Unleashing the Power of Ontologies in Data Integration

The SELEX Case Study

Diego Calvanese

Free University of Bozen-Bolzano

Semantic Days

Stavanger, Norway – 19 May, 2009

Outline of the Talk

1. Ontology-based Data Integration and the QuOnto System

2. Configuration and Data Management (C&DM) at SELEX Sistemi Integrati (SELEX-SI)

3. Experiencing QuOnto for C&DM at SELEX-SI

4. Conclusions

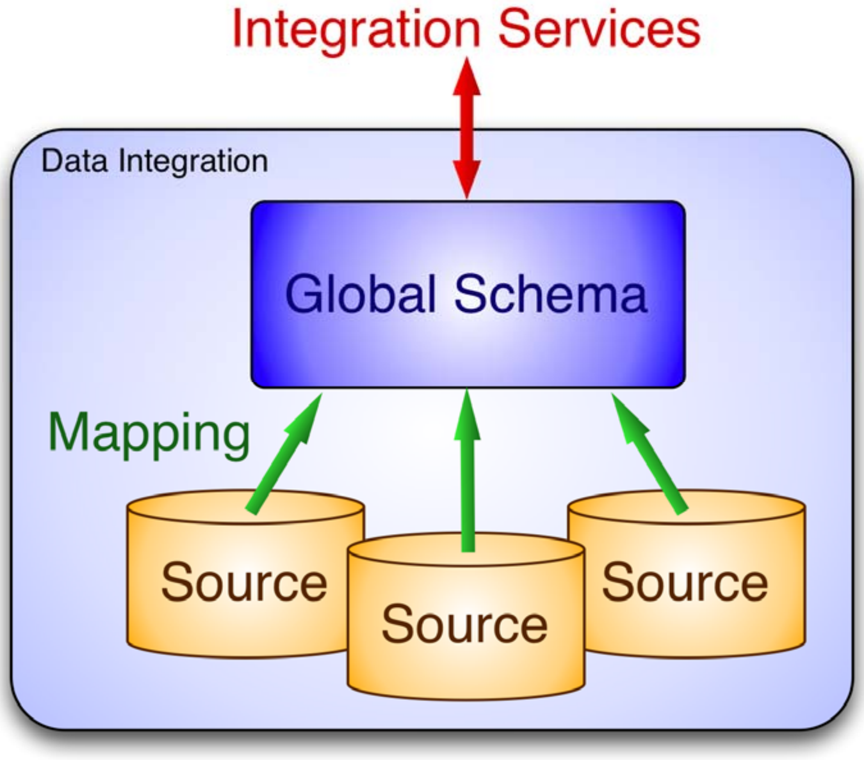
Data Integration

- Data Integration is the problem of providing a unified and transparent access, through a global schema to a collection of data stored in multiple, autonomous, and heterogeneous data sources.

- From [Bernstein & Haas, CACM Sept. 2008]:
  - Large enterprises spend a great deal of time and money on information integration (e.g., 40% of information-technology shops' budget).
  - Market for data integration software estimated to grow from $2.5 billion in 2007 to $3.8 billion in 2012 (+8.7% per year) [IDC. Worldwide Data Integration and Access Software 2008-2012 Forecast. Doc No. 211636 (Apr. 2008)].

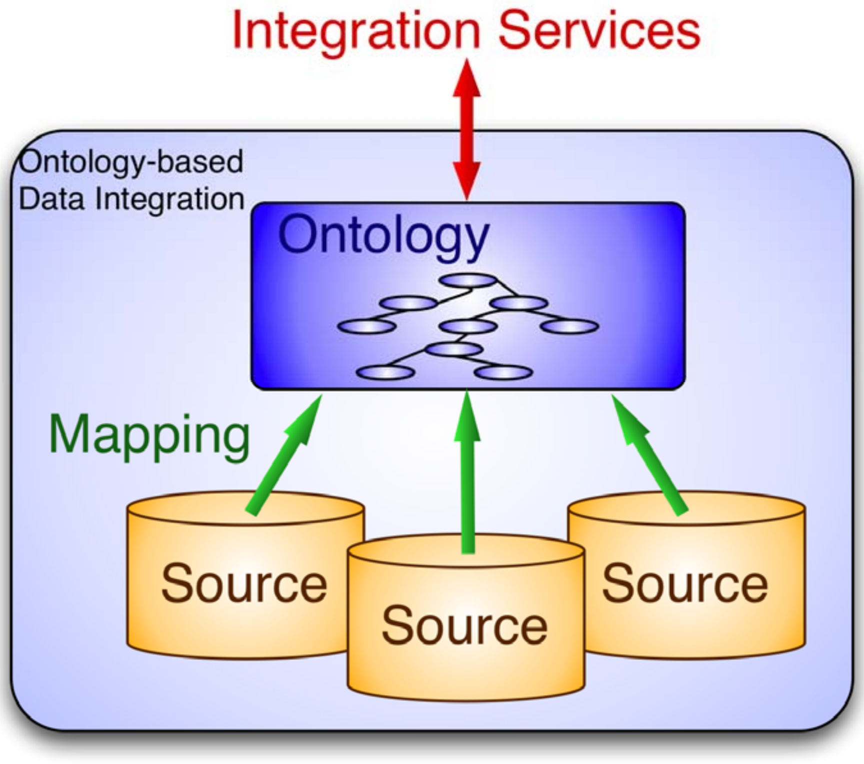
Conceptual Architecture for Data Integration

- A global schema and various data sources.
- Mappings relate data sources to global schema.
- Integration Services (e.g., query answering) are expressed over the global schema.

