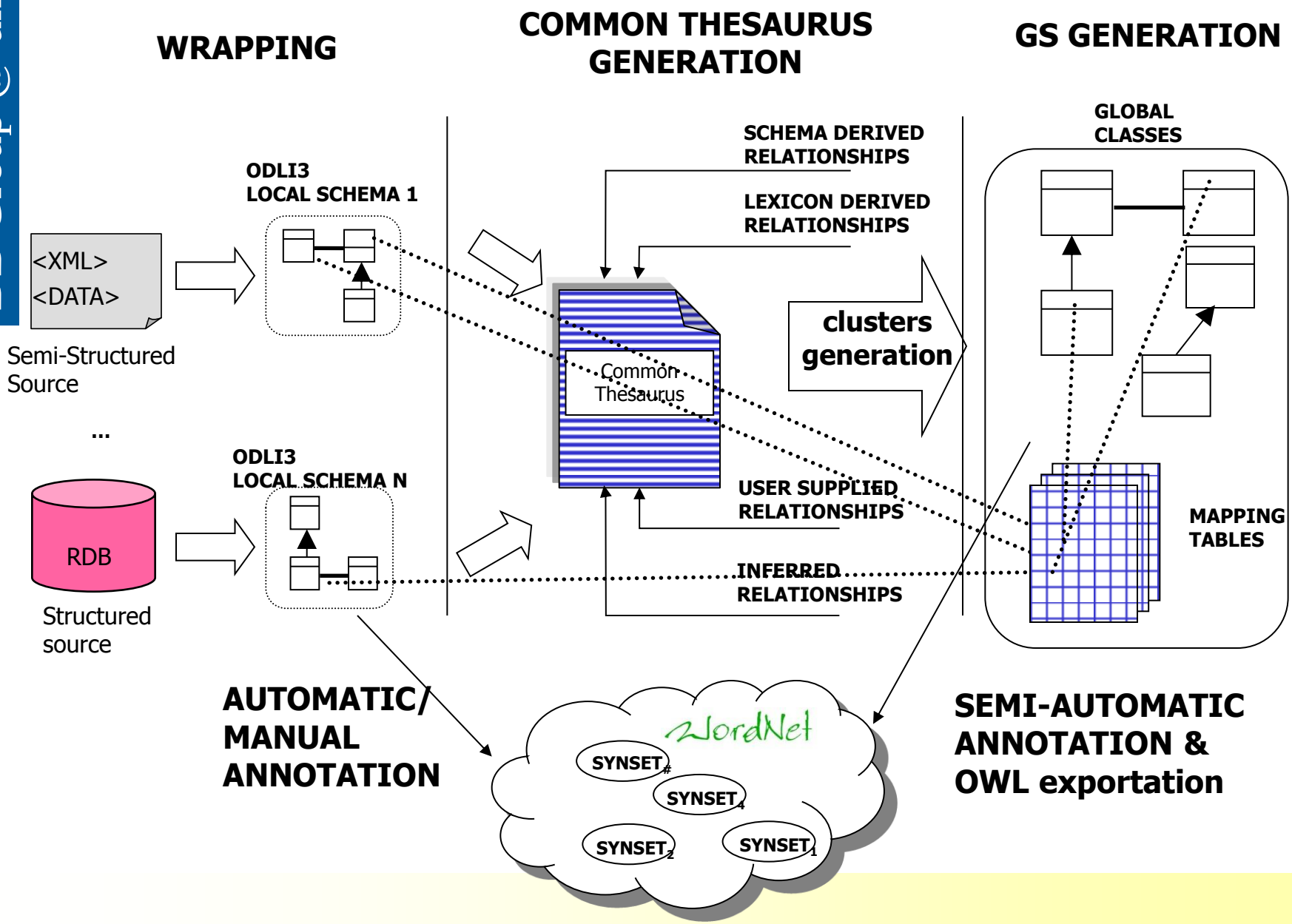
1. Global Schema and Mapping generation

→ 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



Practical Session of Today

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

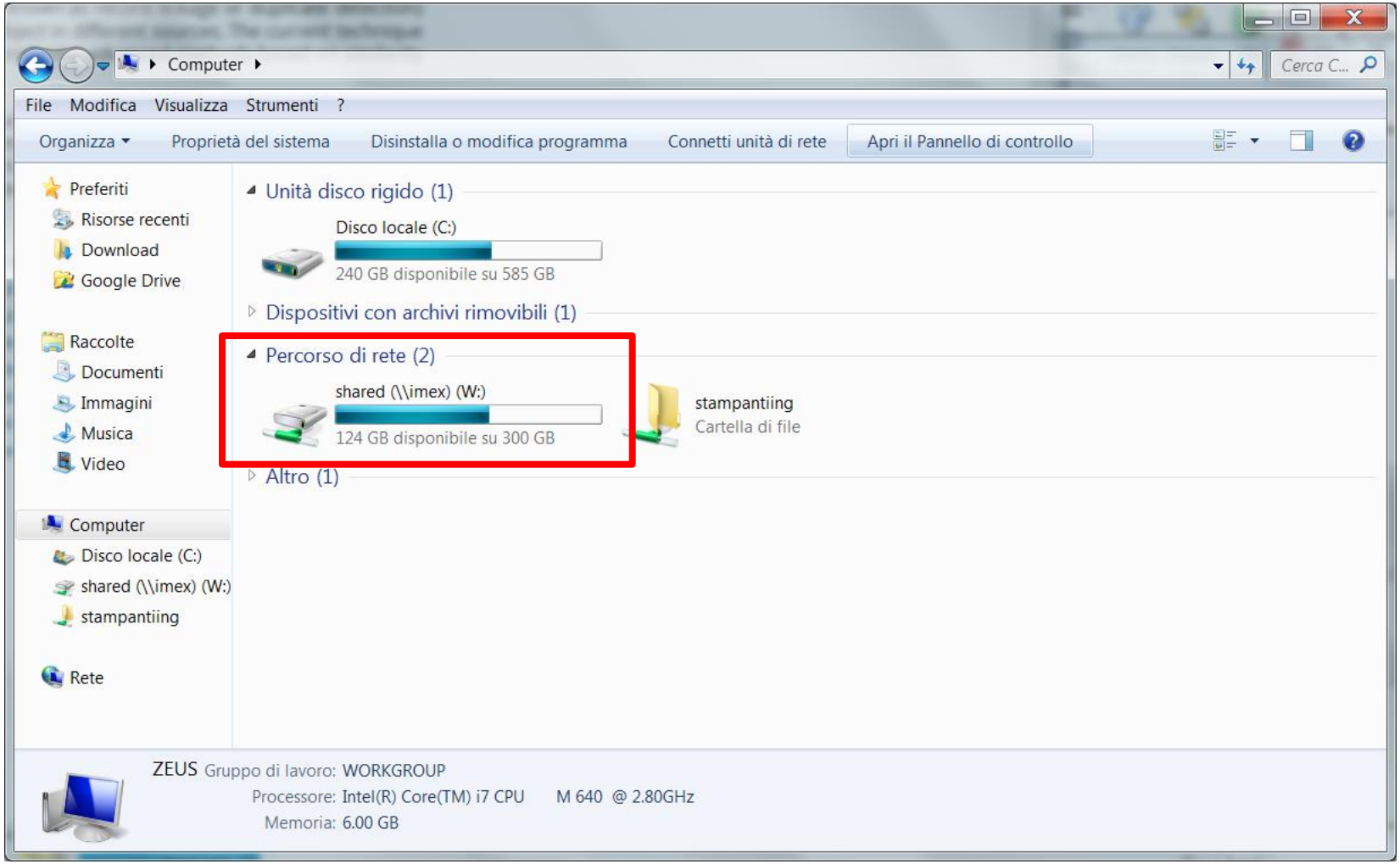
SET UP THE ENVIRONMENT

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

7

Connect to the IMEX server

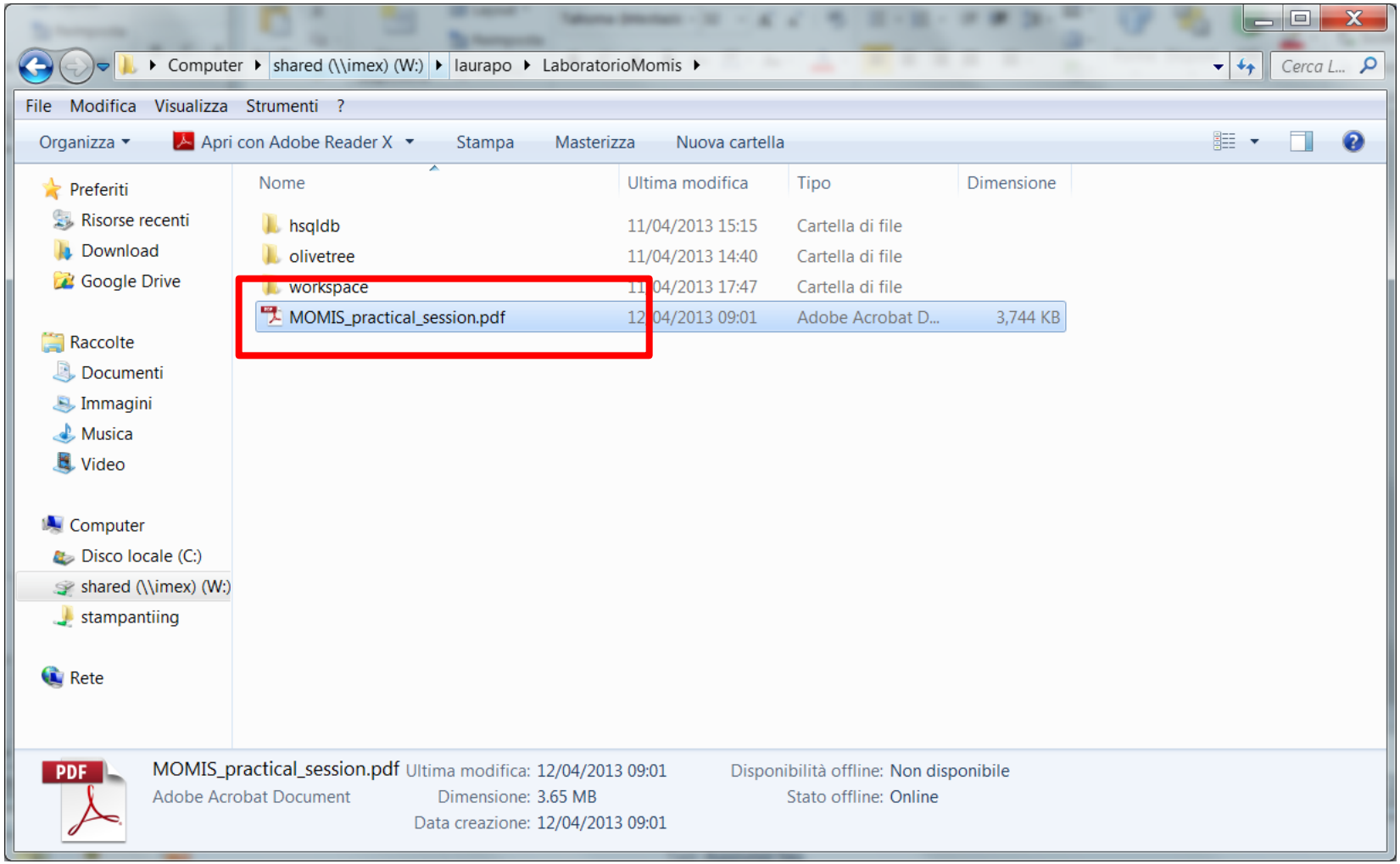


Download this presentation

- Open the folder

  W:\laurapo\LaboratorioMomis\
- Download the presentation

  MOMIS_practical_session.pdf


Computer
shared (\\imex) (W:)
laurapo
LaboratorioMomis
Cerca L...
File
Modifica
Visualizza
Strumenti
?
Organizza
Apri con Adobe Reader X
Stampa
Masterizza
Nuova cartella
Preferiti
Risorse recenti
Download
Google Drive
Raccolte
Documenti
Immagini
Musica
Video
Computer
Disco locale (C:)
shared (\\imex) (W:)
stampantiing
Rete
Nome
Ultima modifica
Tipo
Dimensione
hsqldb
11/04/2013 15:15
Cartella di file
olivetree
11/04/2013 14:40
Cartella di file
workspace
Cartella di file
MOMIS_practical_session.pdf
Adobe Acrobat D...
3,744 KB
PDF
MOMIS_practical_session.pdf
Adobe Acrobat Document
Ultima modifica: 12/04/2013 09:01
Dimensione: 3.65 MB
Data creazione: 12/04/2013 09:01
Disponibilità offline: Non disponibile
Stato offline: Online

Connection to IMEX

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

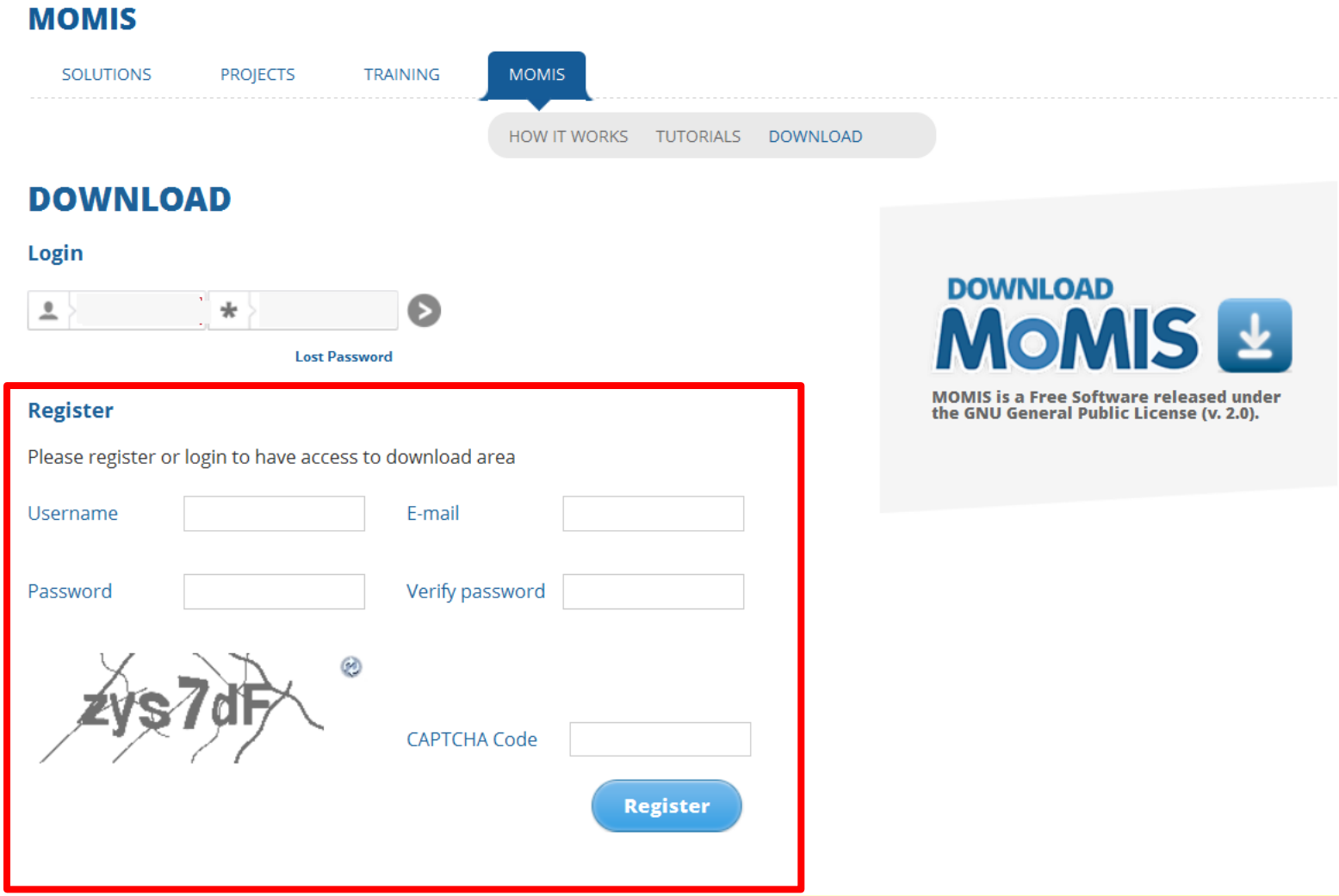
- `net use W: \\imex\shared /user:imex\reader primo99`

11

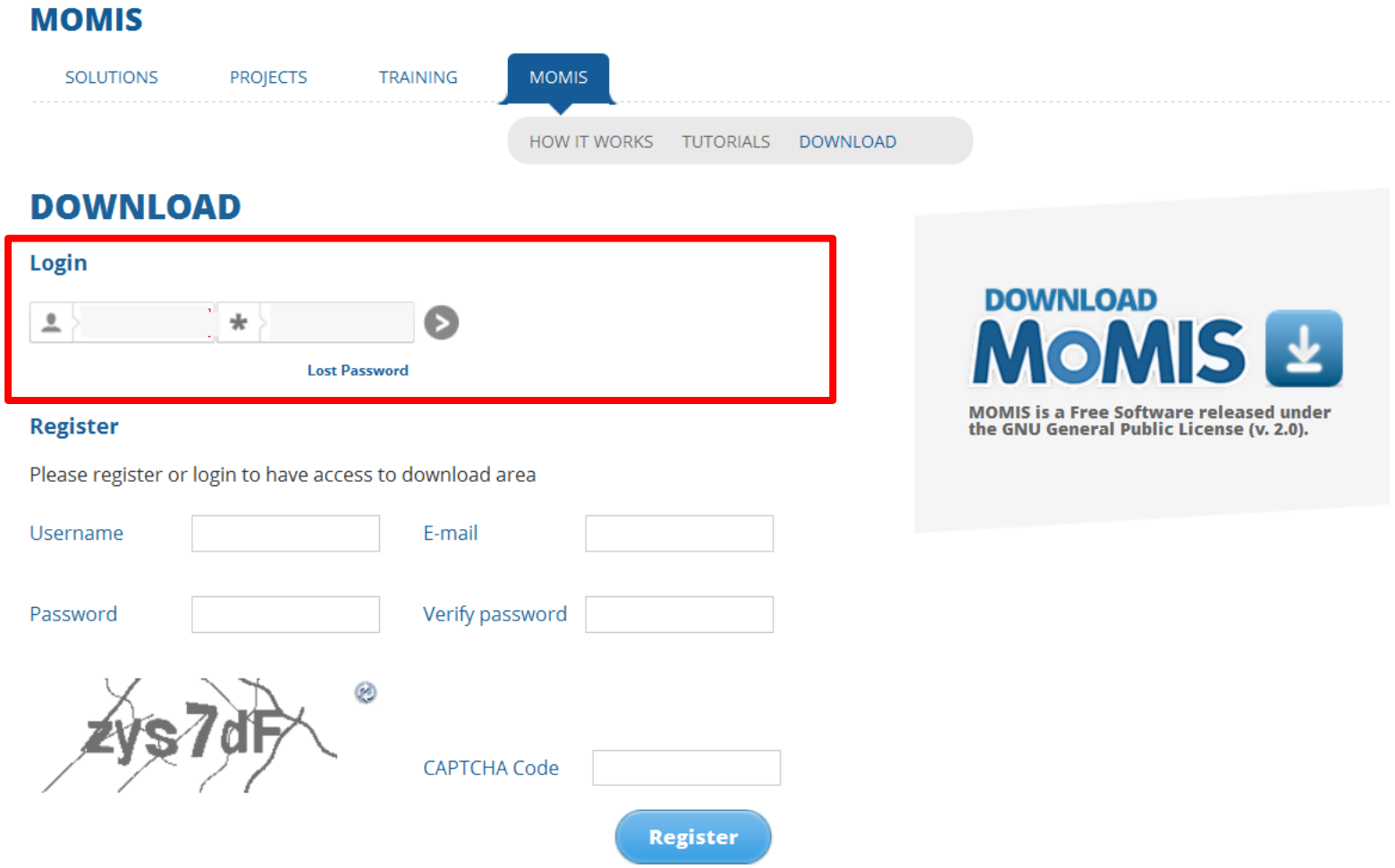
Connect to www.datariver.it

Go to the download page

Register


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
zys7dF
CAPTCHA Code
Register
DOWNLOAD
MOMIS
MOMIS is a Free Software released under the GNU General Public License (v. 2.0).