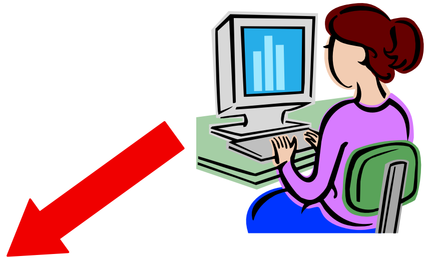


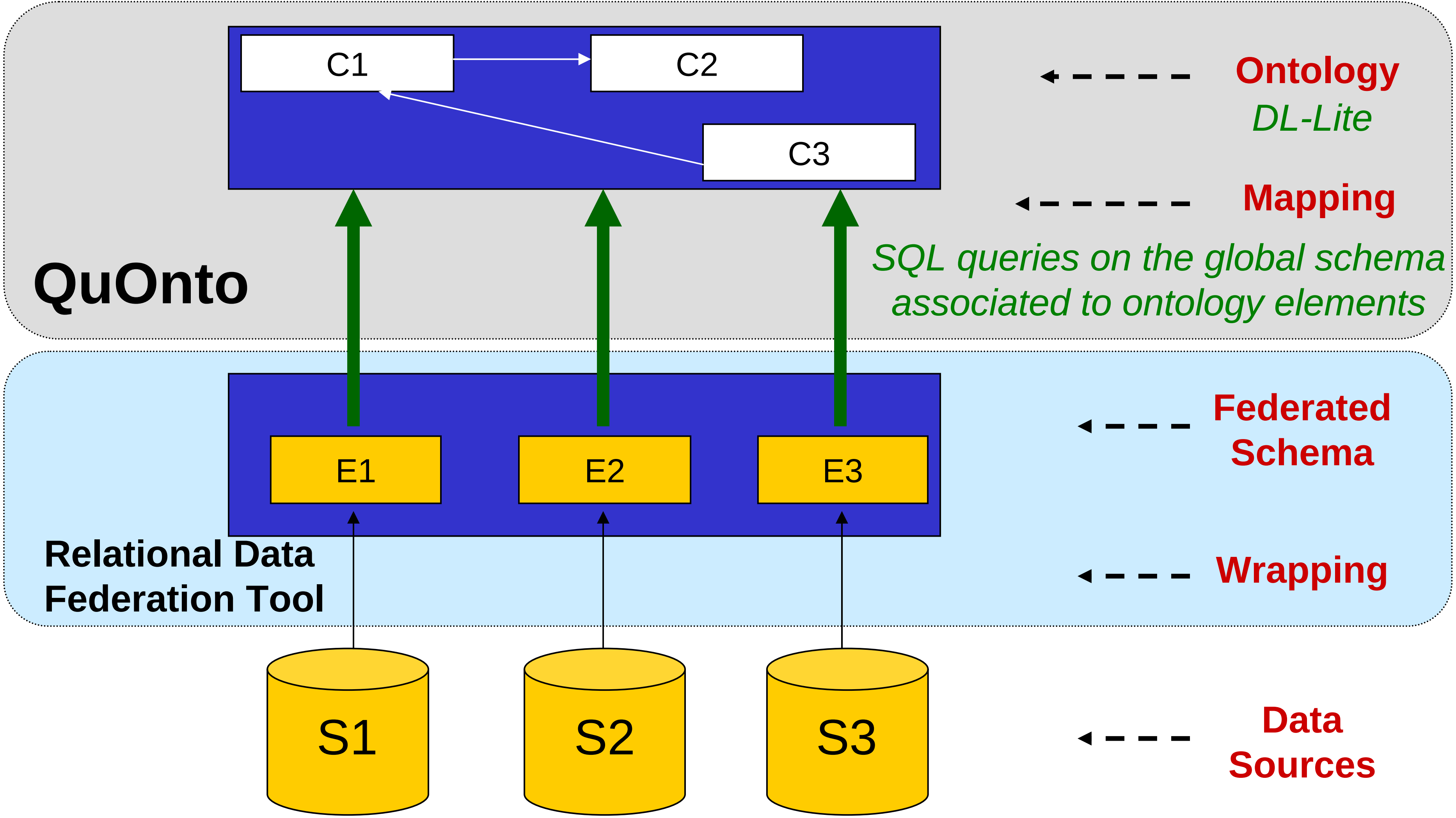
Ontology-based Data Integration

- The global schema is represented through an ontology.
- We assume the data sources to be relational.
- The mappings specify how to construct objects in the ontology from the data items in the sources.



QuOnto Integration Architecture





C1

C2

C3

QuOnto

E1

E2

E3

Relational Data Federation Tool

S1

S2

S3

Ontology

DL-Lite

Mapping

SQL queries on the global schema associated to ontology elements

Federated Schema

Wrapping

Data Sources

Data Integration through QuOnto

- Global schema - *DL-Lite Ontology*

  DL-Lite [C. et al. JAR-07, JODS-08] is a tractable Description Logic (DL) that captures basic ontology languages and allows for query answering through relational database technology.

  The global schema is a set of assertions over concepts and roles, i.e., binary relations between concepts (essentially an UML class diagram).

- Data Sources - represented by a *relational schema*

  This schema can be obtained by means of a data federation tool which manages source wrapping (we call it federated schema).