Login


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
zys7dF
CAPTCHA Code
Register
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:

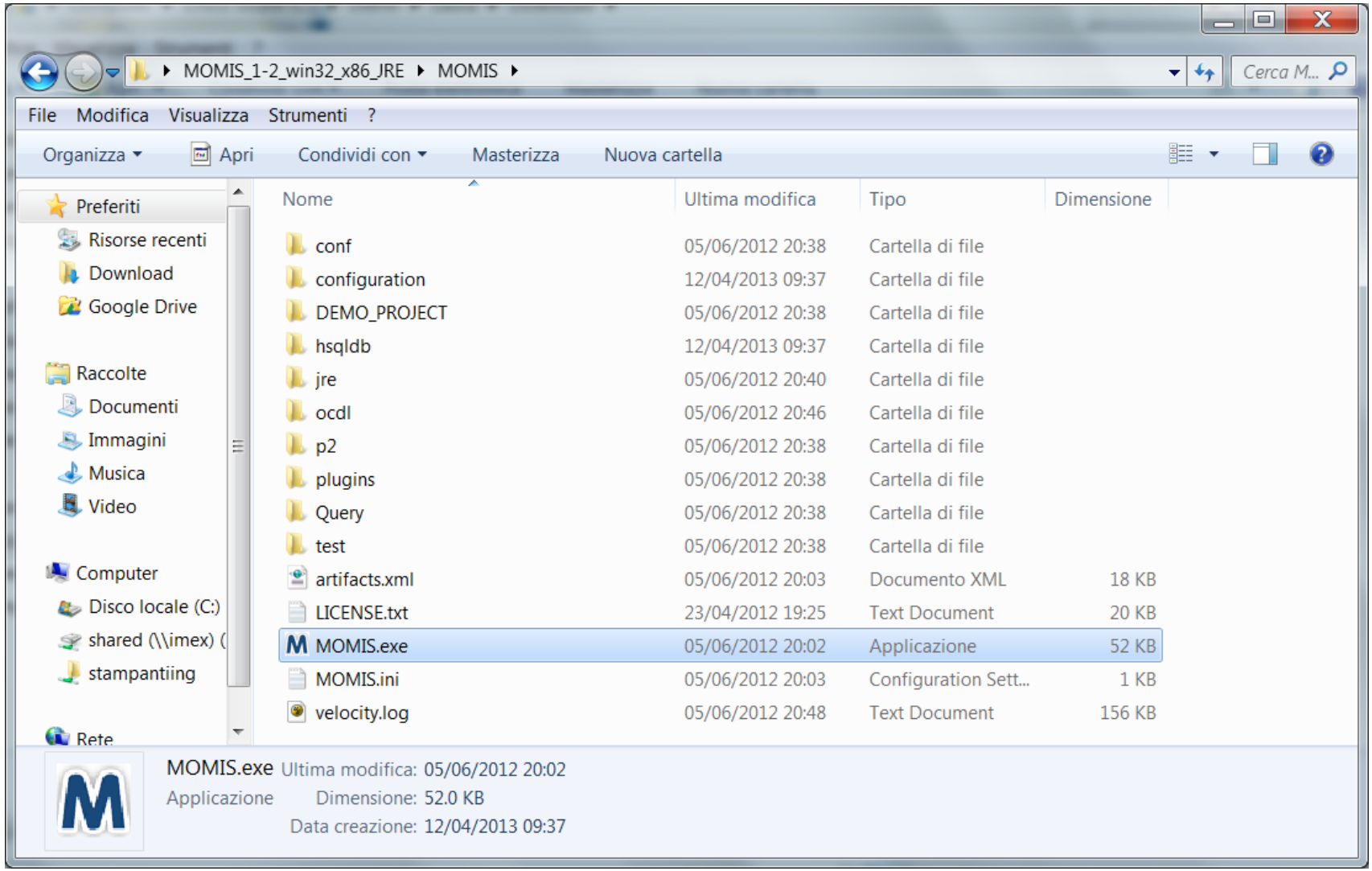
MOMIS requires the Java Runtime Environment (JRE) 1.6 or above. If you already have a JRE (or JDK) installed on your computer, you can download the MOMIS version without the JRE included. If you don't have a JRE installed on your computer, download the MOMIS version with the JRE included.

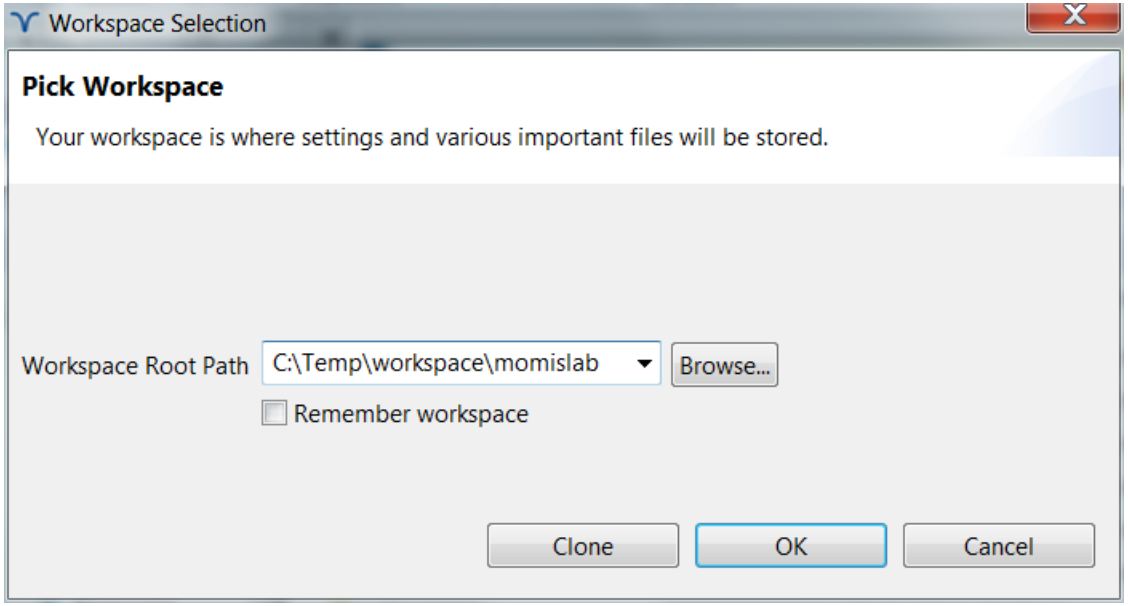
- ✓ MOMIS
  - ✓ momis source code 1.2 Version:1.2.0
  - ✓ momis_1-2_macosx_cocoa_x86_64 Version:1.2.0
  - ✓ momis_1-2_macosx_cocoa_x86 Version:1.2.0
  - ✓ momis_1-2_win32_x86_64 Version:1.2.0
  - ✓ momis_1-2_win32_x86_jre Version:1.2.0
  - ✓ momis_1-2_win32_x86 Version:1.2.0
  - ✓ momis_1-2_linux_gtk_x86_64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

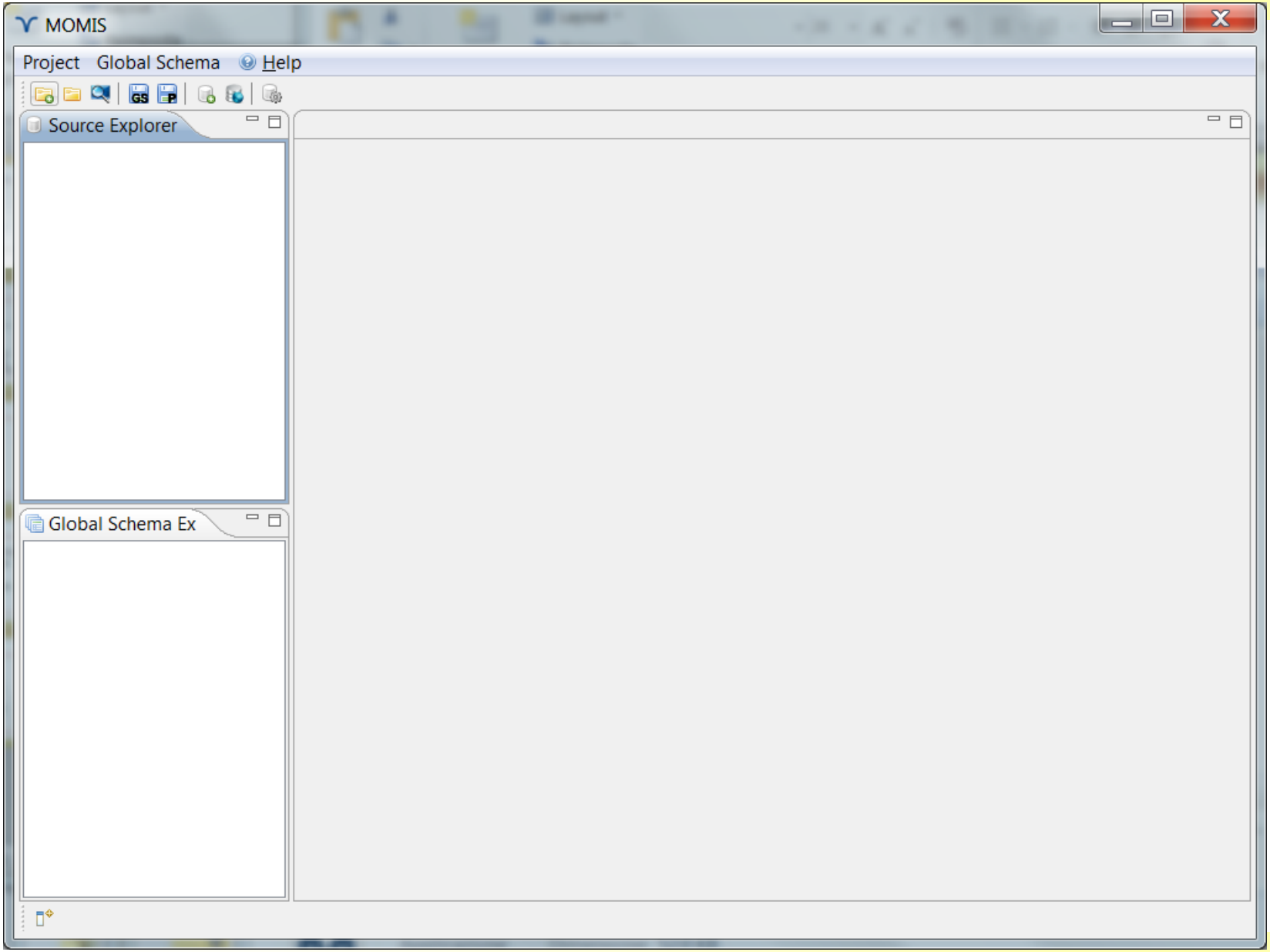
First run of MOMIS

- Extract the MOMIS archive on the desktop
- Execute the file MOMIS.exe that is located in

  `MOMIS_1-2_win32_x86_JRE\MOMIS\`
- Define the workspace
  `C:\Temp\workspace\momislab`


MOMIS_1-2_win32_x86_JRE ▸ MOMIS ▸
Cerca M...
File Modifica Visualizza Strumenti ?
Organizza Apri Condividi con Masterizza Nuova cartella
Preferiti
Risorse recenti
Download
Google Drive
Raccolte
Documenti
Immagini
Musica
Video
Computer
Disco locale (C:)
shared (\\imex) (
stampantiing
Rete
Nome
Ultima modifica
Tipo
Dimensione
conf 05/06/2012 20:38 Cartella di file
configuration 12/04/2013 09:37 Cartella di file
DEMO_PROJECT 05/06/2012 20:38 Cartella di file
hsqldb 12/04/2013 09:37 Cartella di file
jre 05/06/2012 20:40 Cartella di file
ocdl 05/06/2012 20:46 Cartella di file
p2 05/06/2012 20:38 Cartella di file
plugins 05/06/2012 20:38 Cartella di file
Query 05/06/2012 20:38 Cartella di file
test 05/06/2012 20:38 Cartella di file
artifacts.xml 05/06/2012 20:03 Documento XML 18 KB
LICENSE.txt 23/04/2012 19:25 Text Document 20 KB
MOMIS.exe 05/06/2012 20:02 Applicazione 52 KB
MOMIS.ini 05/06/2012 20:03 Configuration Sett... 1 KB
velocity.log 05/06/2012 20:48 Text Document 156 KB
MOMIS.exe Ultima modifica: 05/06/2012 20:02
Applicazione Dimensione: 52.0 KB
Data creazione: 12/04/2013 09:37


Workspace Selection
Pick Workspace
Your workspace is where settings and various important files will be stored.
Workspace Root Path
C:\Temp\workspace\momislab
Browse...
Remember workspace
Clone
OK
Cancel


MOMIS
Project
Global Schema
Help
Source Explorer
Global Schema Ex

TEST 1

Test 1

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

22

Download the workspace

- 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`

23

Download the sources

- Access the following folder on IMEX:

  `W:\laurapo\LaboratorioMomis`

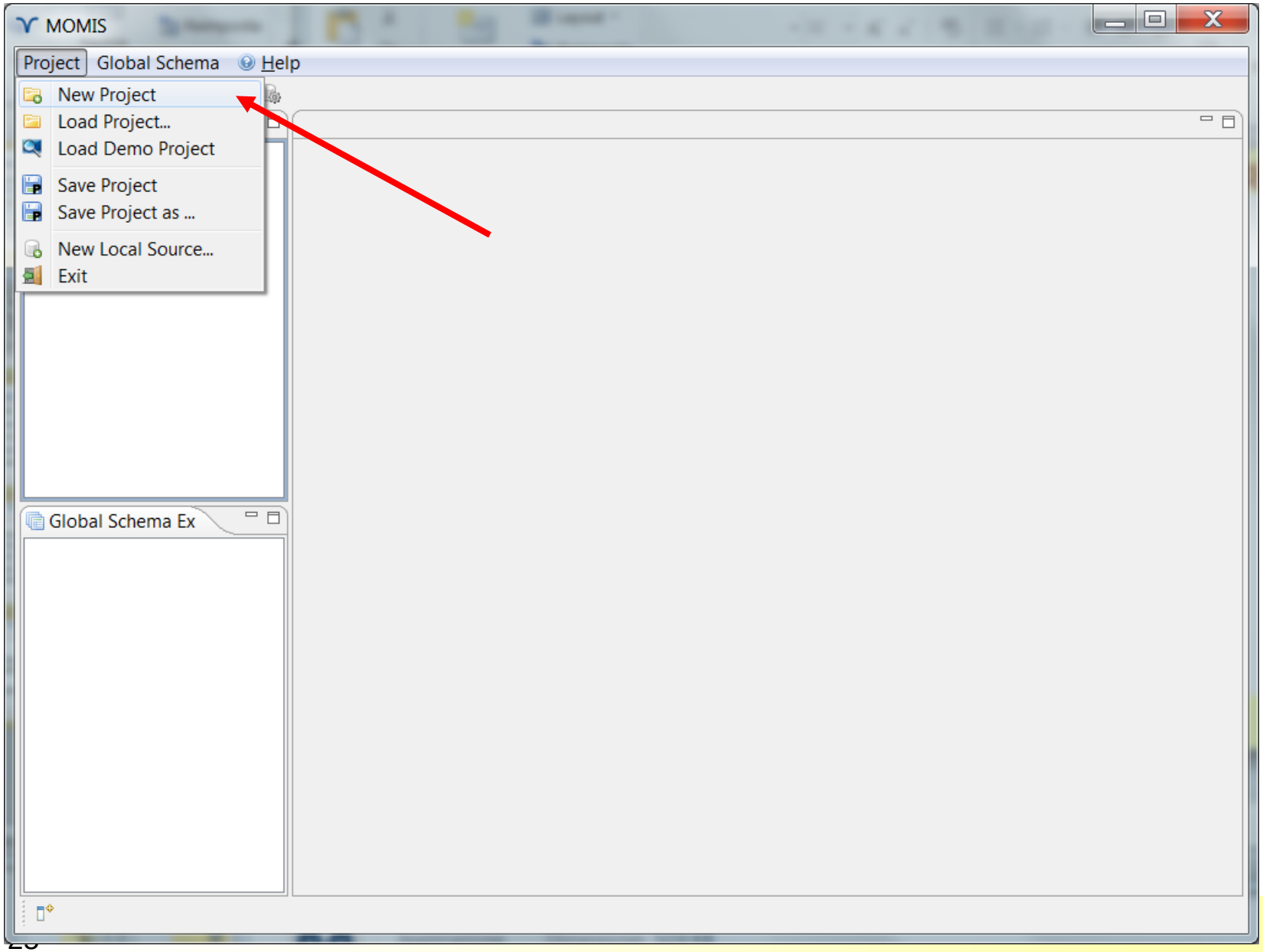
- Copy the folder `olivetree` in your local folder

`MOMIS_1-2_win32_x86_JRE\MOMIS\test\sourcesDb\`

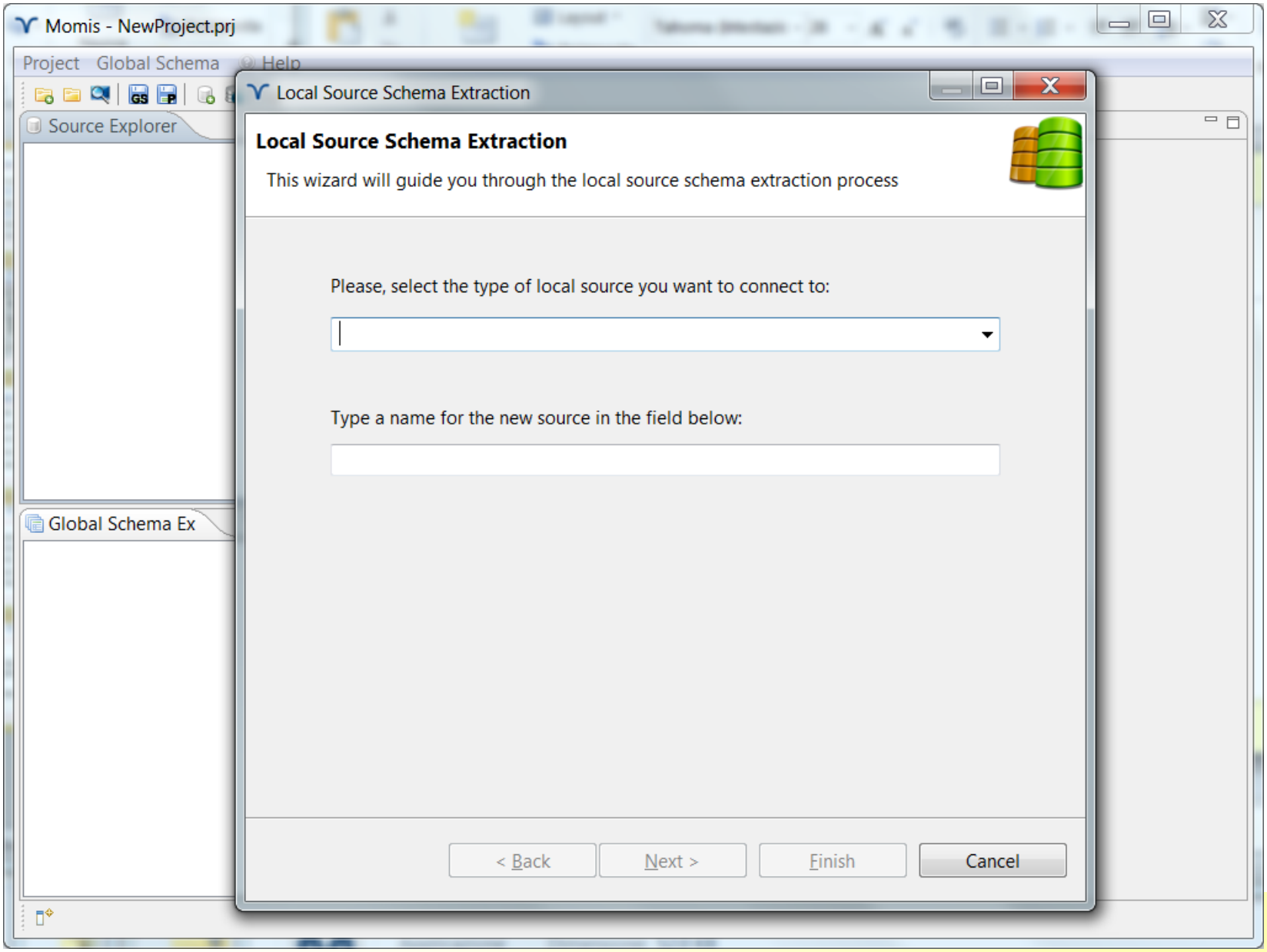
- Run MOMIS

24

Create a new project



Add new sources

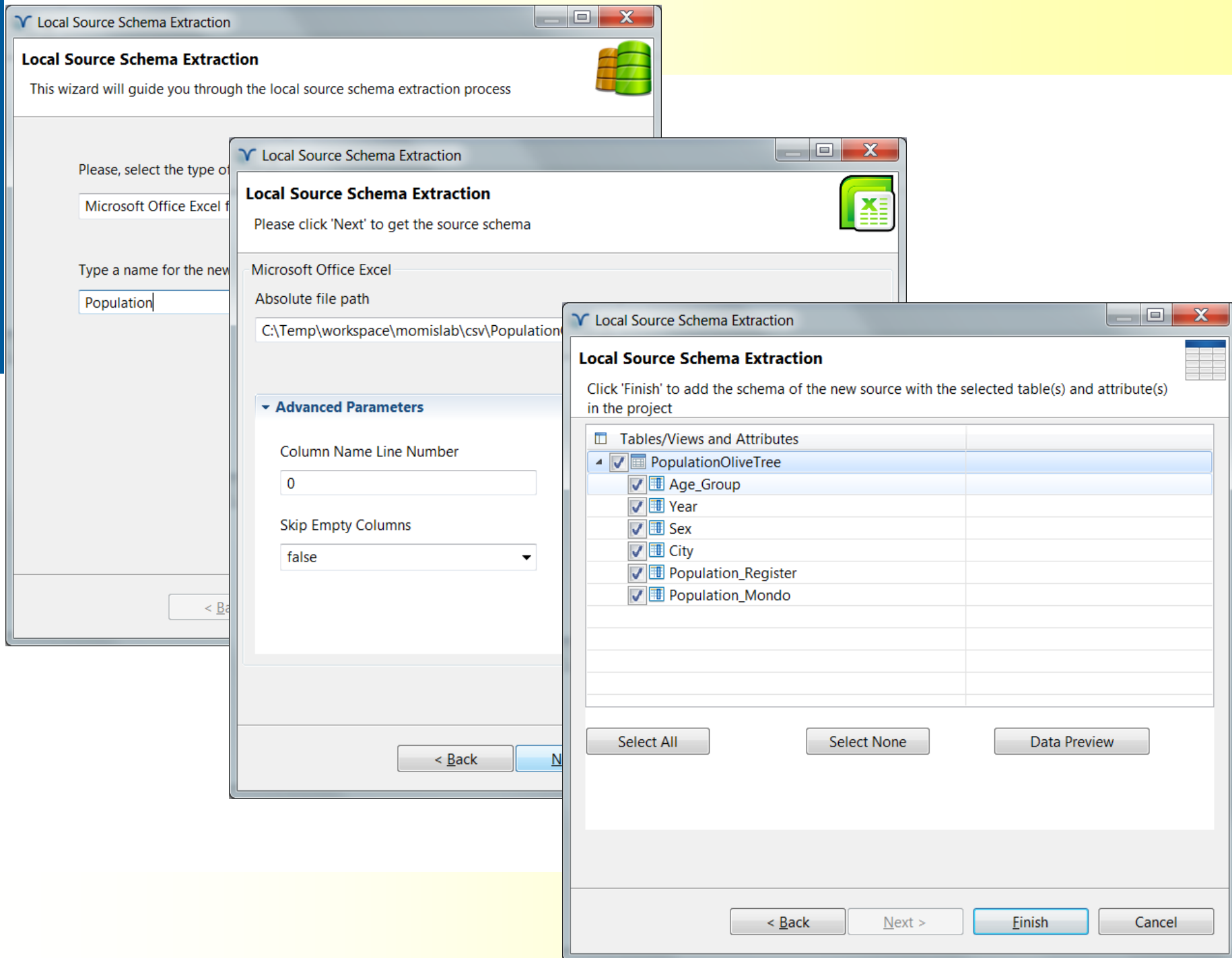


Set up the connection to an Excel file

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

27

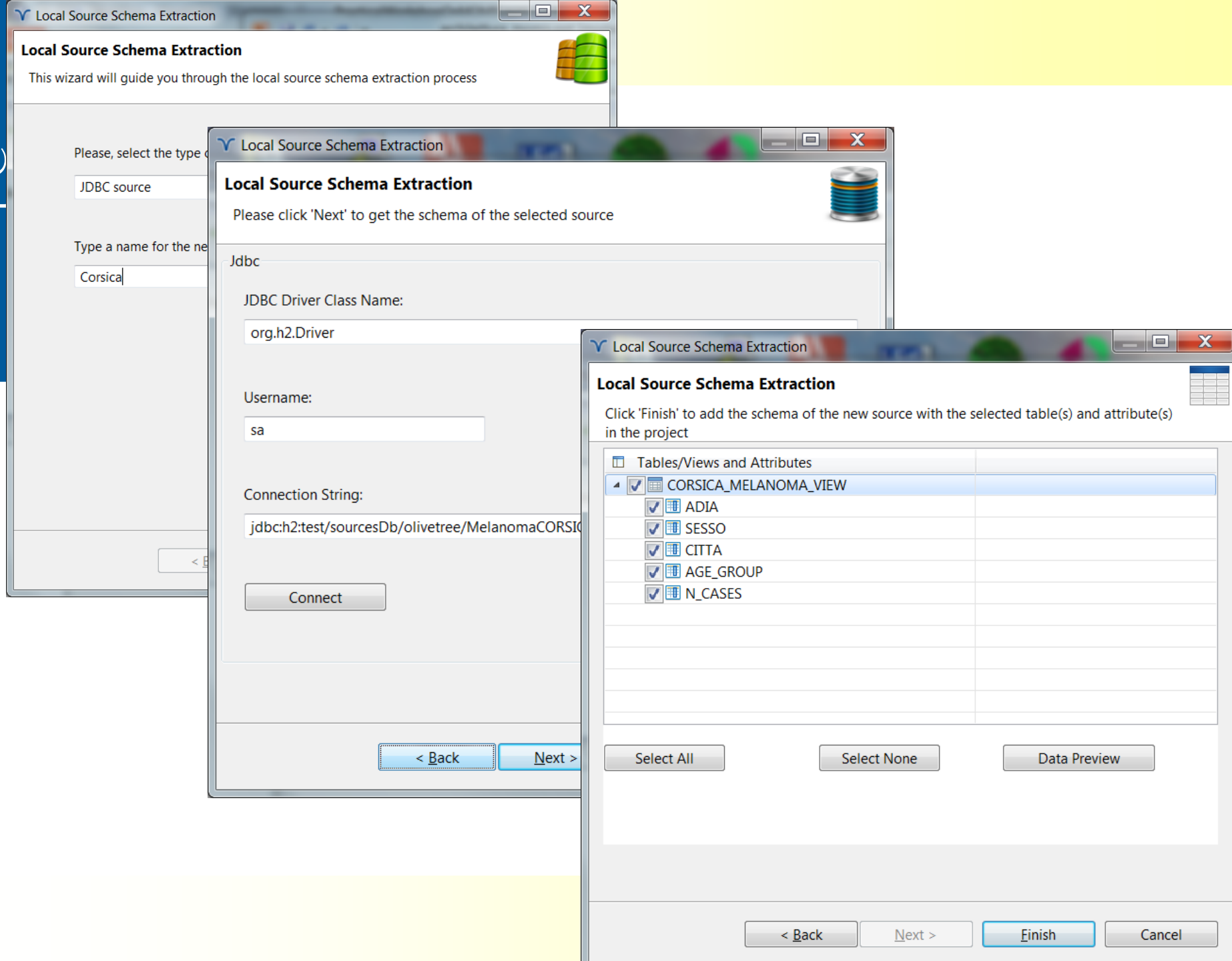
DB Group @ unimo



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/MelanomaCORSICA/db
  - **Upload Driver:** `C:\Temp\workspace\momislab\lib\h2-1.2.125.jar`
- Select connect
- Select the table `CorsicaMelanomaView`

29

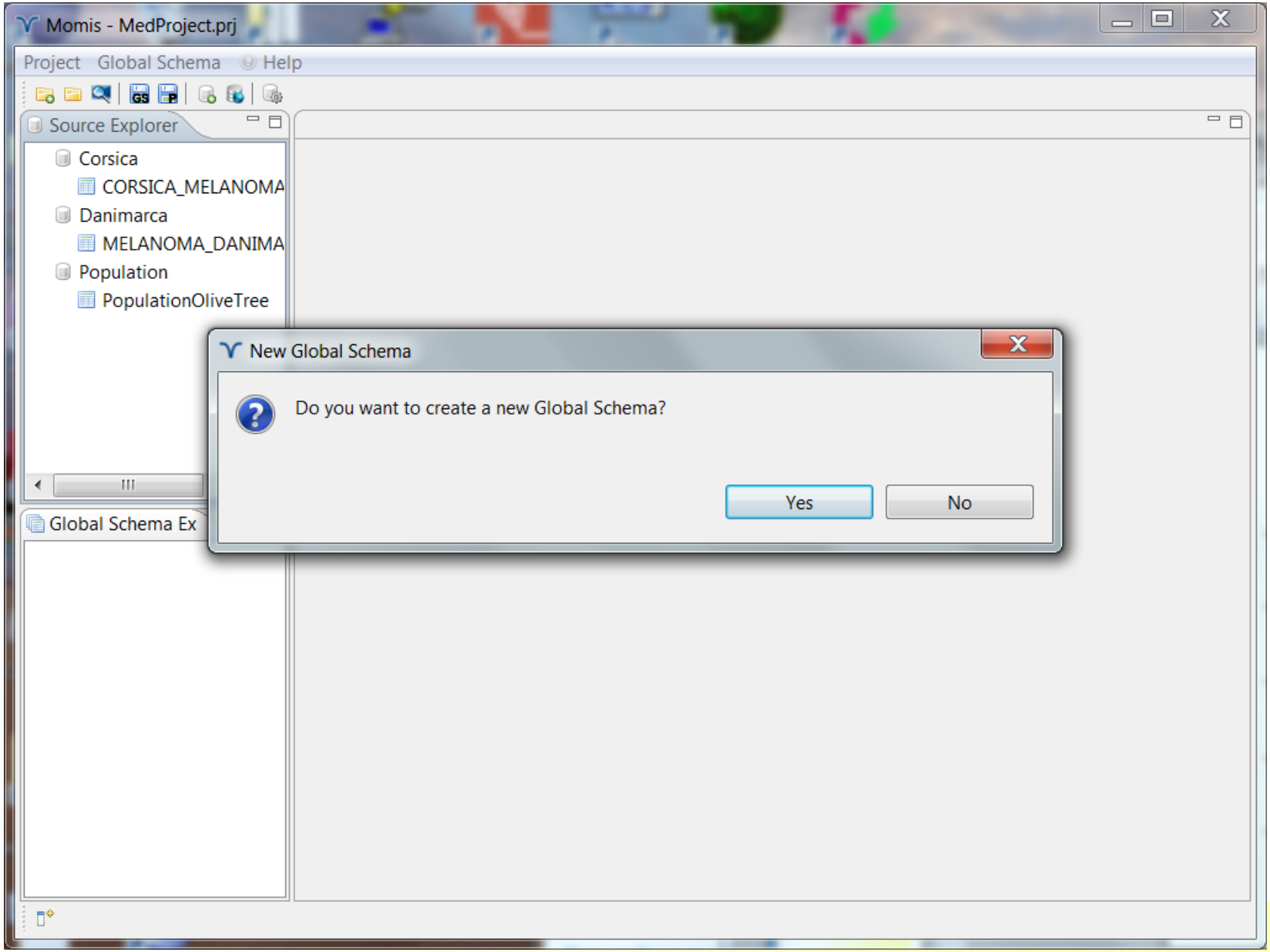

Local Source Schema Extraction
Local Source Schema Extraction
This wizard will guide you through the local source schema extraction process
Please, select the type
JDBC source
Type a name for the ne
Corsica
Local Source Schema Extraction
Local Source Schema Extraction
Please click 'Next' to get the schema 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
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
CORSICA_MELANOMA_VIEW
ADIA
SESSO
CITTA
AGE_GROUP
N_CASES
Select All
Select None
Data Preview
< Back
Next >
Finish
Cancel

Set up the connection to the second H2 database

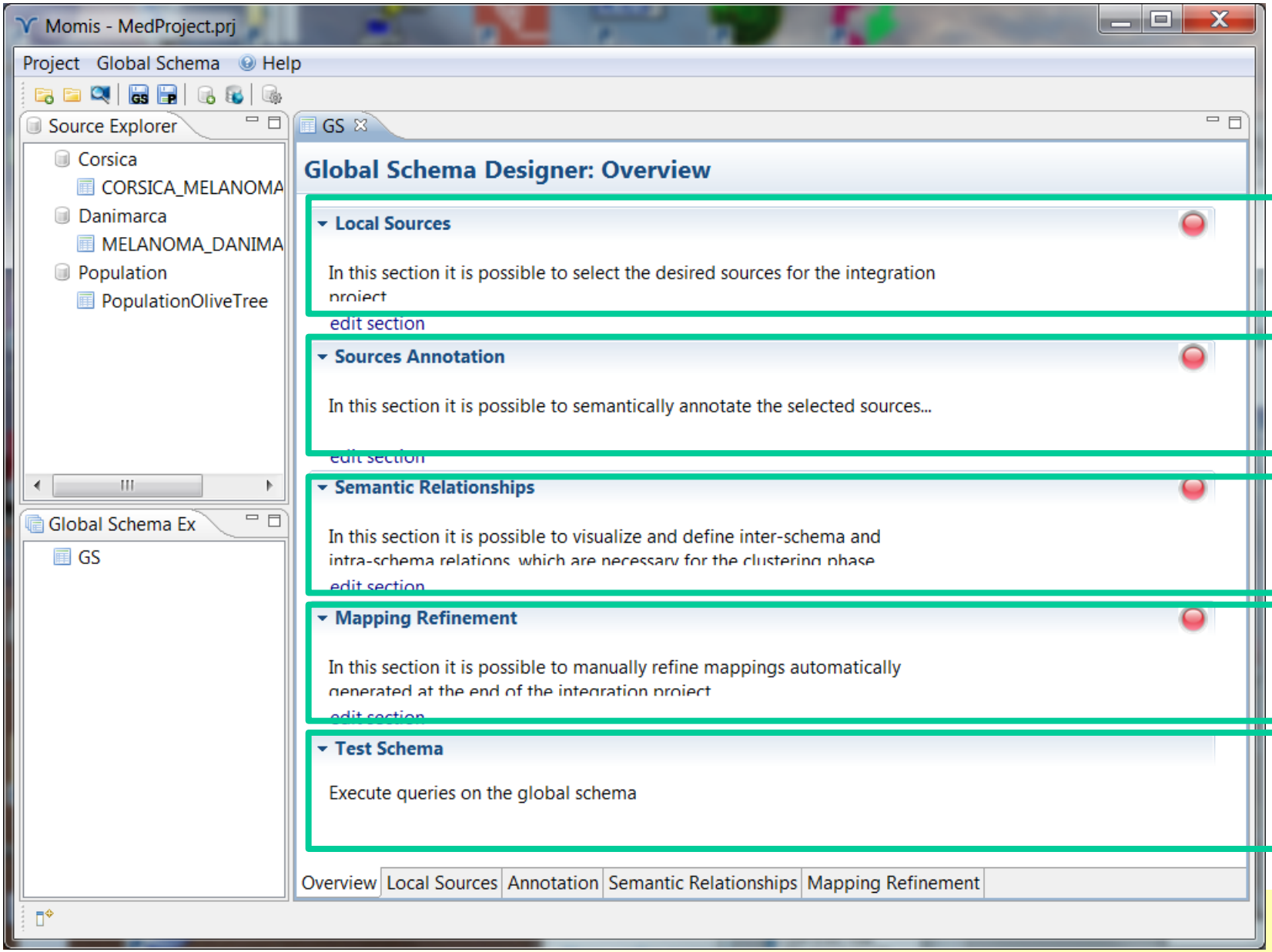
- Set up the parameters
  - **type**: Jdbc Source
  - **name**: Danimarca
  - **JDBC Driver class**: org.h2.Driver
  - **user**: sa
  - **Conn String**: jdbc:h2:test/sourcesDb/olivetree/MelanomaDANIMARCA/db
- Select connect
- Select the table `MelanomaDanimarcaView`