- Mappings - *Global-As-View (GAV)*

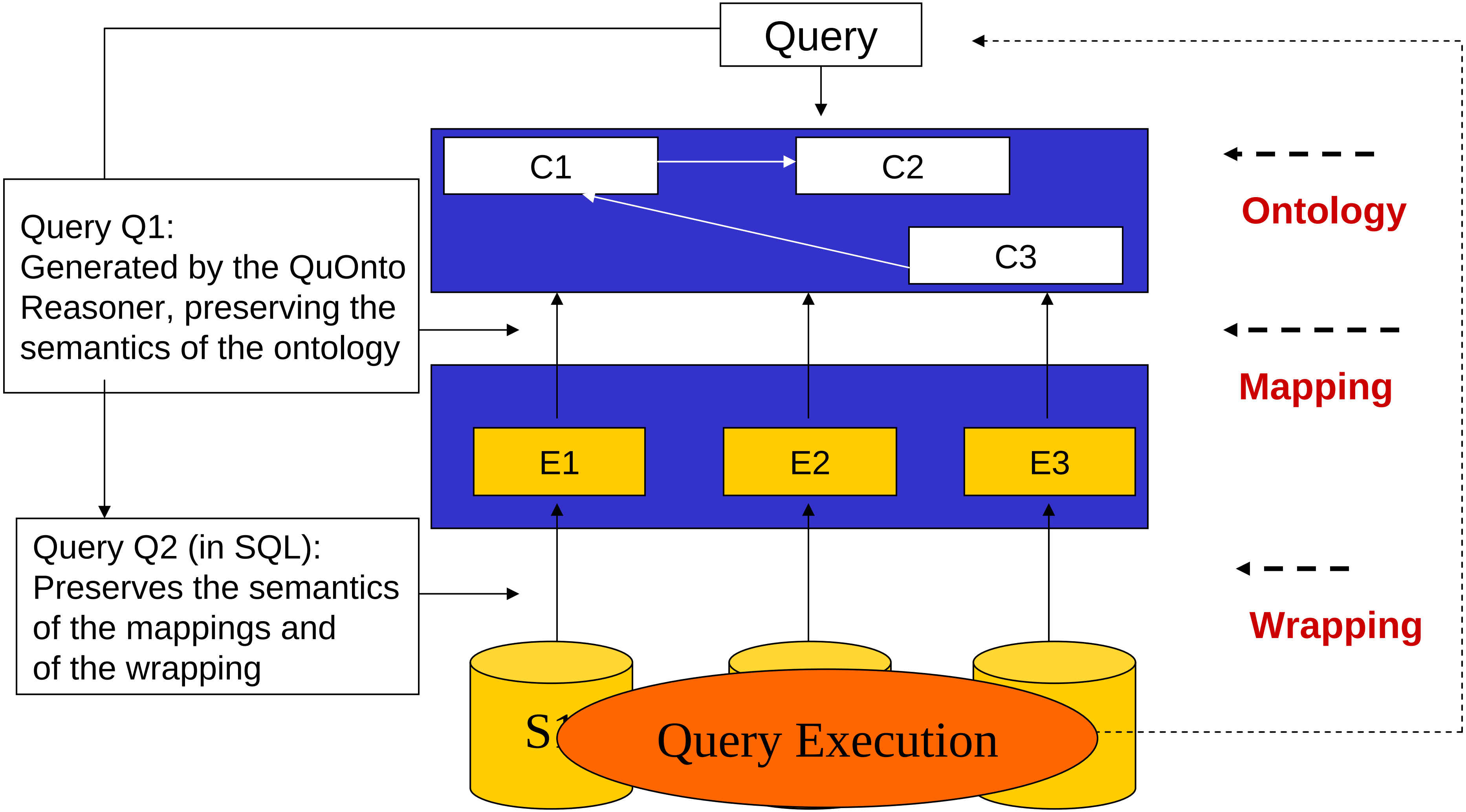
  I.e., a set of assertions of the form

  $$\boldsymbol{C} \leftarrow \boldsymbol{Q}$$

  *where **C** is an element of the global schema and **Q** is an SQL query over the federated schema.*

If we go beyond the above expressiveness the system loses its nice computational properties [C. et al. SKDB-08, KR-06].

Query answering in QuOnto



Query

C1

C2

C3

Ontology

Query Q1:
Generated by the QuOnto Reasoner, preserving the semantics of the ontology

Mapping

E1

E2

E3

Query Q2 (in SQL):
Preserves the semantics of the mappings and of the wrapping

Wrapping

S1

Query Execution

Outline of the Talk

1. Ontology-based Data Integration and the QuOnto System
2. Configuration and Data Management (C&DM) at SELEX Sistemi Integrati (SELEX-SI)
3. Experiencing QuOnto for C&DM at SELEX-SI
4. Conclusions

C&DM at SELEX-SI

- SELEX-SI is a Finmeccanica company that is world leader in the provision of integrated defence, air traffic, and mission critical systems, with customer base in over 150 countries.
- C&DM is a technical management model that governs the entire products' life cycle, enforcing product consistency with respect to requirements, design, and operational data.
- SELEX-SI produces and maintains systems with very long life cycle, which require a correct configuration management after delivery → C&DM is "The hub of the wheel" in SELEX-SI.
- C&DM in SELEX-SI involves three main processes: Project & Product, Manufacturing, and In-Service Config. Management.
- In this case study, we mainly focused on Manufacturing and In-Service CM, and in particular on:
  - component design and production
  - component deployment
  - analysis of component's obsolescence

Data Integration for C&DM

- Currently, *several different tools* are used for the various C&DM processes (e.g., RDBMS-based tools like SAP R3, SAP Customer Support, Odb, or XML-based tools like eDEA).
- This results in a set of heterogeneous data sources, completely autonomous or weakly integrated, managing overlapping data.
- Data integration is *manually* performed by *C&DM experts*, with great efforts in terms of time and resources and no guarantees on reliability and effectiveness of the retrieved information.
- Desiderata: Simplify and automatize the data integration process!
- Our Solution: Integrate C&DM data sources through the *QuOnto* ontology-based data integration management system

Outline of the Talk

1. Ontology-based Data Integration and the QuOnto System

2. Configuration and Data Management (C&DM) at SELEX Sistemi Integrati (SELEX-SI)

3. Experiencing QuOnto for C&DM at SELEX-SI
   i. Federating C&DM Data Sources
   ii. Designing the SELEX-SI C&DM Ontology
   iii. Relating Data Sources to the SELEX-SI C&DM Ontology through QuOnto
   iv. Querying the System

4. Conclusions

Data Exports from C&DM Tools

Project & Product Configuration Management Tools

- UGS TEAMCENTER: data on apparatus components and their configuration states, seen at the design level, exported in HTML format (*~2 MB export with ~10.000 records*)

Manufacturing Configuration Management Tools

- SAP R3: data partially overlapping with USG Teamcenter data, as well as data on components obsolescence, exported in Excel format *(~3 MB export with ~30.000 records)*

In-Service Configuration Management Tools

- SAP CS: data on physical components realized from design items, exported in Excel format *(~1 MB export with ~5.000 records)*
- Edea: XML data on the deployment of physical components, partially overlapping with SAP CS data *(~5 MB DB dump with ~5.000 nodes)*
- Odb: relational data on components obsolescence, possible substitutions, and requests of purchasing or producing new components *(~110 MB SQLServer DB Dump with ~50.000 tuples)*

Data Export: Example (From SAP R3)

Materiale			R603B					OF			Alt.		Imp.	1	
			ATCR 33S VERS.BASE										Valid	01/01/1900	
Qtà imp.			1,000		NR			Qtà base			1,000			NR	
Lv	Pos.	PN											Quant	UM CtP	TYPE
		Definizione													
1	0	05R107B									2,000			NR	L
		DIGITAL RECEIVER UNIT													
1	0	05R108B									1,000			NR	L
		RF/IF													
1	0	06R077									1,000			NR	L
		SOLID STATE TX 10S BASE													
1	0	Y.NT.1074				YNT	0					16,000		PZ	Y
		PASSIVAZIONE													
		YNTGENN000													
2	0	701370G1		SUPP.CONN.											
2	0	194692P1									32,000			NR	L
		PERNO													
		DOCENGG100													
3	0	19E004P112									0,160			KG	O
		MANCA DESCRIZIONE													
		OBSGENN000													
2	0	19E004P115									0,160			KG	O
		19E004P118													
		OBSGENN000													
2	0	19E004P118									0,160			KG	L

Federated Schema

- The data federation tool used in this case study is the IBM WebSphere Federation Server (FS), which provides support to wrap in relational format heterogeneous data, such as Excel, XML, HTML, textual data.
- All data sources to be integrated are represented in WebSphere FS by means of non-materialized relational views called nicknames.
- Each nickname is the output of a semi-automatic process of wrapping. Roughly:
  - A nickname is associated to each Excel sheet and to each HTML file.
  - A nickname is associated to each XML document representing data at its nodes, whereas other nicknames represent the father-child relation between document nodes.
  - A nickname is associated to each SQLServer relational table.
- Resulting federated schema: relational schema with 50 relations, each with around 15 attributes.

Federation: Example (TAB_1_SAP_R3 T)

```
+-----------------------+---------------+-------+-----------+-----------+-------------+
| Field                 | Type          | Null  | Key       | Default | Extra       |
+-----------------------+---------------+-------+------------+----------+------------+
| PN                    | varchar(50) | YES  | MUL       | NULL   |             |
| DEFINIZIONE           | varchar(50) | YES  |           | NULL    |             |
| VERSIONE              | varchar(50) | YES  |           | NULL    |             |
| QUANT                 | varchar(50) | YES  |           | NULL    |             |
| TIPOLOGIA             | varchar(50) | YES  |           | NULL    |             |
| UM_CTP                | varchar(50) | YES  |           | NULL    |             |
|.................................................................................... |
|.................................................................................... |
+----------------------+----------------+-------+-----------+----------+------------+
```

Outline of the Talk

1. Ontology-based Data Integration and the QuOnto System

2. Configuration and Data Management (C&DM) at SELEX Sistemi Integrati (SELEX-SI)

3. Experiencing QuOnto for C&DM at SELEX-SI
   i. Federating C&DM Data Sources
   ii. Designing the SELEX-SI C&DM Ontology
   iii. Relating Data Sources to the SELEX-SI C&DM Ontology through QuOnto
   iv. Querying the System

4. Conclusions

Ontology Design (Some Relevant Elements)