31

Select to create a new GS



Start the integration process following the steps



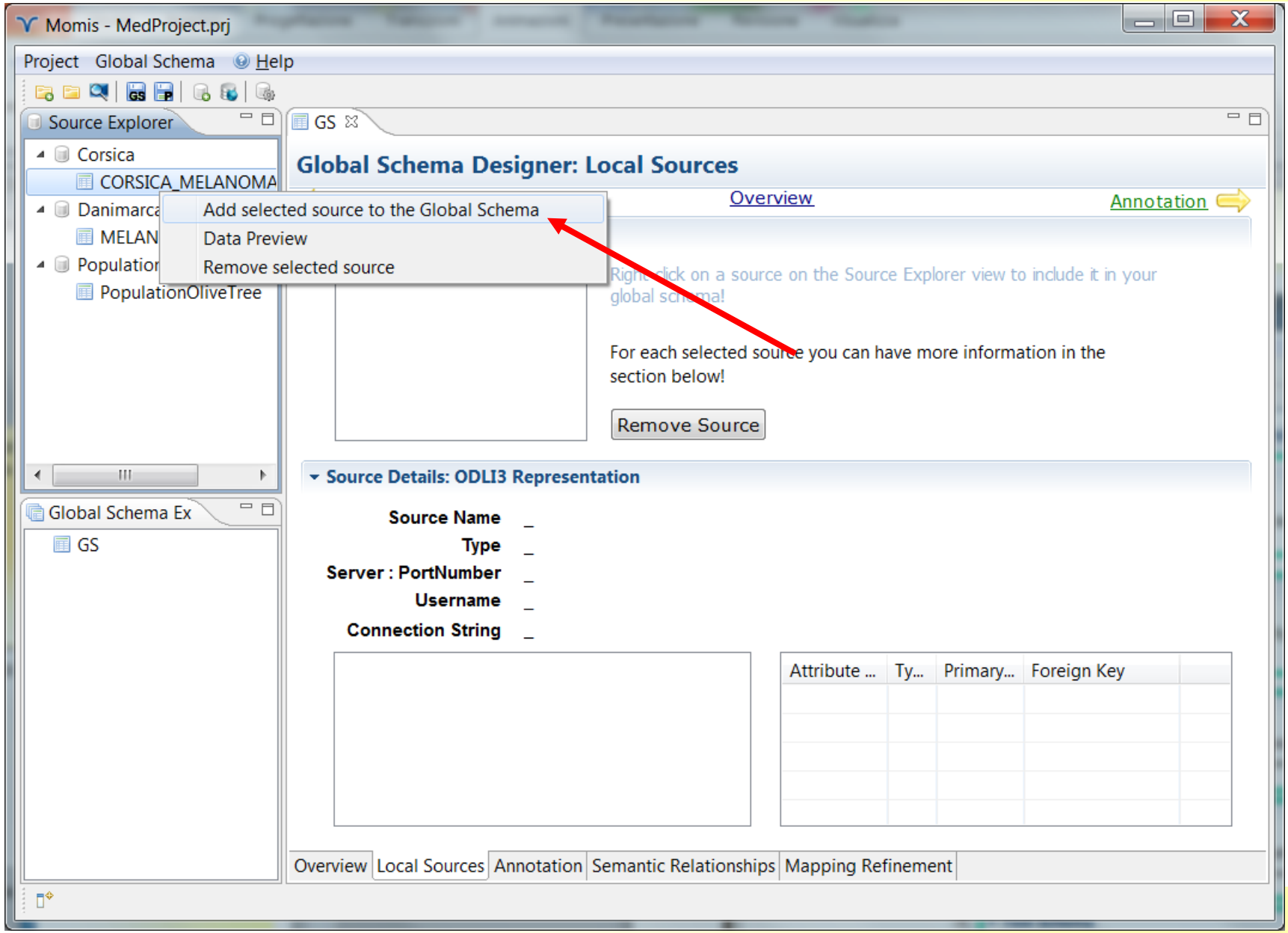
USE OF MOMIS – Test 1

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

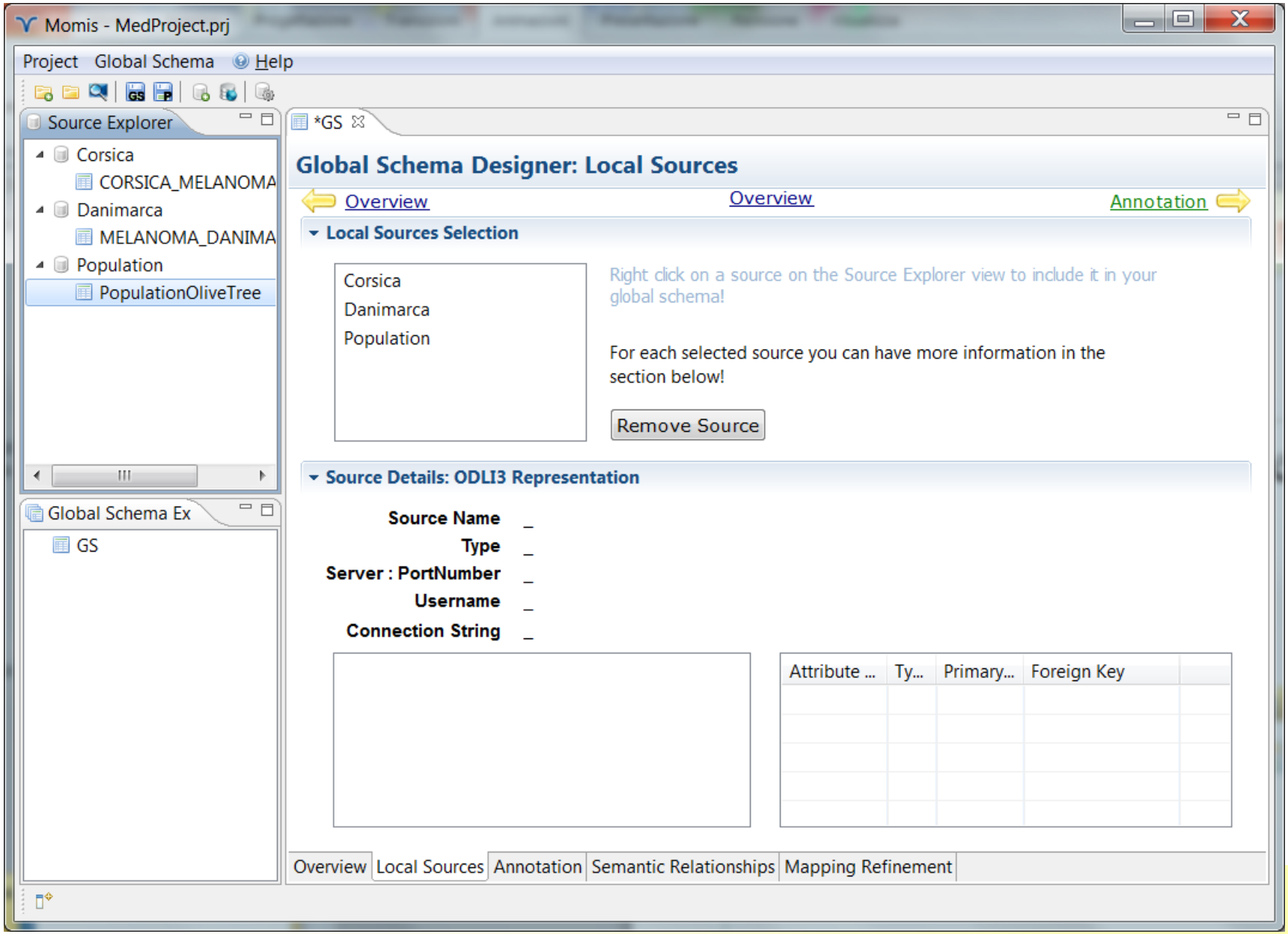
- Local source selection:
- Annotation:
  - Simple
  - Adding new terms and new meanings
- Relationship discovery:
  - Automatic computation
  - Adding new relationships
  - Compute inferred relationships
- Cluster generation and Mapping refinement
- Query the final GS

34

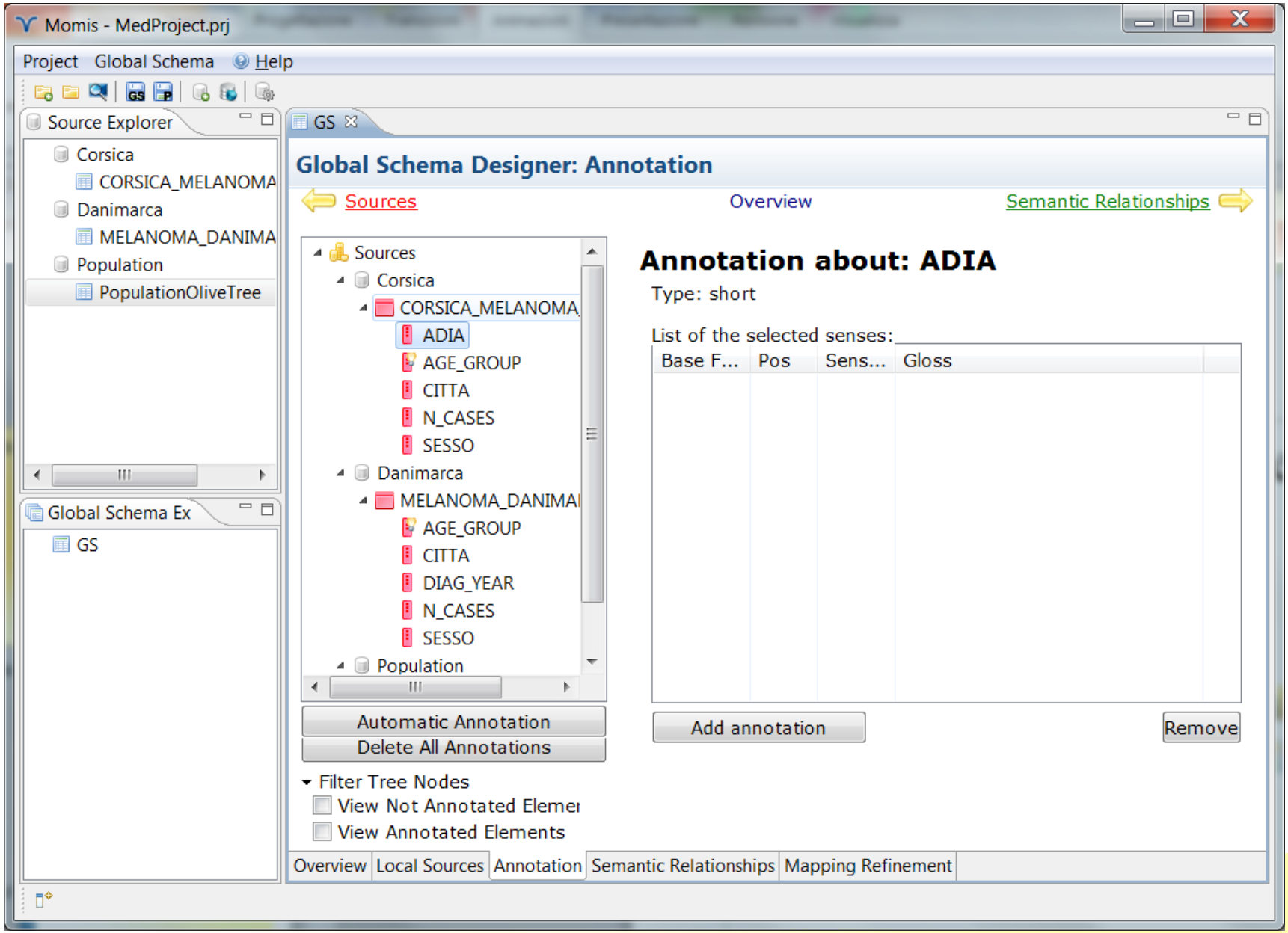
Step1 - Add all the 3 sources to the GS



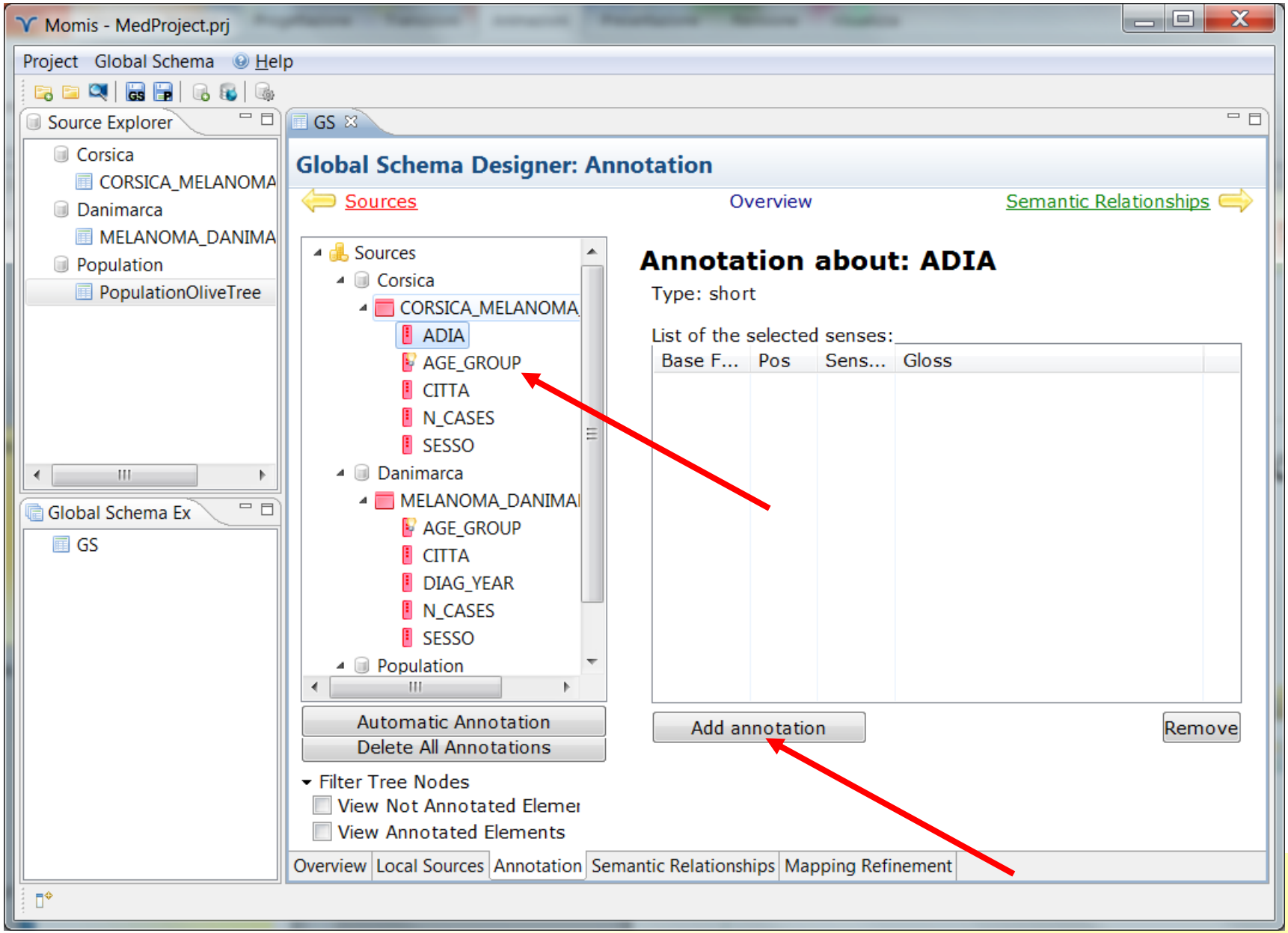
Proceed with the second step: Annotation



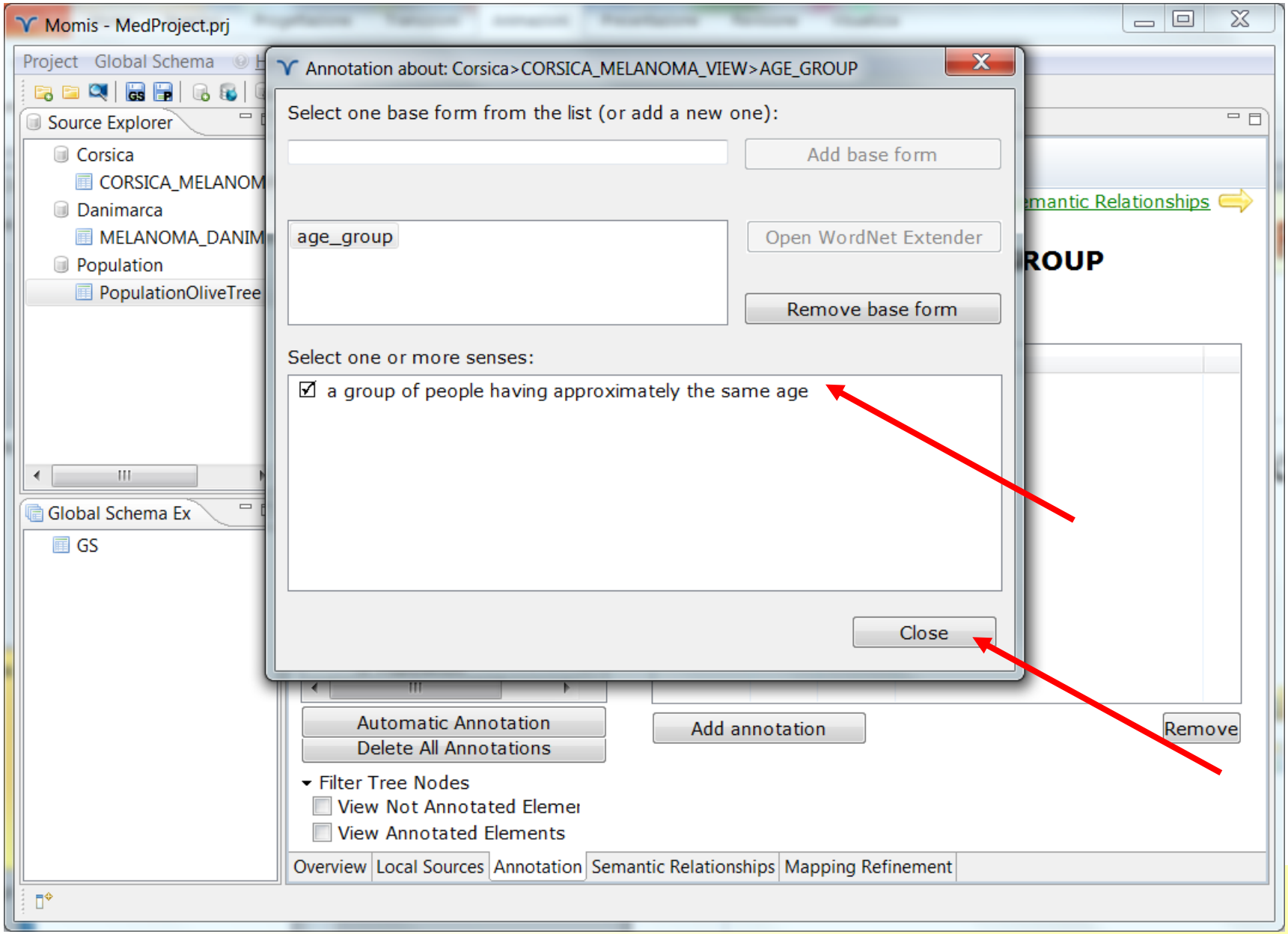
Step 2 - Annotation



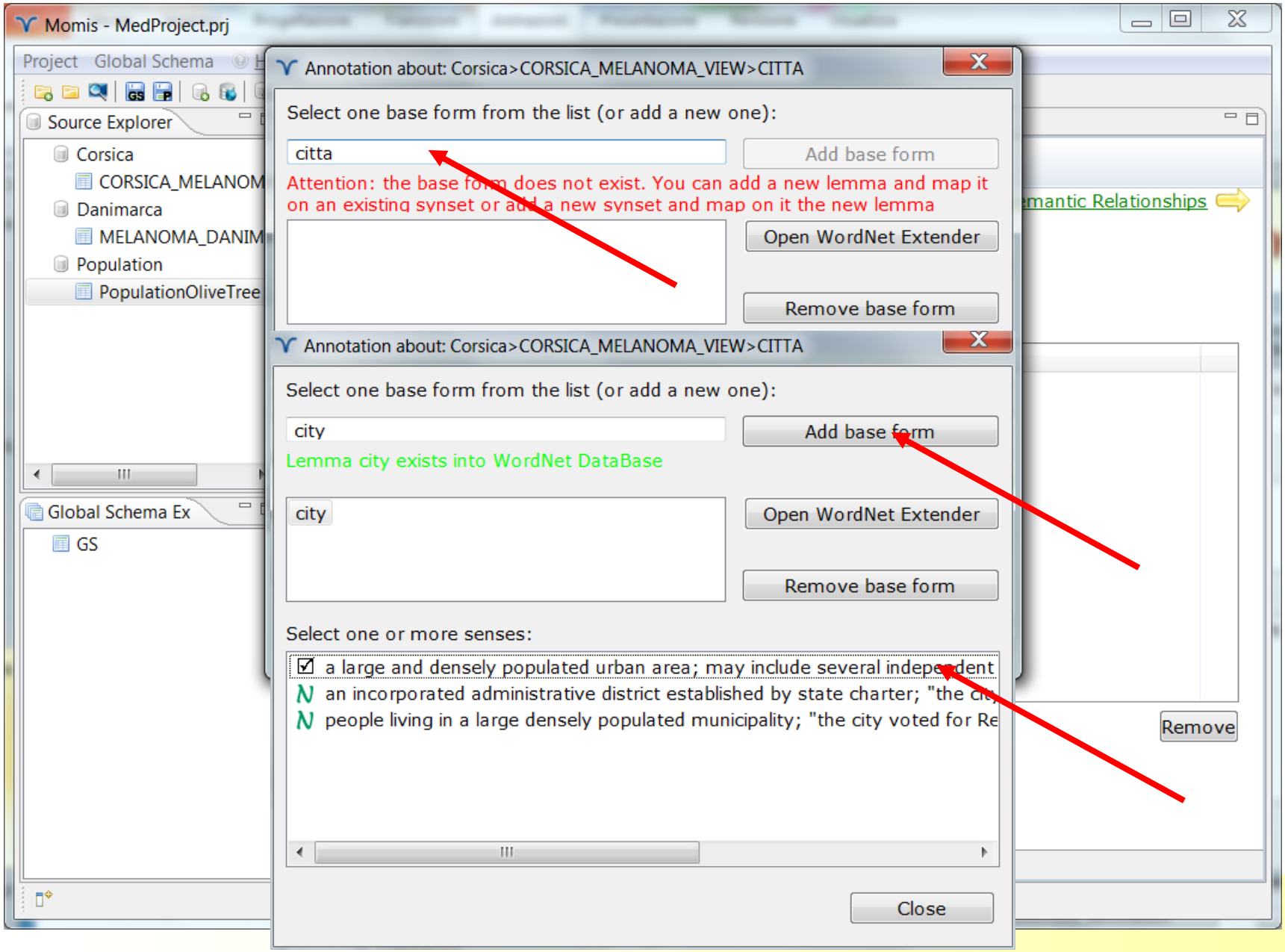
Select annotations for existing lemma



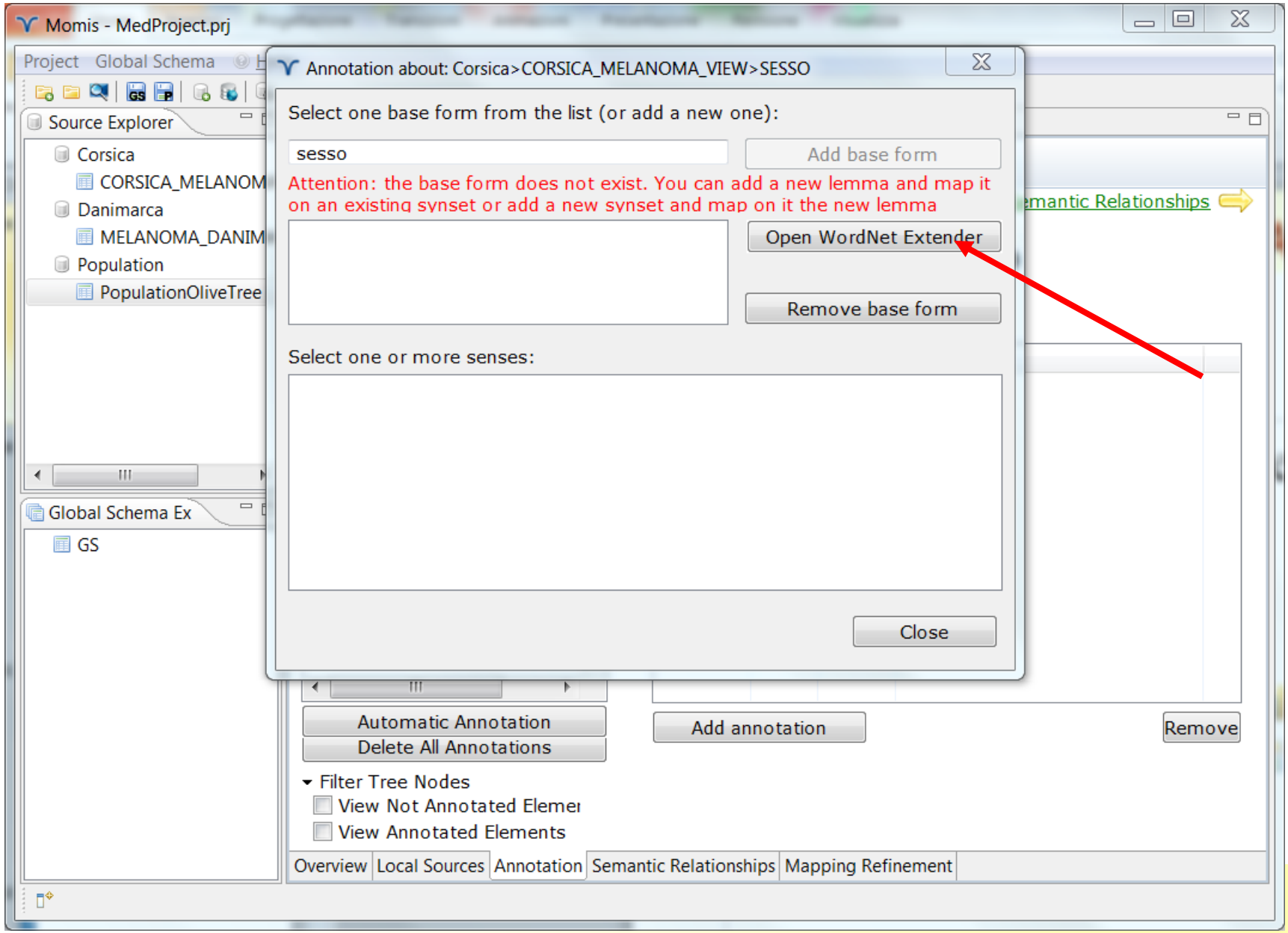
Select annotations for existing lemma



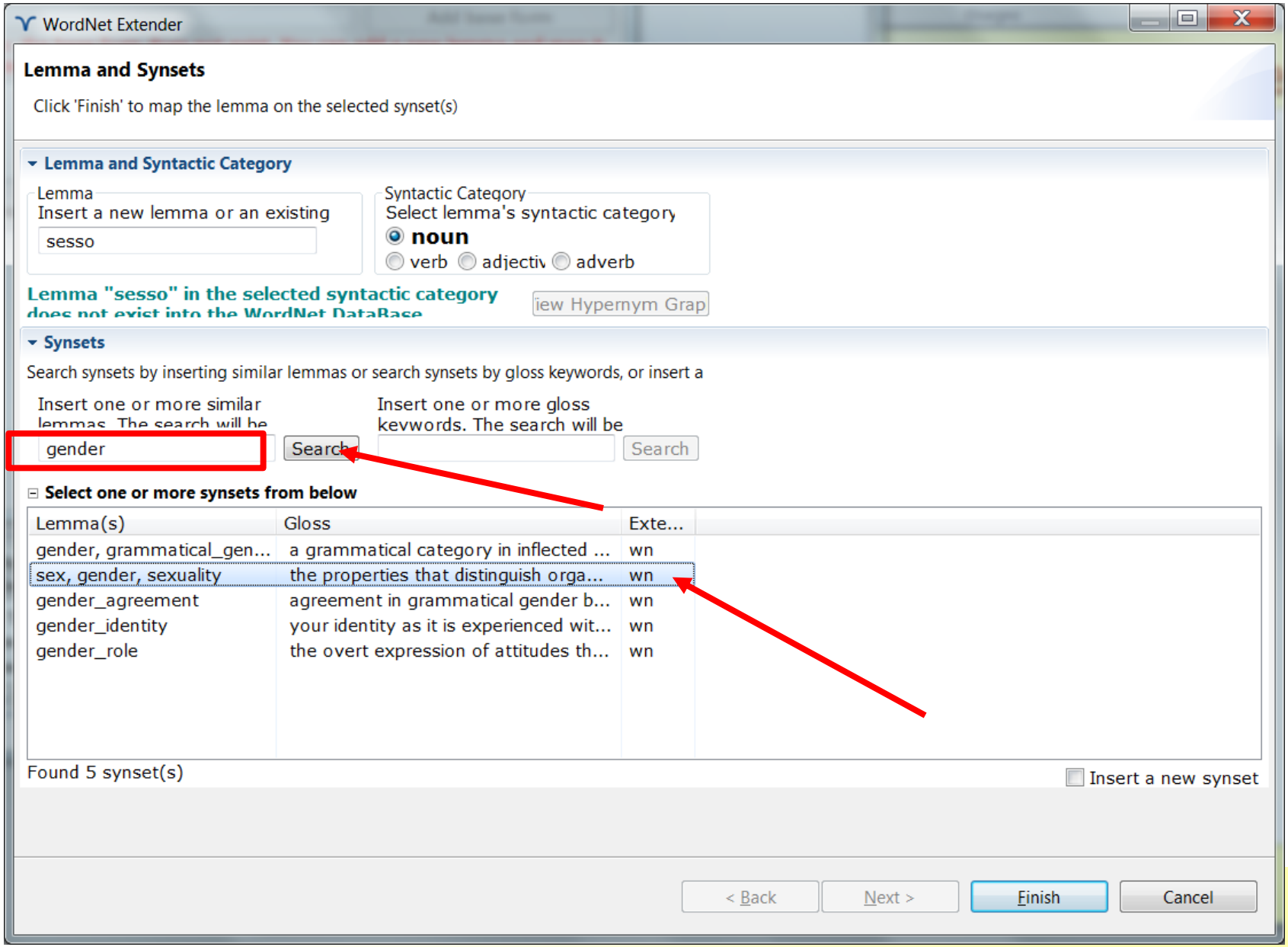
Change base form for lemma that are not present in WordNet

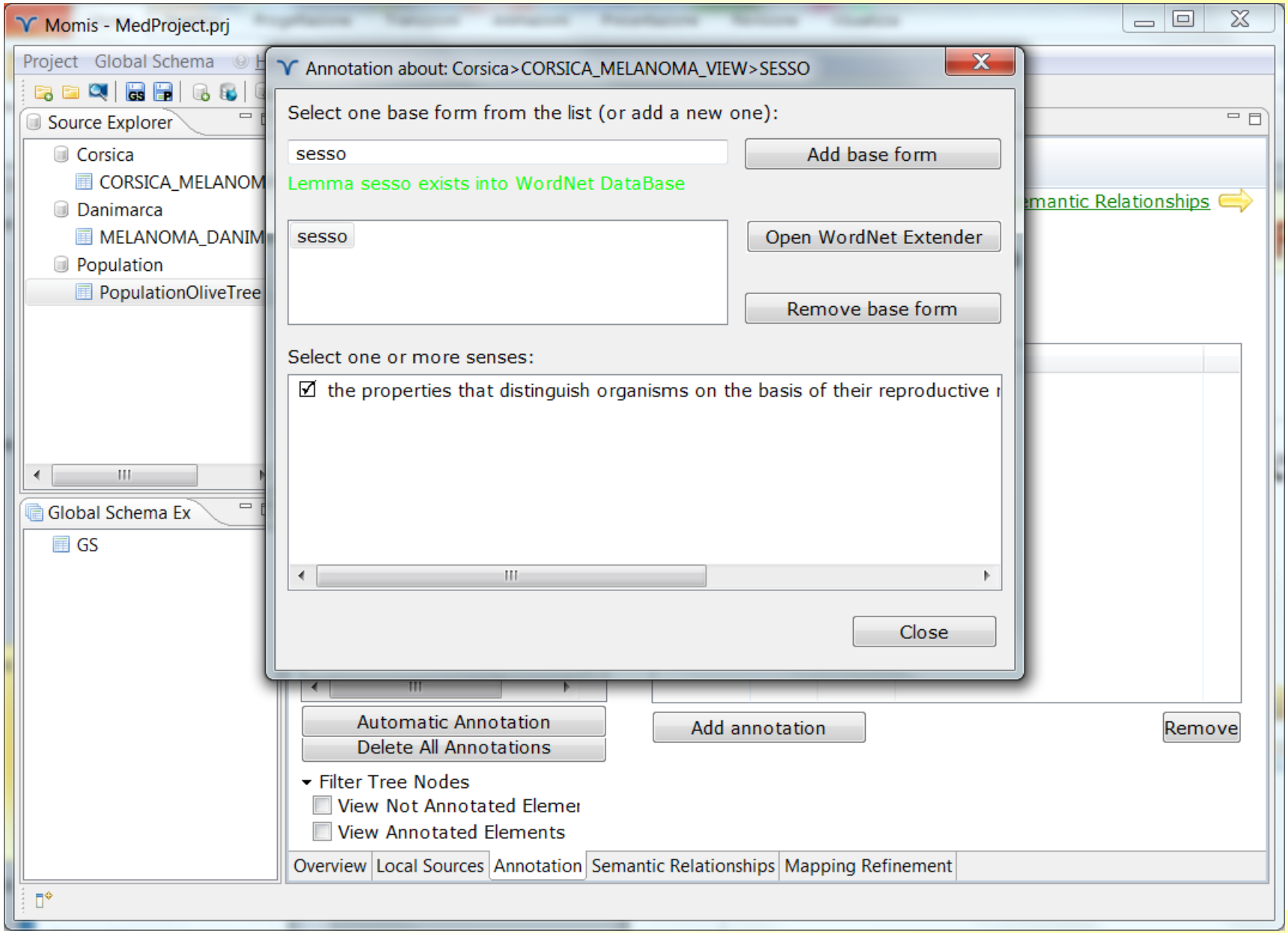


or...Add new lemma to WordNet

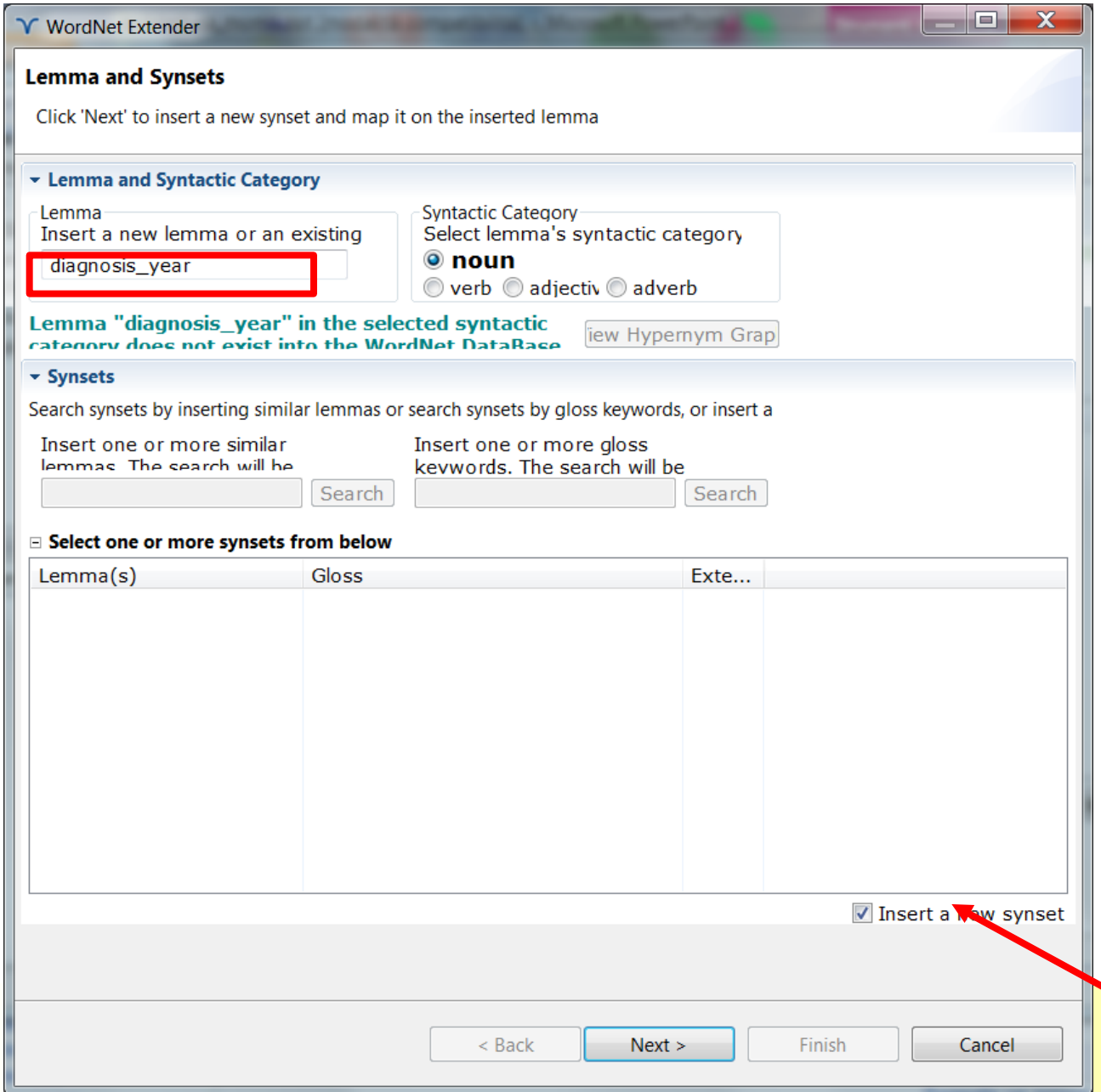


...by using WordNet Extender

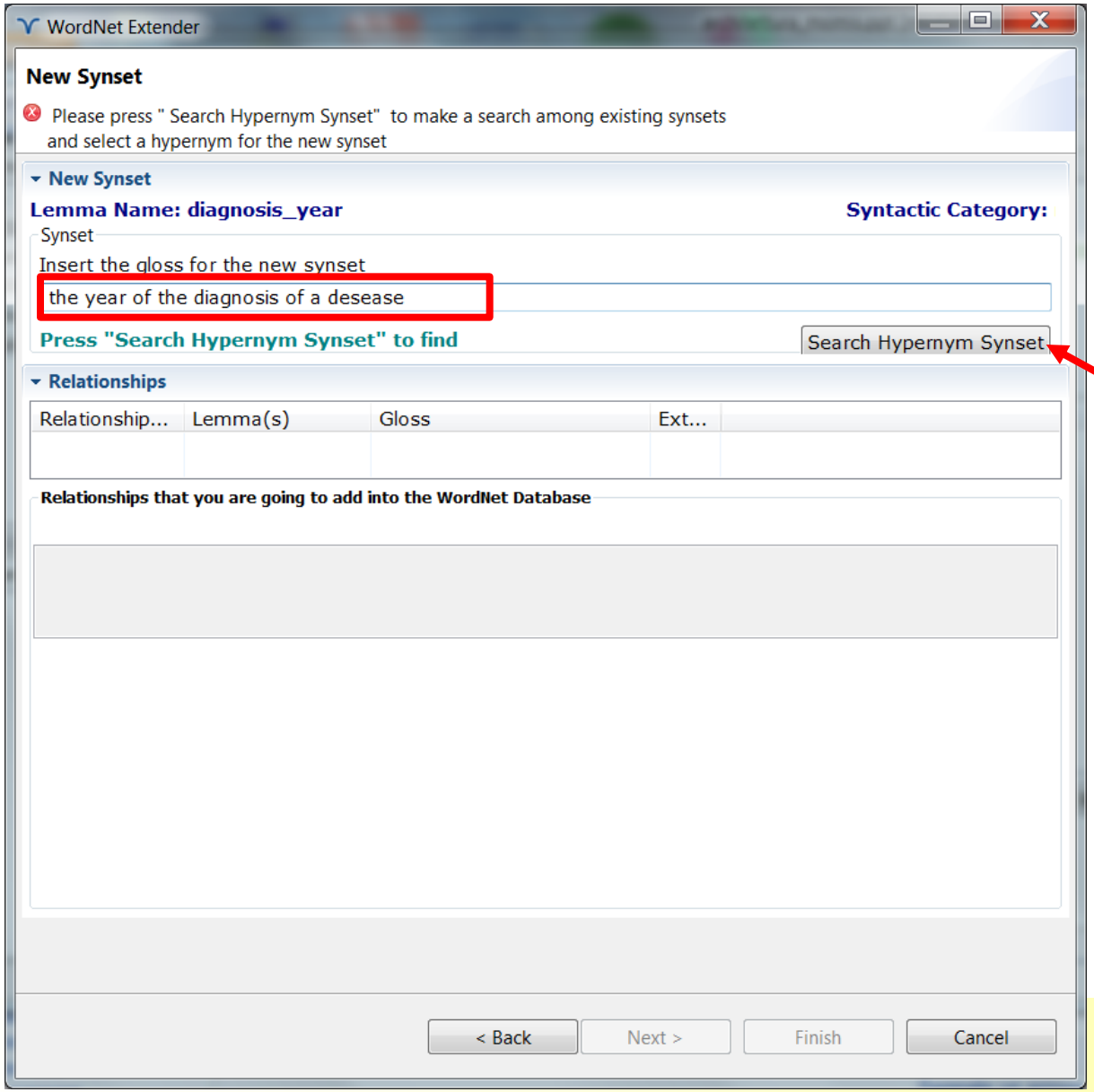



Momis - MedProject.prj
Project
Global Schema
Source Explorer
Corsica
CORSICA_MELANOM
Danimarca
MELANOMA_DANIM
Population
PopulationOliveTree
Global Schema Ex
GS
Annotation about: Corsica>CORSICA_MELANOMA_VIEW>SESSO
Select one base form from the list (or add a new one):
sesso
Add base form
Lemma sesso exists into WordNet DataBase
sesso
Open WordNet Extender
Remove base form
Select one or more senses:
the properties that distinguish organisms on the basis of their reproductive
Close
mantic Relationships
Automatic Annotation
Delete All Annotations
Add annotation
Remove
Filter Tree Nodes
View Not Annotated Eleme
View Annotated Elements
Overview
Local Sources
Annotation
Semantic Relationships
Mapping Refinement