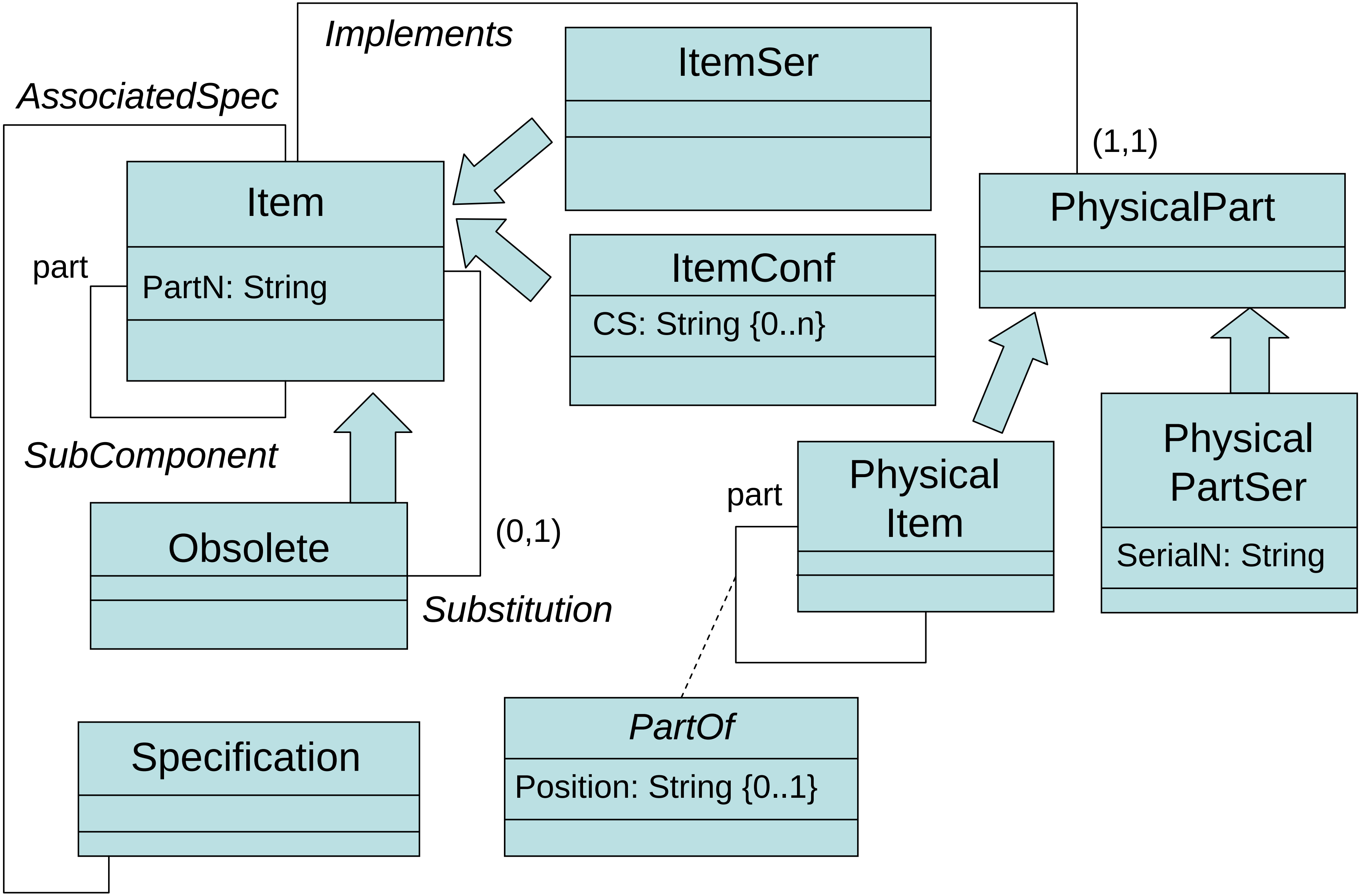
Concepts

- (virtual) Item: everything that is used in a project (e.g., a component). It can be configurable (can have several configuration states - CS) and serializable (each corresponding implementation has a serial number)
- Physical Part: implementation of a virtual item, possibly associated to a serial number (if it corresponds to a serializable item)
- Physical Item: physical part deployed in a larger component
- Specification: state of obsolescence of an item
- Obsolete: obsolete item (items that are no longer available). A substitution (role Substitution) for it can be specified

Roles

- SubComponent: relation between an item and the items that are its sub-components
- Implements: relation between an item and the physical parts that implement it
- PartOf: relation between a physical item and its parts (phys. items), possibly associated to the position that the part has in the physical item

Ontology Fragment (UML Approximation)



Ontology fragment DL-Lite Specification

$\mathit{ItemSer} \sqsubseteq \mathit{Item}$

$\mathit{ItemConf} \sqsubseteq \mathit{Item}$

$\mathit{Obsolete} \sqsubseteq \mathit{Item}$

$\mathit{Item} \sqsubseteq \delta(\mathit{PartN})$

$\mathrm{func}(\mathit{PartN})$

$\rho(\mathit{PartN}) \sqsubseteq \mathit{String}$

$\exists\, \mathit{Implements} \sqsubseteq \mathit{Item}$

$\exists\, \mathit{Implements}^{-} \sqsubseteq \mathit{PhysicalPart}$

$\exists\, \mathit{Substitution} \sqsubseteq \mathit{Item}$

$\exists\, \mathit{Substitution}^{-} \sqsubseteq \mathit{Obsolete}$

$\exists\, \mathit{SubComponent} \sqsubseteq \mathit{Item}$

$\exists\, \mathit{SubComponent}^{-} \sqsubseteq \mathit{Item}$

$\mathit{PhysicalPartSer} \sqsubseteq \mathit{PhysicalPart}$

$\mathit{PhysicalItem} \sqsubseteq \mathit{PhysicalPart}$

$\exists\, \mathit{partOf} \sqsubseteq \mathit{PhysicalItem}$

$\exists\, \mathit{partOf}^{-} \sqsubseteq \mathit{PhysicalItem}$

$\rho(\mathit{Position}) \sqsubseteq \mathit{String}$

$\mathrm{func}(\mathit{Position})$

$\mathit{PhysicalPartSer} \sqsubseteq \delta(\mathit{SerialN})$

$\rho(\mathit{SerialN}) \sqsubseteq \mathit{String}$

$\exists\, \mathit{AssociatedSpec} \sqsubseteq \mathit{Item}$

$\exists\, \mathit{AssociatedSpec}^{-} \sqsubseteq \mathit{Specification}$

$\mathrm{func}(\mathit{Implements}^{-})$

$\mathit{PhysycalPart} \sqsubseteq \exists\, \mathit{Implements}^{-}$

$\mathit{PhysycalPartSer} \sqsubseteq \exists\, \mathit{Implements1}$

$\exists\, \mathit{Implements1} \sqsubseteq \mathit{ItemSer}$

$\exists\, \mathit{Implements1} \sqsubseteq \neg\, \mathit{ItemConf}$

Outline of the Talk

1. Ontology-based Data Integration and the QuOnto System

2. Configuration and Data Management (C&DM) at SELEX Sistemi Integrati (SELEX-SI)

3. Experiencing QuOnto for C&DM at SELEX-SI
   i. Federating C&DM Data Sources
   ii. Designing the SELEX-SI C&DM Ontology
   iii. Relating Data Sources to the SELEX-SI C&DM Ontology through QuOnto
   iv. Querying the System

4. Conclusions

Mapping Assertions

- The instances of Item are objects constructed with the part numbers retrieved from TAB_1_SAP_R3 of the SAP_R3 system

```
Item(f(T.pn)) ←  SELECT T.pn
                 FROM TAB_1_SAP_R3  T
```

- The instances of Substitution are defined as follows

```
Substitution(f(T.pn), f(T.def)) ←  SELECT T.pn, T.def
                                   FROM TAB_1_SAP_R3 T
                                   WHERE T.tipologia = 'O'
                                         AND T.pn IN (
                                               SELECT T2.pn
                                               FROM TAB_1_SAP_R3 T2)
```

Outline of the Talk

1. Ontology-based Data Integration and the QuOnto System

2. Configuration and Data Management (C&DM) at SELEX Sistemi Integrati (SELEX-SI)

3. Experiencing QuOnto for C&DM at SELEX-SI
    - i. Federating C&DM Data Sources
    - ii. Designing the SELEX-SI C&DM Ontology
    - iii. Relating Data Sources to the SELEX-SI C&DM Ontology through QuOnto
    - iv. Querying the System

4. Conclusions

Query 1

Return the pairs of part numbers <p,p'> such that p is the part number of an obsolete item for which there exists a deployed implementation, and p' is the part number of a substitution of p

q(it, sub) :- PartN(X, it), Obsolete(X), Substitution(X, Y),
PartN(Y, sub), Implements(X,Z), PhysicalItem(Z)

Query 1 specified over the federated schema

```
SELECT T_SPEC.SPEC_ID as part_number, T_SPEC.SPEC_DEF as substitution,
FROM PAOLO."R603B_PROD_SAP_INFO$" AS "R603B_PROD_SAP_INFO$", PAOLO.T_SPEC AS
          T_SPEC
WHERE "R603B_PROD_SAP_INFO$".N_COMPONENTI = T_SPEC.SPEC_ID AND
          "R603B_PROD_SAP_INFO$".TIPOLOGIA = 'O'
UNION
SELECT T_SPEC.SPEC_ID as part_number, T_SPEC.SPEC_DEF as substitution,
FROM PAOLO.T_SPEC AS T_SPEC, PAOLO."X08R009_PROD_SAP_INFO$" AS
          "X08R009_PROD_SAP_INFO$"
WHERE T_SPEC.SPEC_ID = "X08R009_PROD_SAP_INFO$".N_COMPONENTI AND
          "X08R009_PROD_SAP_INFO$".O = 'O'
UNION
SELECT T_SPEC.SPEC_ID as part_number, T_SPEC.SPEC_DEF as substitution,
FROM PAOLO.T_SPEC AS T_SPEC, PAOLO."U08011971_PROD_SAP_INFO$" AS
          "U08011971_PROD_SAP_INFO$"
WHERE T_SPEC.SPEC_ID = "U08011971_PROD_SAP_INFO$".N_COMPONENTI AND
          "U08011971_PROD_SAP_INFO$".O = 'O'
UNION......... Complete Query
```

Query 2 (importance of reasoning)

Return all items

$$q(X) \text{ :- } Item(X)$$

- If we evaluate the query q without exploiting the reasoning capabilities of QuOnto, we get 577 objects in the answer.
- These are indeed the items directly mapped on the concept Item.
- The ontology specifies also that
  - Obsolete items are items ($Obsolete \sqsubseteq Item$)
  - Objects that are used as item substitutions are items ($\exists Substitution \sqsubseteq Item$)
  - ...
- Exploiting this knowledge (and the mappings specified on *Obsolete*, *Substitution*, *etc.)* through a *sound and complete* query answering algorithm, we get 1562 objects in the answer.

Conclusions

Our experience can be considered successful from different point of views, in particular:

- Access (i.e., query answering) to distributed and heterogeneous data has been centralized and automatized.
- Exploiting the conceptual representation of the domain of interest (i.e., the *DL-Lite* global schema), non-experts can now have both a more clear picture of the domain and access to data integration features.
- Exploiting reasoning capabilities of QuOnto, implicit knowledge automatically comes into play to produce complete answers to user queries.

Ongoing and future work

- Extending the C&DM ontology.
- Adding other C&DM data sources.
- Testing new QuOnto features: answering complex (i.e., EQL) queries, constraints, data update.

Thank You!

People involved in this work:

- *Alfonso Amoroso*[1]
- *Giuseppe De Giacomo*[2]
- *Gennaro Esposito*[1]
- *Domenico Lembo*[2]
- *Paolo Urbano*[2]
- *Raffaele Vertucci*[2]

1) SELEX Sistemi Integrati
2) Sapienza Univ. of Rome

Ontology Design

We proceeded both Bottom-up and Top-down:

- Bottom-up: we constructed an ontology for each data source and then fused such ontologies towards the design of the global one.
- Top-down: we iteratively refined the ontology according to specific user requirements.
  In particular, we considered the most relevant queries the user want to ask to the system, e.g.,
  - ✓ *Find out obsolete components that are installed and possible components substitutions.*
  - ✓ *For a given component find out the physical apparatus in which it is installed.*
  - ✓ *Find out the obsolescence status as it is indicated in OdB for the components that are obsolete according to SAP R3.*
  - ✓ *Compare configuration states of each component as they are indicated in each system (e.g., Teamcenter, Edea).*