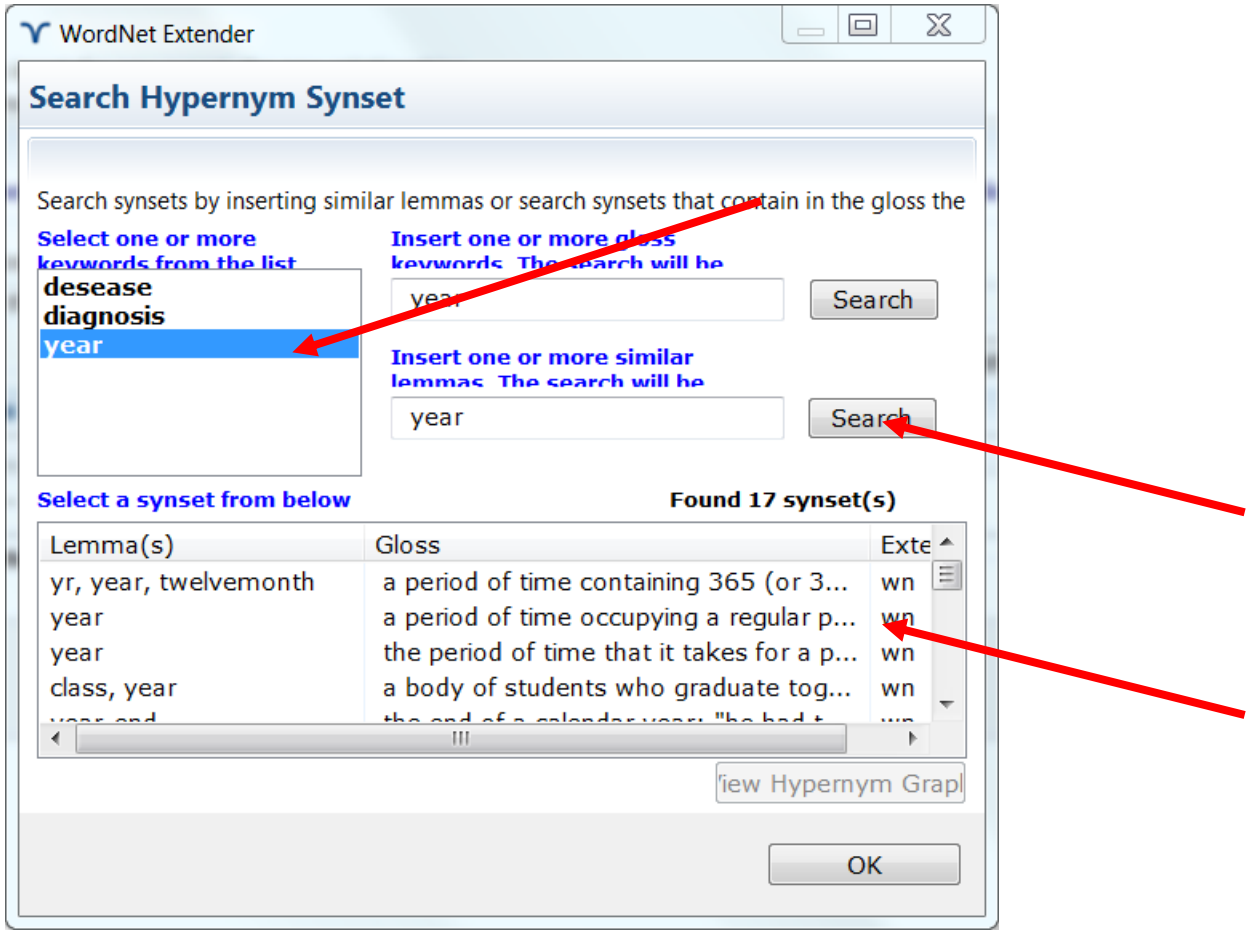
...also new meanings can be specify



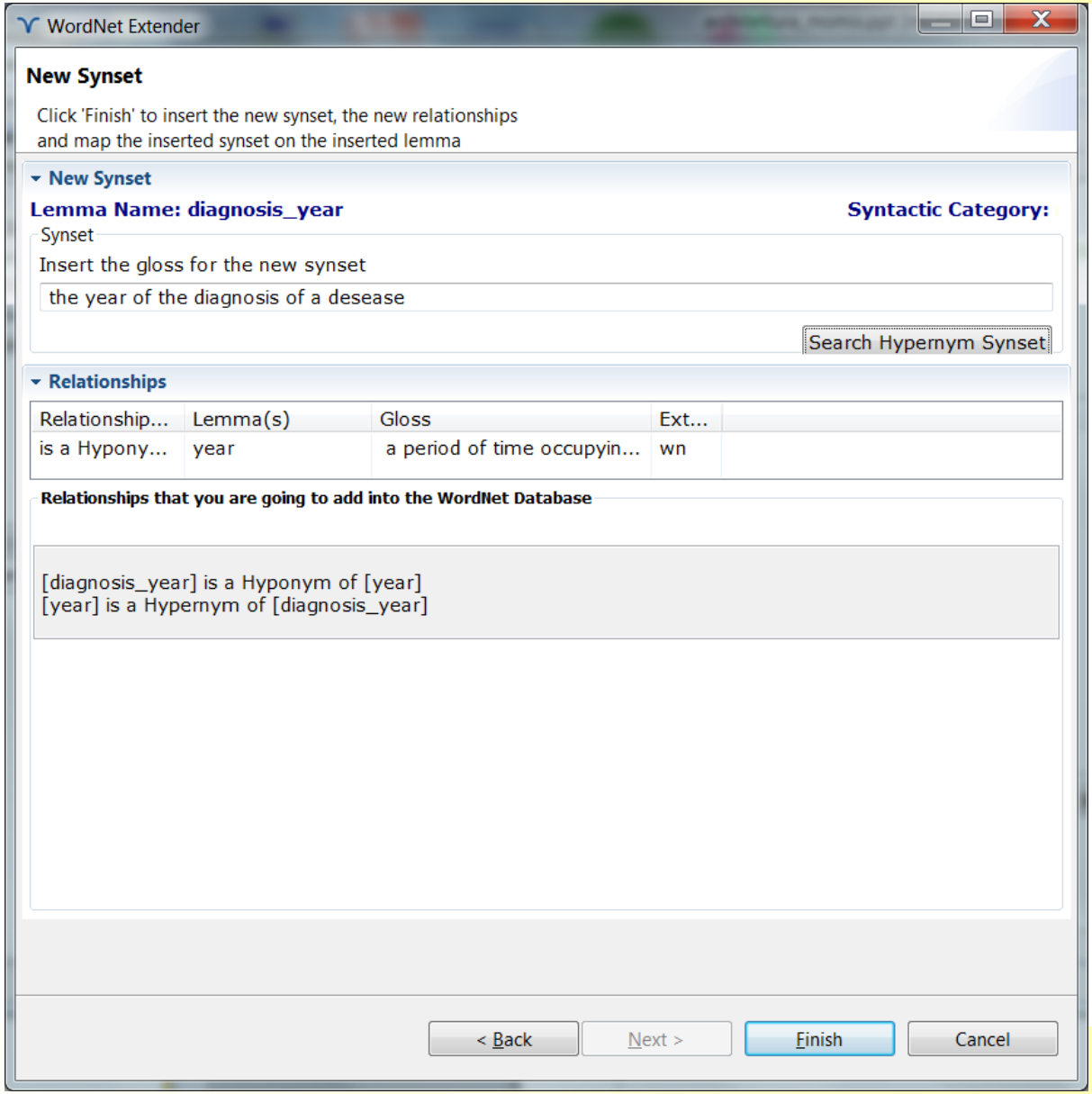
Define the new meaning

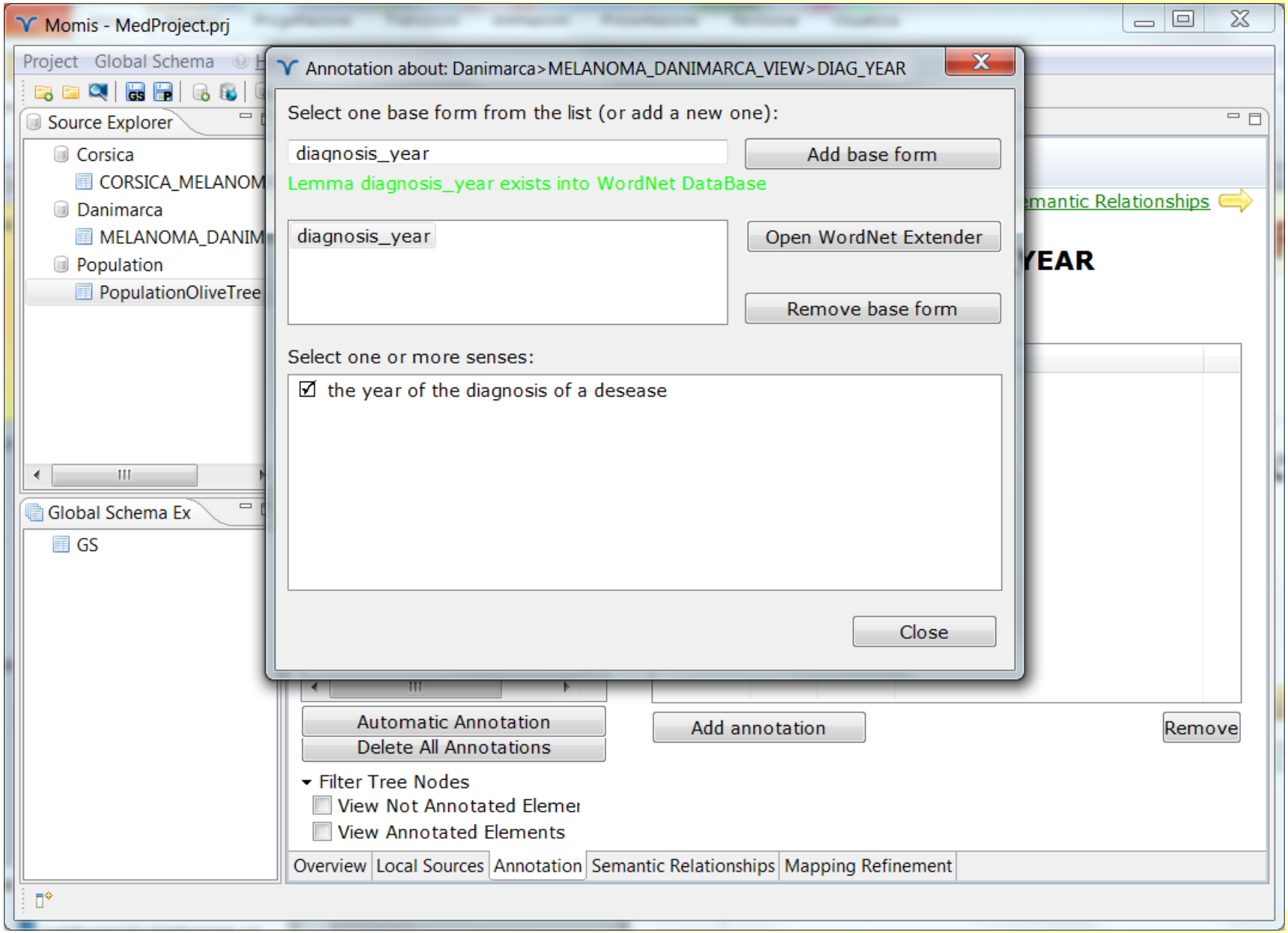


Select the correct hypernyms

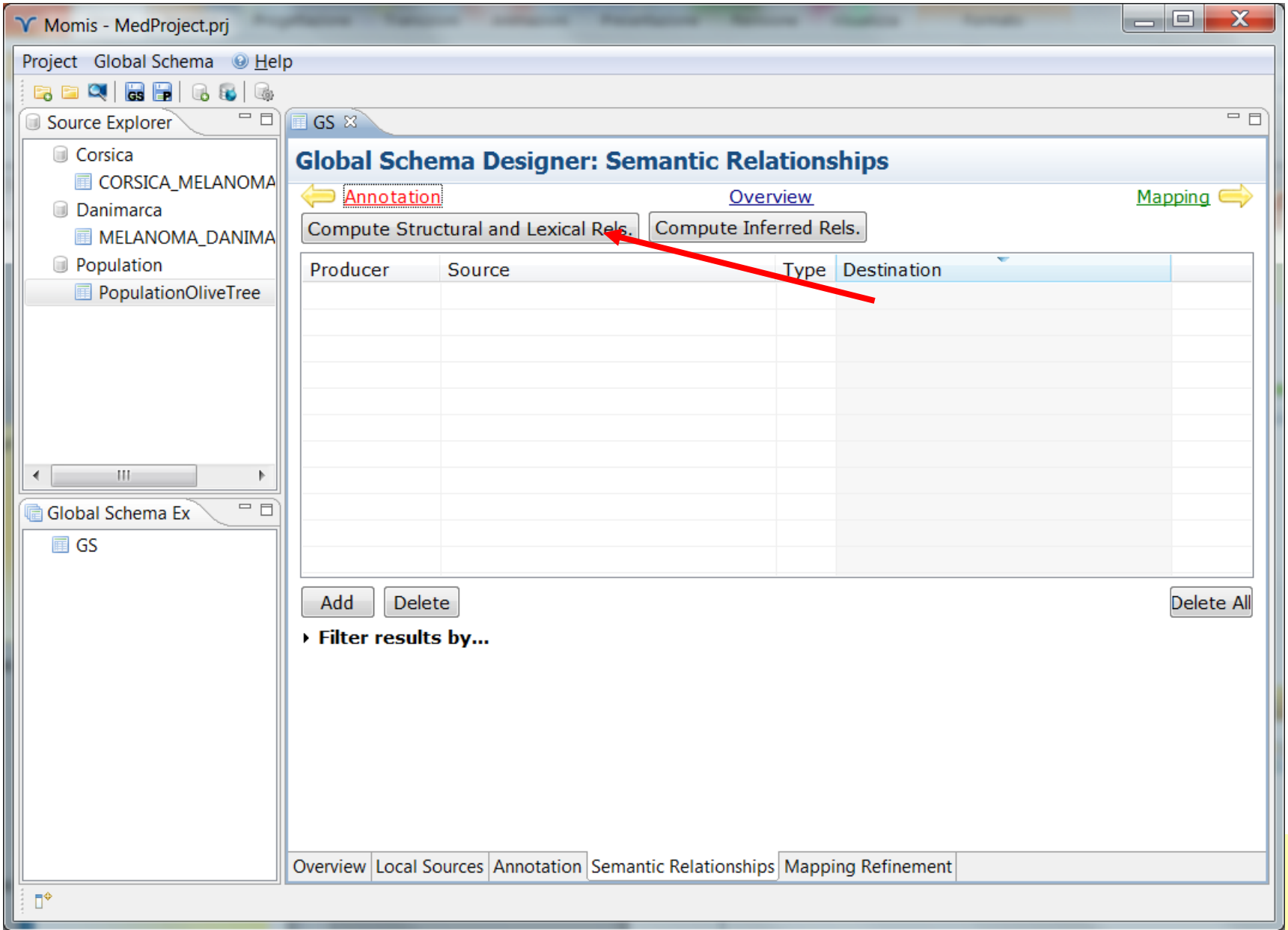


so that the new meaning is connected in the WordNet hypernym graph

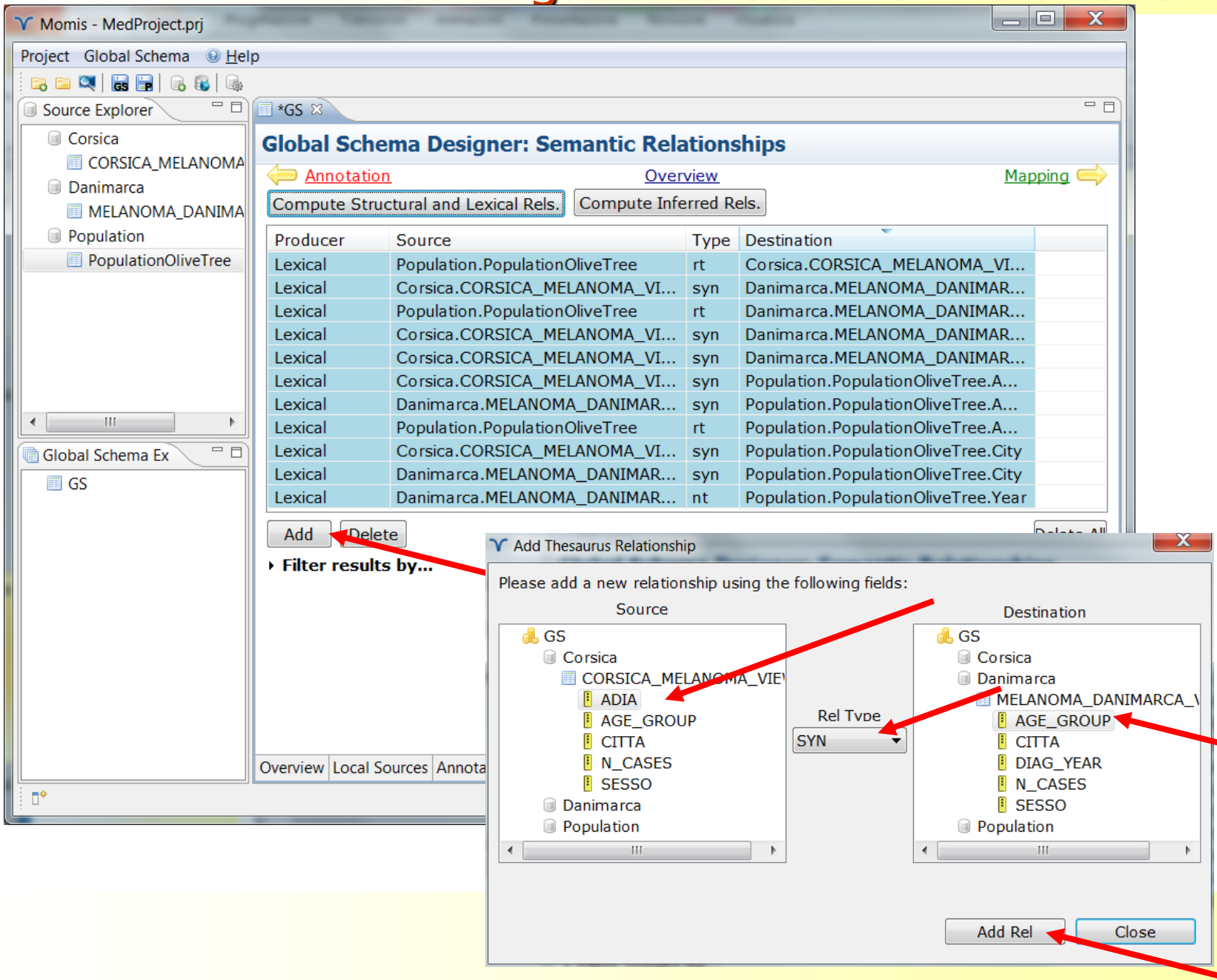



Momis - MedProject.prj
Project
Global Schema
Source Explorer
Corsica
CORSICA_MELANOM
Danimarca
MELANOMA_DANIM
Population
PopulationOliveTree
Global Schema Ex
GS
Annotation about: Danimarca>MELANOMA_DANIMARCA_VIEW>DIAG_YEAR
Select one base form from the list (or add a new one):
diagnosis_year
Add base form
Lemma diagnosis_year exists into WordNet DataBase
diagnosis_year
Open WordNet Extender
Remove base form
Select one or more senses:
the year of the diagnosis of a desease
Close
Semantic Relationships
YEAR
Automatic Annotation
Delete All Annotations
Add annotation
Remove
Filter Tree Nodes
View Not Annotated Elemer
View Annotated Elements
Overview
Local Sources
Annotation
Semantic Relationships
Mapping Refinement

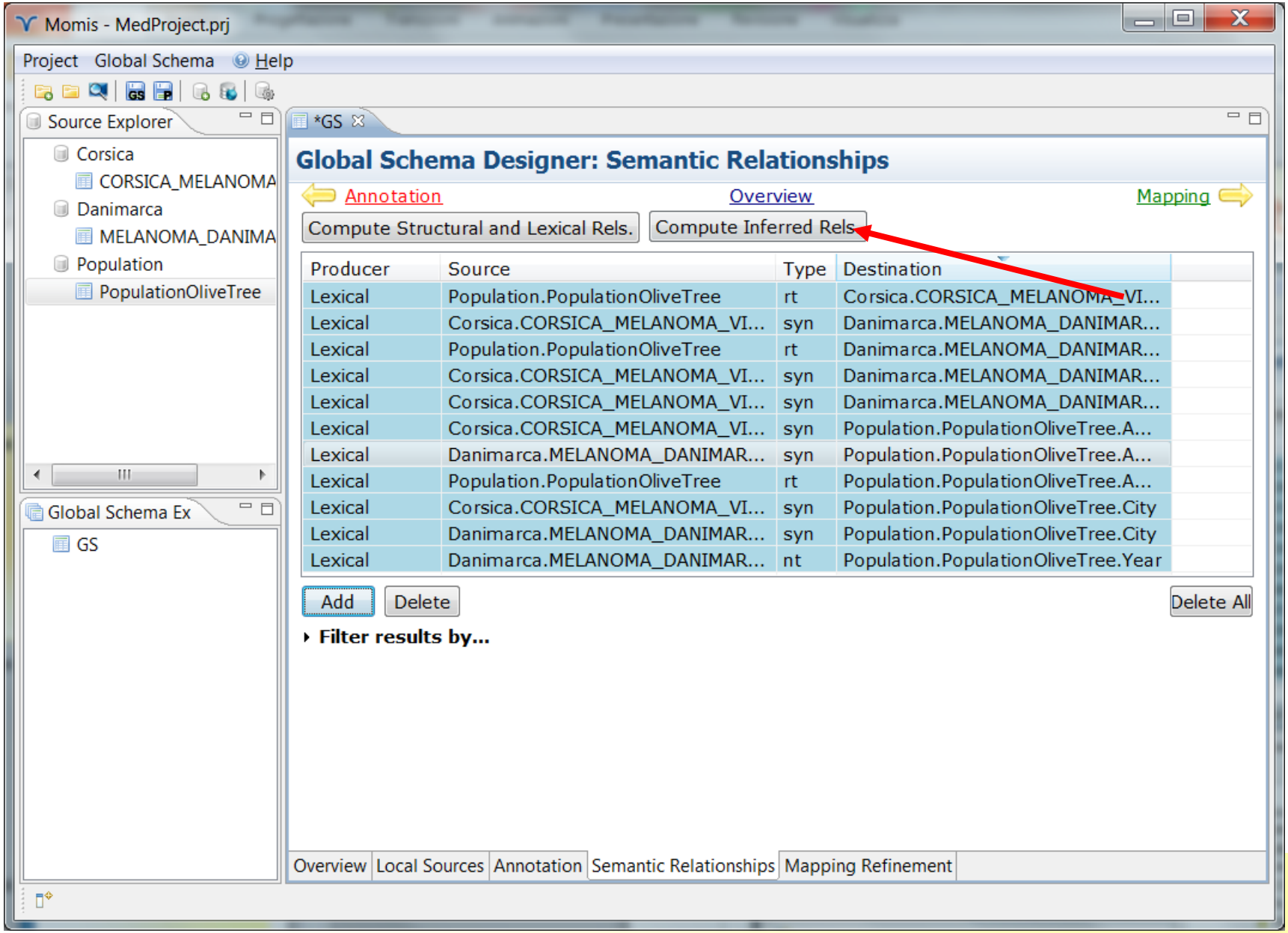
Once the annotation is finished you can proceed to step 3 – the computation of sematic relationship



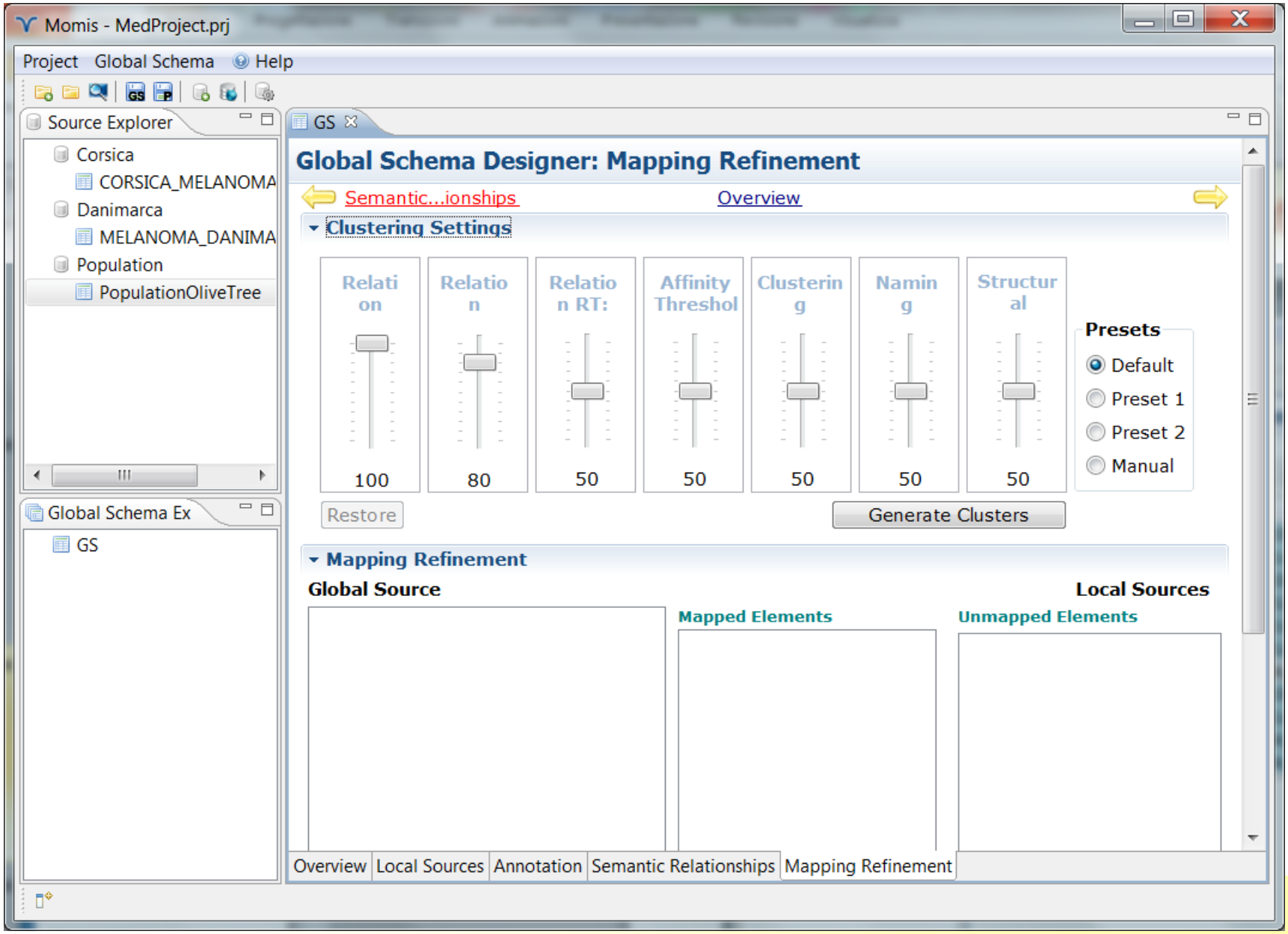
You might also add new relationships



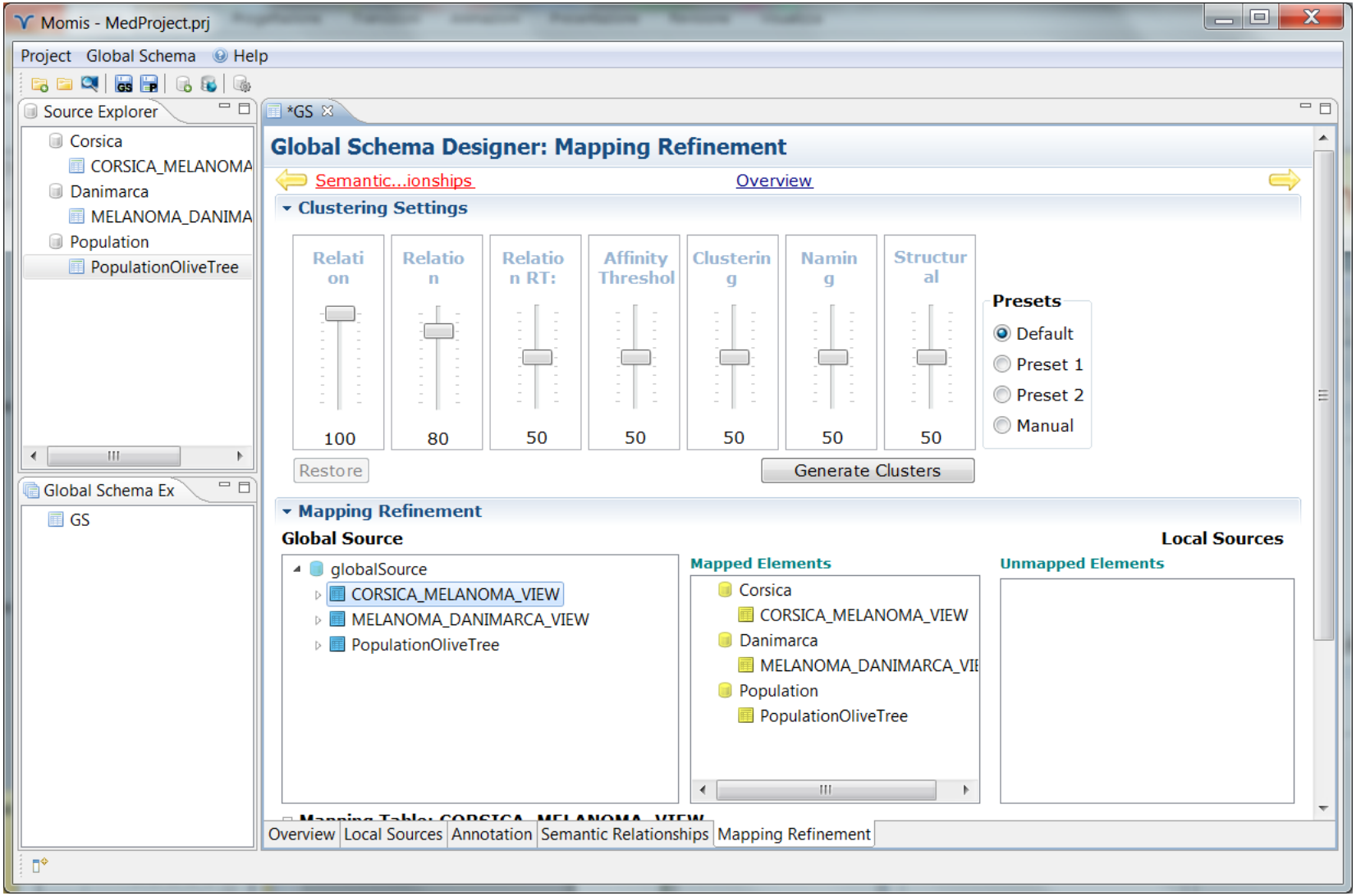
in the end, compute the inferred relationships



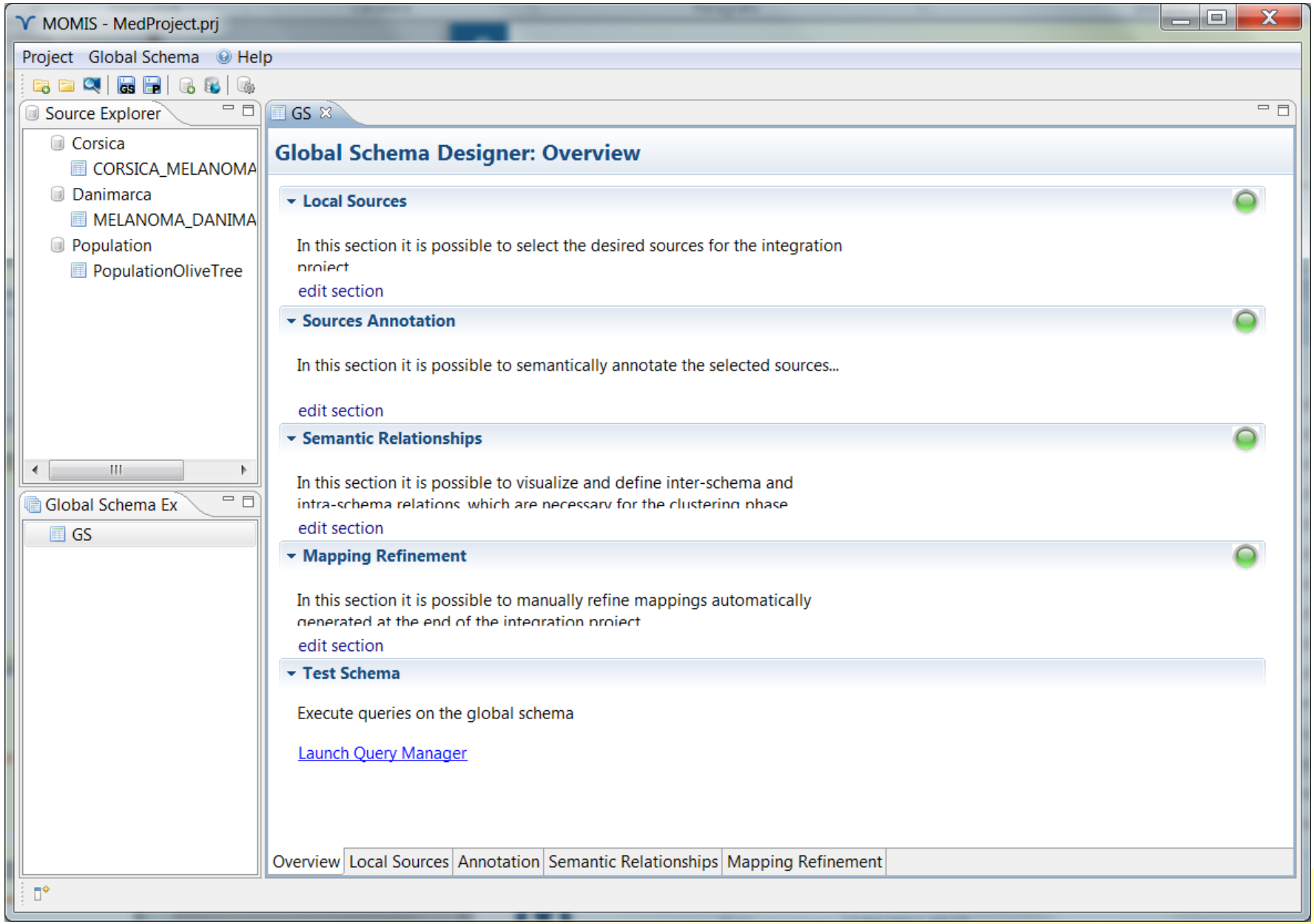
Step 4 – generation of clusters

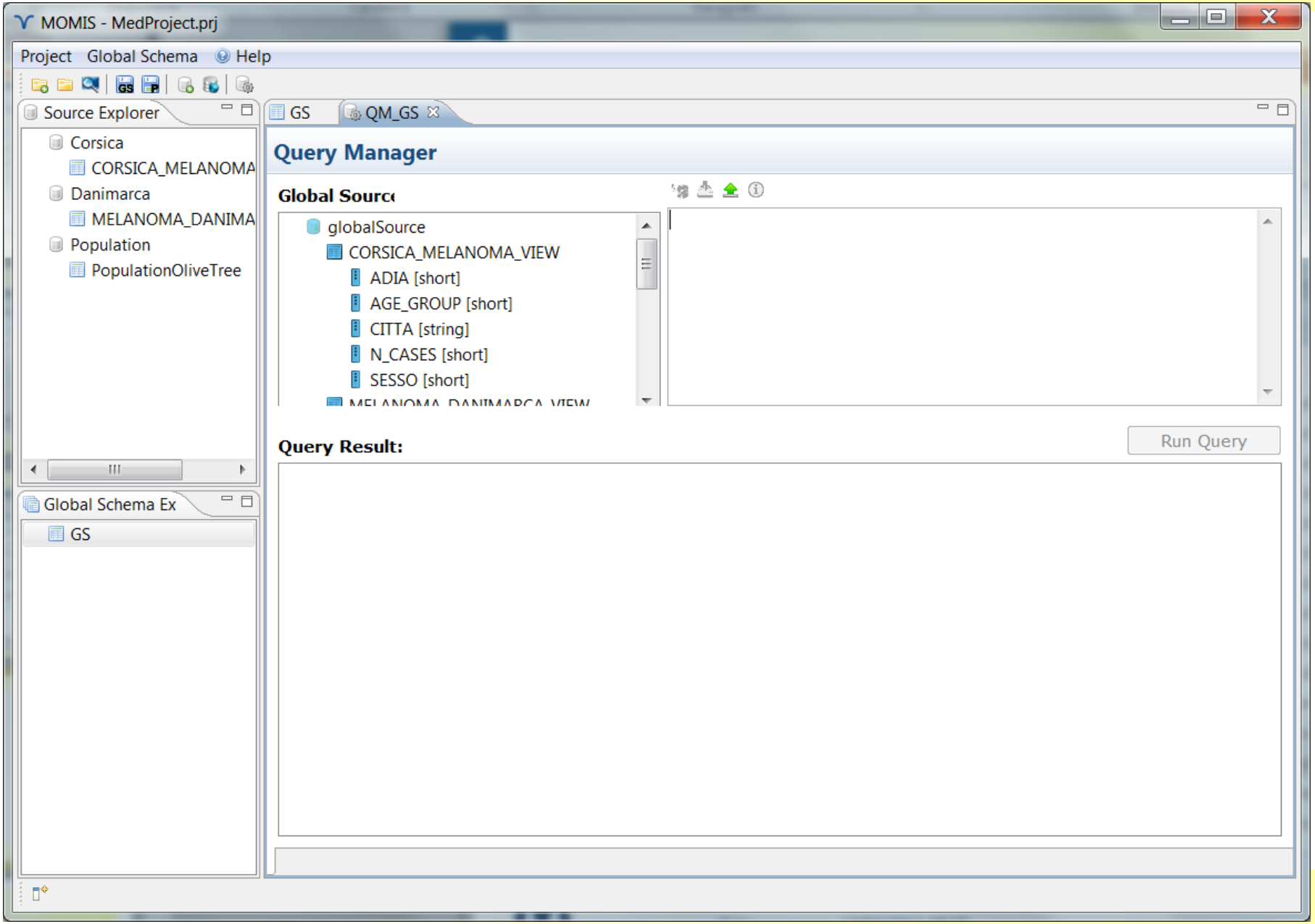


Refine the mapping



Query the GS


MOMIS - MedProject.prj
Project
Global Schema
Help
Source Explorer
Corsica
CORSICA_MELANOMA
Danimarca
MELANOMA_DANIMA
Population
PopulationOliveTree
Global Schema Ex
GS
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


MOMIS - MedProject.prj
Project
Global Schema
Help
Source Explorer
Corsica
CORSICA_MELANOMA
Danimarca
MELANOMA_DANIMA
Population
PopulationOliveTree
Global Schema Ex
GS
GS
QM_GS
Query Manager
Global Sourc
globalSource
CORSICA_MELANOMA_VIEW
ADIA [short]
AGE_GROUP [short]
CITTA [string]
N_CASES [short]
SESSO [short]
Query Result:
Run Query