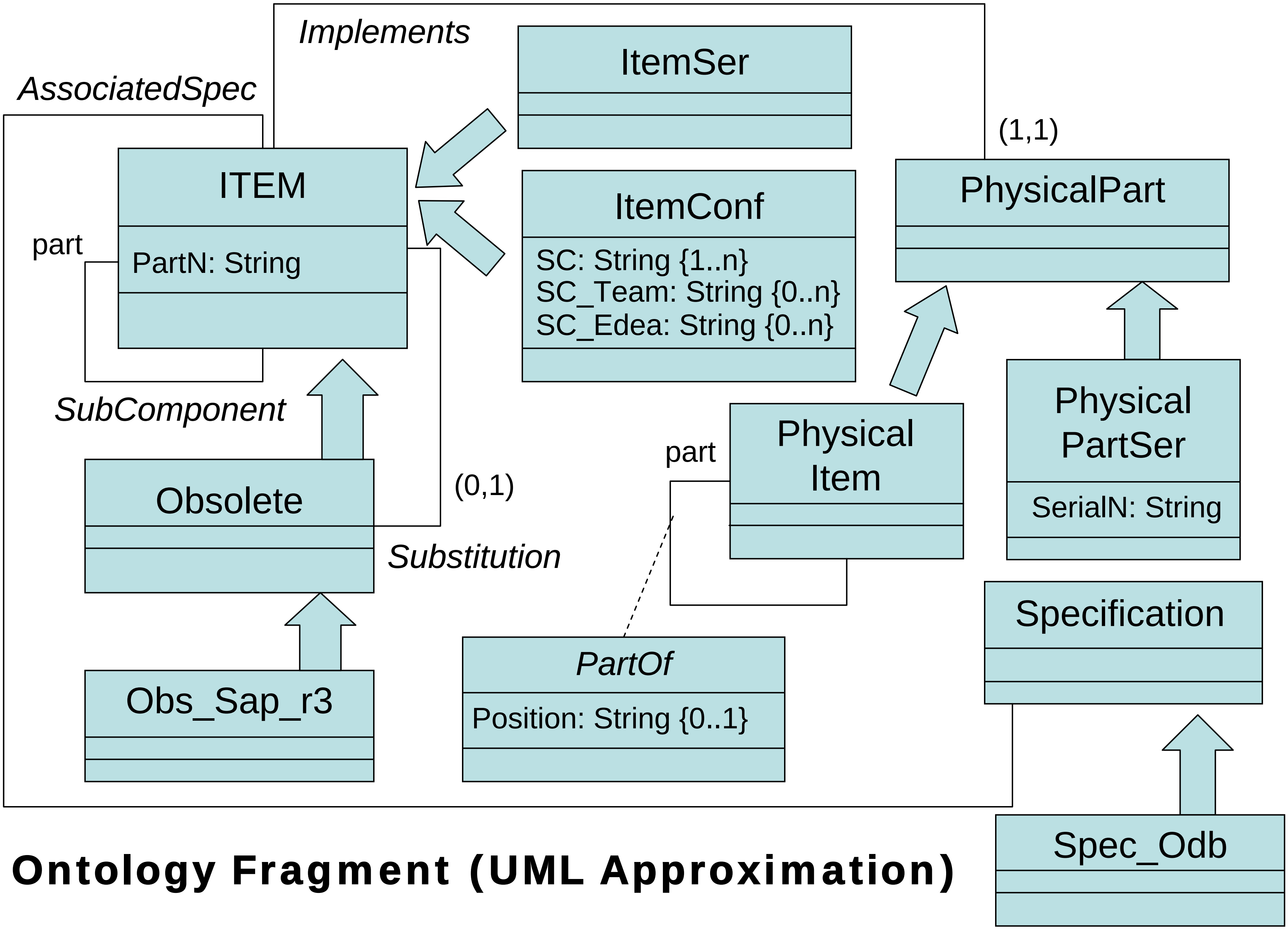
Ontology Design (continued)

- The queries to the ontology often need to refer to the C&DM tools:

  Return the obsolescence status as it is indicated in OdB for the components that are obsolete according to SAP R3.

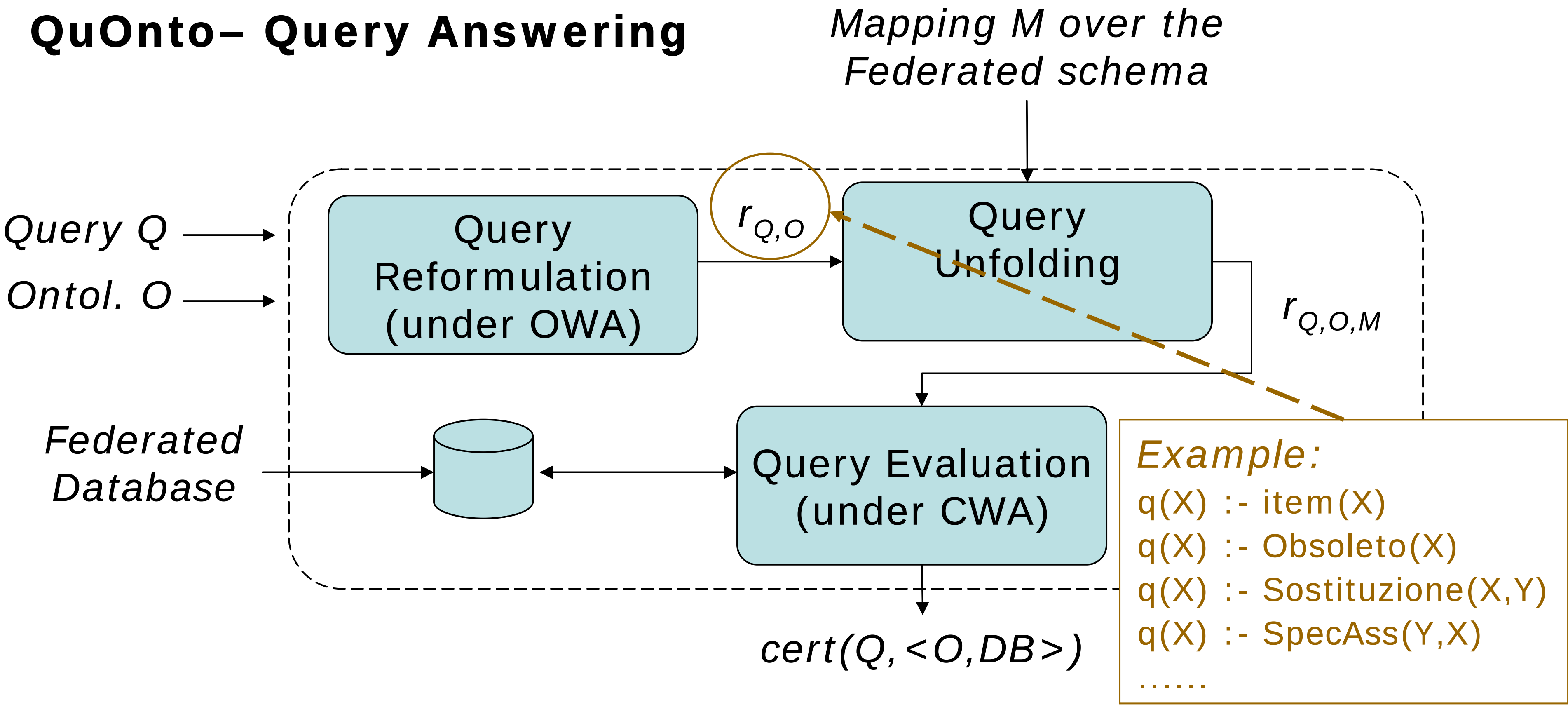
- Therefore, we introduced in the ontology some elements that explicitly refer to C&DM tools, e.g.,
  - Obs_sap_r3: items that are obsolete according to SAP R3,
  - Spec_Odb: state of obsolescence as specified in OdB,
  - SC_team, SC_Edea: attributes of the concept item that indicate the configuration state in TEAMCENTER and Edea.