TEST 2

Download the glossary

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

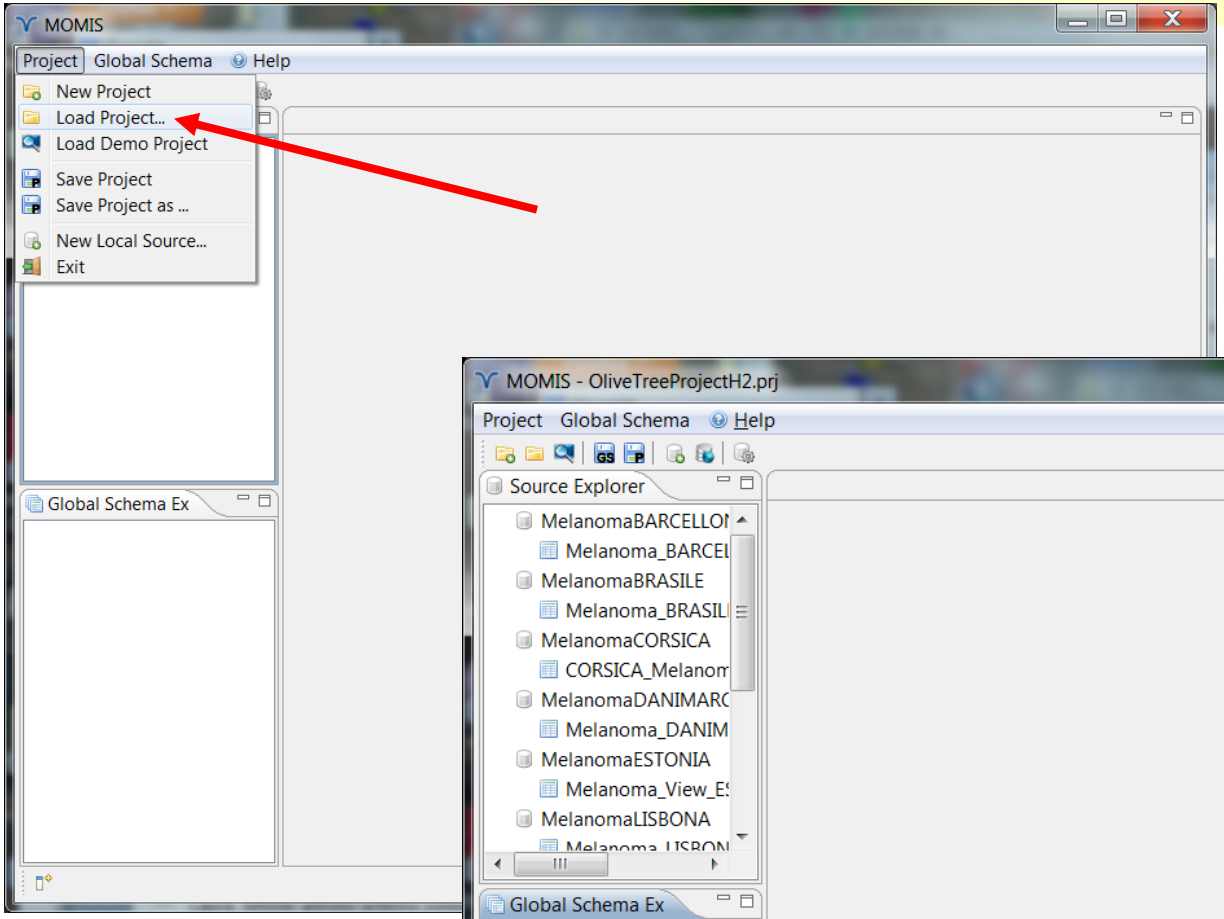
`MOMIS_1-2_win32_x86_JRE\MOMIS\hsql\`

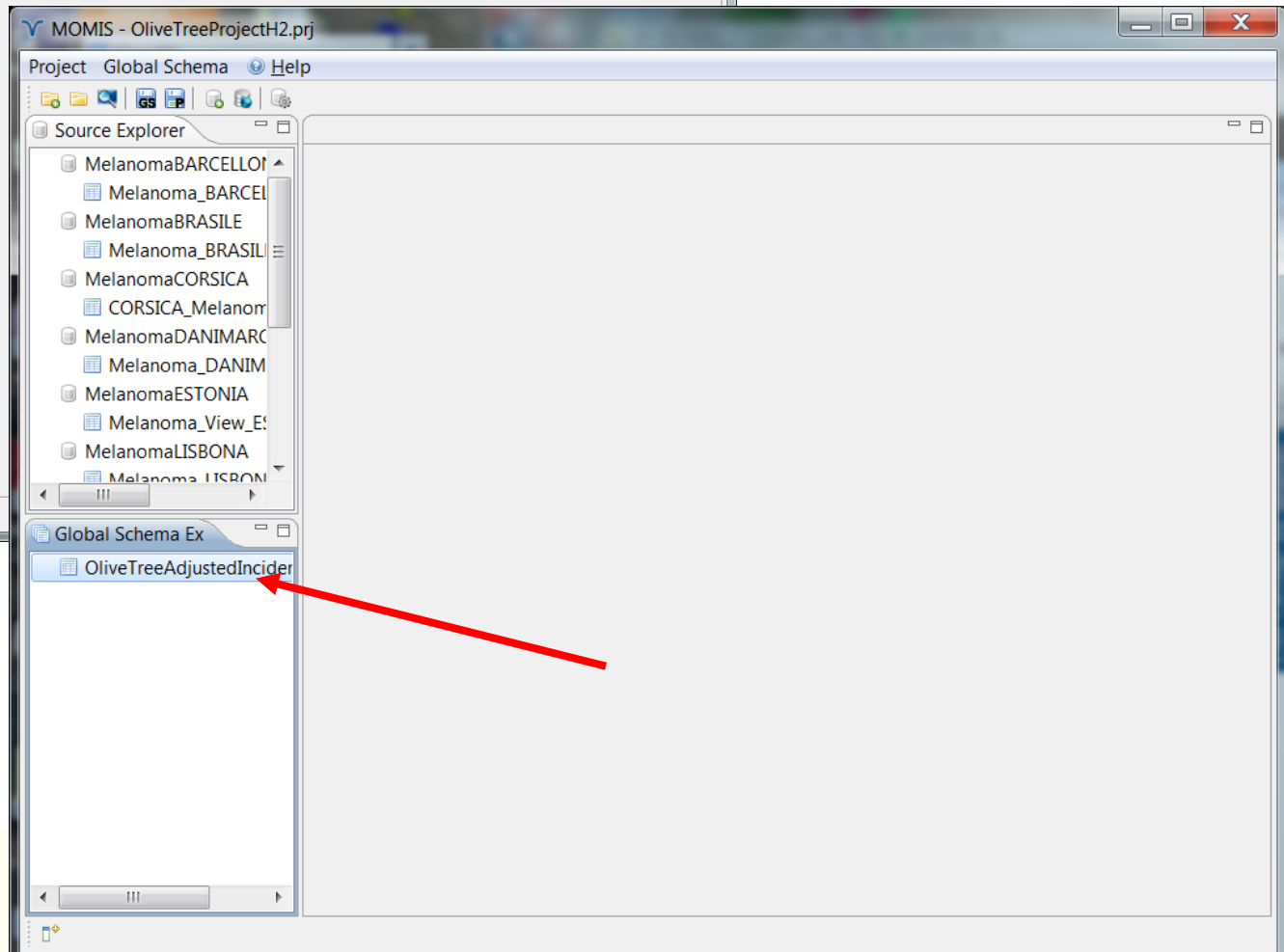
57

Explore and Query an existing project

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

58

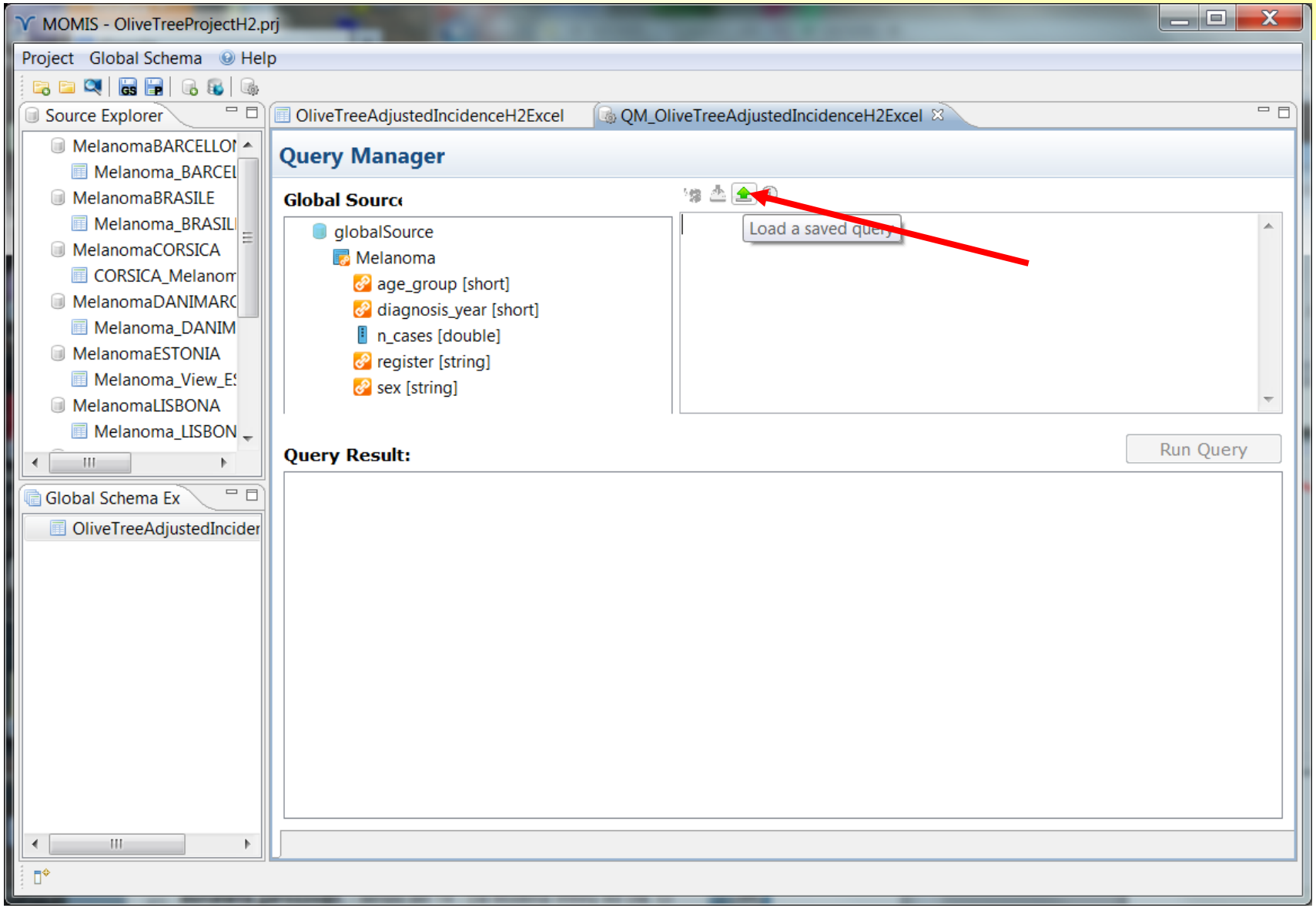

MOMIS
Project
Global Schema
Help
New Project
Load Project...
Load Demo Project
Save Project
Save Project as ...
New Local Source...
Exit
Global Schema Ex


MOMIS - OliveTreeProjectH2.prj
Project
Global Schema
Help
Source Explorer
MelanomaBARCELLO
Melanoma_BARCEL
MelanomaBRASILE
Melanoma_BRASIL
MelanomaCORSICA
CORSICA_Melanom
MelanomaDANIMARC
Melanoma_DANIM
MelanomaESTONIA
Melanoma_View_ES
MelanomaLISBONA
Global Schema Ex
OliveTreeAdjustedIncider

Query Manager

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

60


MOMIS - OliveTreeProjectH2.prj
Project
Global Schema
Help
Source Explorer
MelanomaBARCELLO
Melanoma_BARCEL
MelanomaBRASILE
Melanoma_BRASIL
MelanomaCORSICA
CORSICA_Melanom
MelanomaDANIMARC
Melanoma_DANIM
MelanomaESTONIA
Melanoma_View_ES
MelanomaLISBONA
Melanoma_LISBON
Global Schema Ex
OliveTreeAdjustedIncider
OliveTreeAdjustedIncidenceH2Excel
QM_OliveTreeAdjustedIncidenceH2Excel
Query Manager
Global Sourc
globalSource
Melanoma
age_group [short]
diagnosis_year [short]
n_cases [double]
register [string]
sex [string]
Load a saved query
Query Result:
Run Query

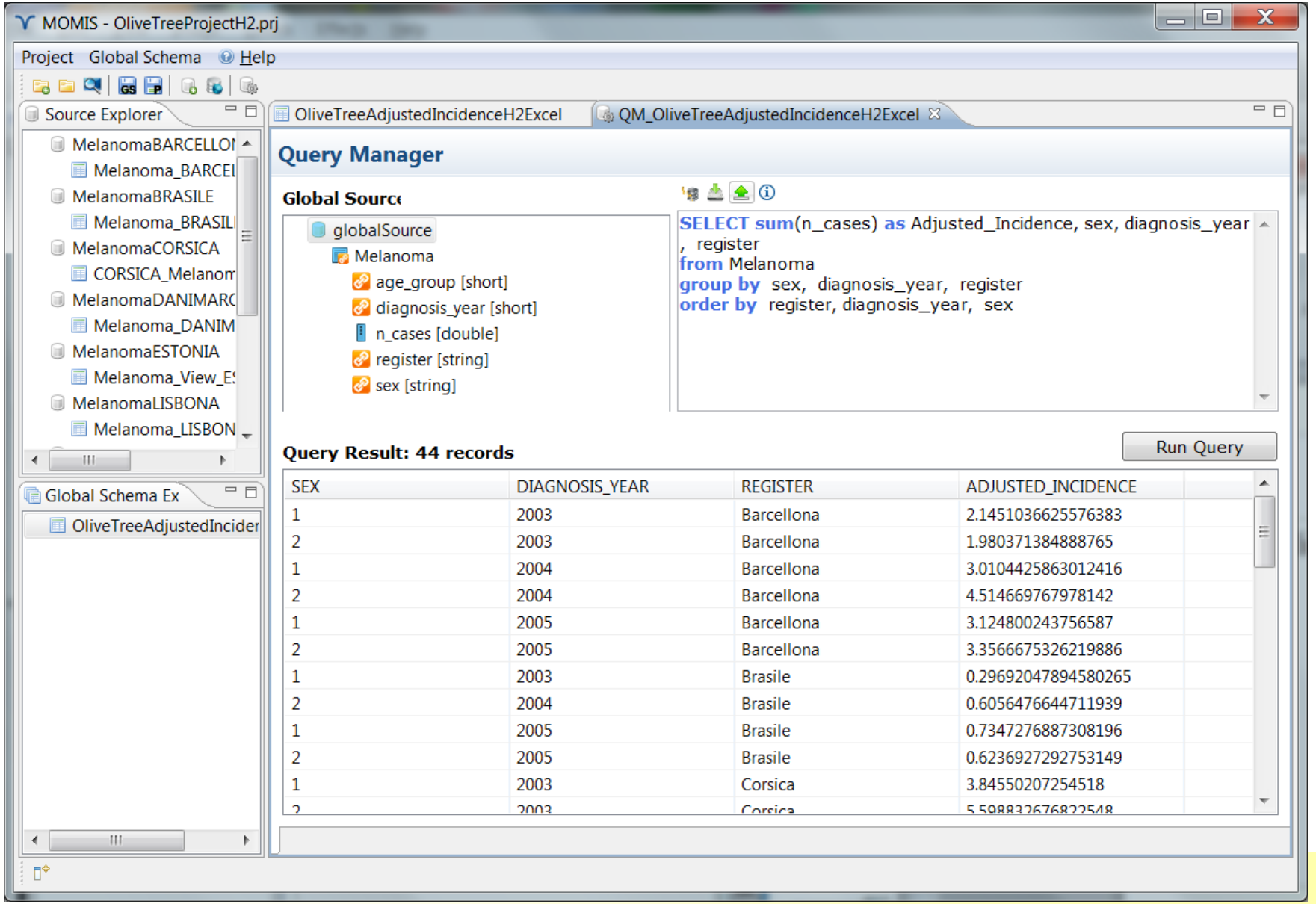
An example

The query

```
AdjustedIncidenceAllRegistries.oql
```

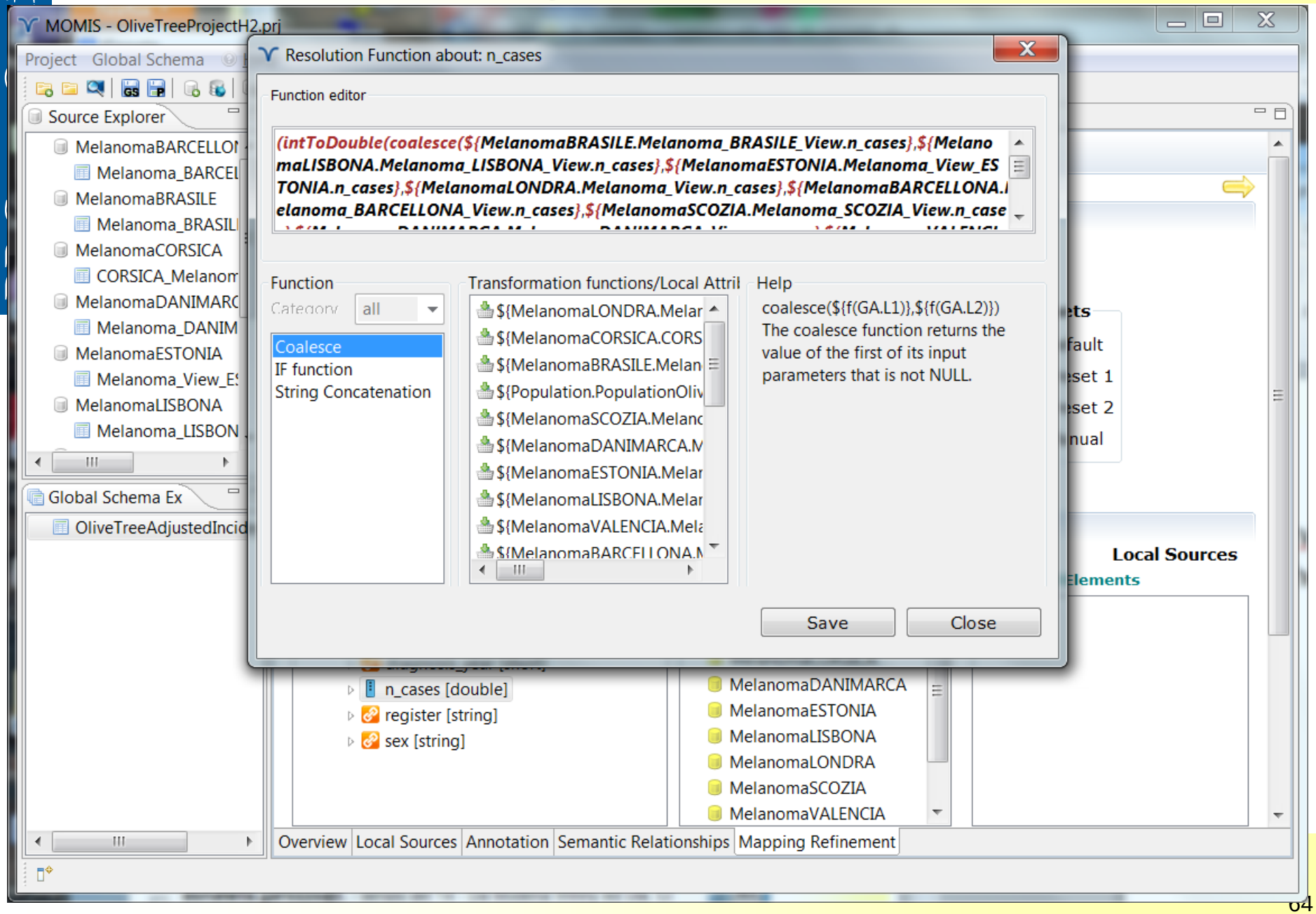
shows the incidence of occurrence of melanomas for a given geographical area and age group.

Incidence of Melanoma



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	5.598832676822548

The formula is expressed in the resolution function



ADDITIONAL MATERIAL

Common Thesaurus

- Set of intensional and extensional relationships expressing intra-schema and inter-schema knowledge
  - **Intensional Relationships**
    between class and attribute names (T)
    < Ti SYN Tj > *Synonymy*
    < Ti NT Tj > *(Narrower Term - NT)*
    < Ti RT Tj > *(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 ⇒ SYN
    - Hyponymy ⇒ NT
    - Meronymy and Correlation ⇒ RT

Annotation and Lexicon-derived Relationships

(Narrower Term)
NT

	Word form		
Meaning (synset)	*Book*	*Volume*	*Publication*
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 doorstop"	✕	✕	
the amount of 3-dimensional space occupied by an object		●	
a copy of a printed work offered for distribution			✕

Hyponymy (is a kind of)

Lexicon derived relationships

Book	SYN	Volume
Book	NT	Publication

Extending WordNet

- **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.

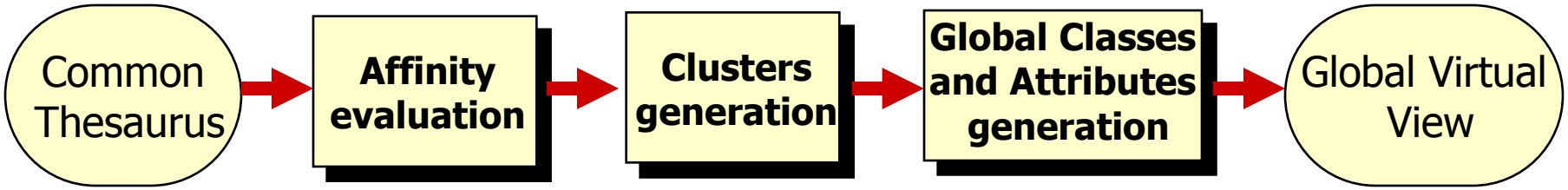
- The designer may add a new meaning (for an existing word-form or for a new one) by:
  - writing the gloss explicitly, or
  - using an existing synset chosen among a list of candidates obtained by an explicit search (using one or more keywords) or by exploiting similarity search techniques.

- 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=(**L**,**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=(**L**,**GA**), a *Mapping Table* (MT) is generated, to represent the mappings between global and local attributes
  - MT is a table **GA**x**L** : 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)

Mapping Table Refinement

- 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** $\mathbf{V_C}$ over the local classes of C.

- The integration designer, supported by the Ontology Builder graphical interface, can implicitely **define** $\mathbf{V_C}$ 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

Data fusion: Full Join Merge

- 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

SUM

Join Conditions

Object identifier

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

Resolution Functions

AVG

```
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

Full Join Merge

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  AS A,
         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			

Global Query Management

- **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 funtions to FOJ

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

Query Resul: Application of the residual conditions to RES_FOJ

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