  Obs_sap_r3 is a specialization of Obsolete. SC_team and SC_Edea are specializations of SC, and are attributes that represent the (general) configuration state.



Ontology Fragment (UML Approximation)

QuOnto– Query Answering


Mapping M over the Federated schema
Query Q
Ontol. O
Query Reformulation (under OWA)
$r_{Q,O}$
Query Unfolding
$r_{Q,O,M}$
Federated Database
Query Evaluation (under CWA)
cert(Q,<O,DB>)

Example:
q(X) :- item(X)
q(X) :- Obsoleto(X)
q(X) :- Sostituzione(X,Y)
q(X) :- SpecAss(Y,X)
……

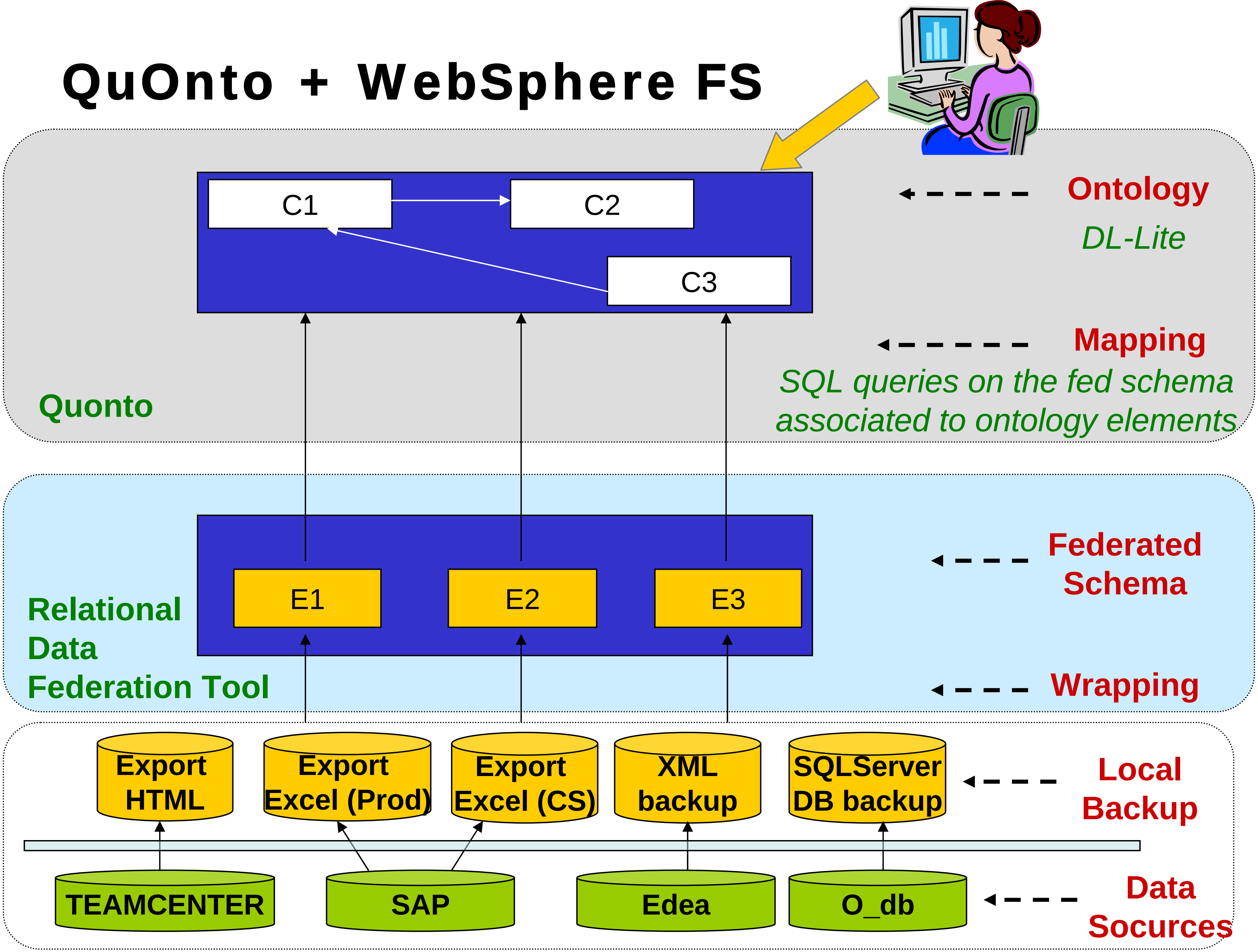
Query Answering is done in three phases

1. Query Reformulation: produces the query $r_{Q,O}$ that compiles the ontology *O* in the input query *Q*.
2. Query Unfolding: produces the query $r_{Q,O,M}$ over the federated schema.
3. Query evaluation: evaluates $r_{Q,O,M}$ over the federated database.

QuOnto & WebSphere FS

- QuOnto is developed in Java.
- It can access through JDBC any Relational Data Federation tool.
- In this case study we used the IBM WebSphere Federation Server (FS).
- WebSphere FS allows for semi-automatic design of wrappers to represent in relational format heterogeneous data, such as Excel, XML, HTML, textual data.
- Data Sources are thus seen by QuOnto as if they were a single relational database managed by WebSphere FS.

QuOnto + WebSphere FS



Quonto

C1

C2

C3

Ontology

DL-Lite

Mapping

SQL queries on the fed schema associated to ontology elements

Relational Data Federation Tool

E1

E2

E3

Federated Schema

Wrapping

Export HTML

Export Excel (Prod)

Export Excel (CS)

XML backup

SQLServer DB backup

Local Backup

TEAMCENTER

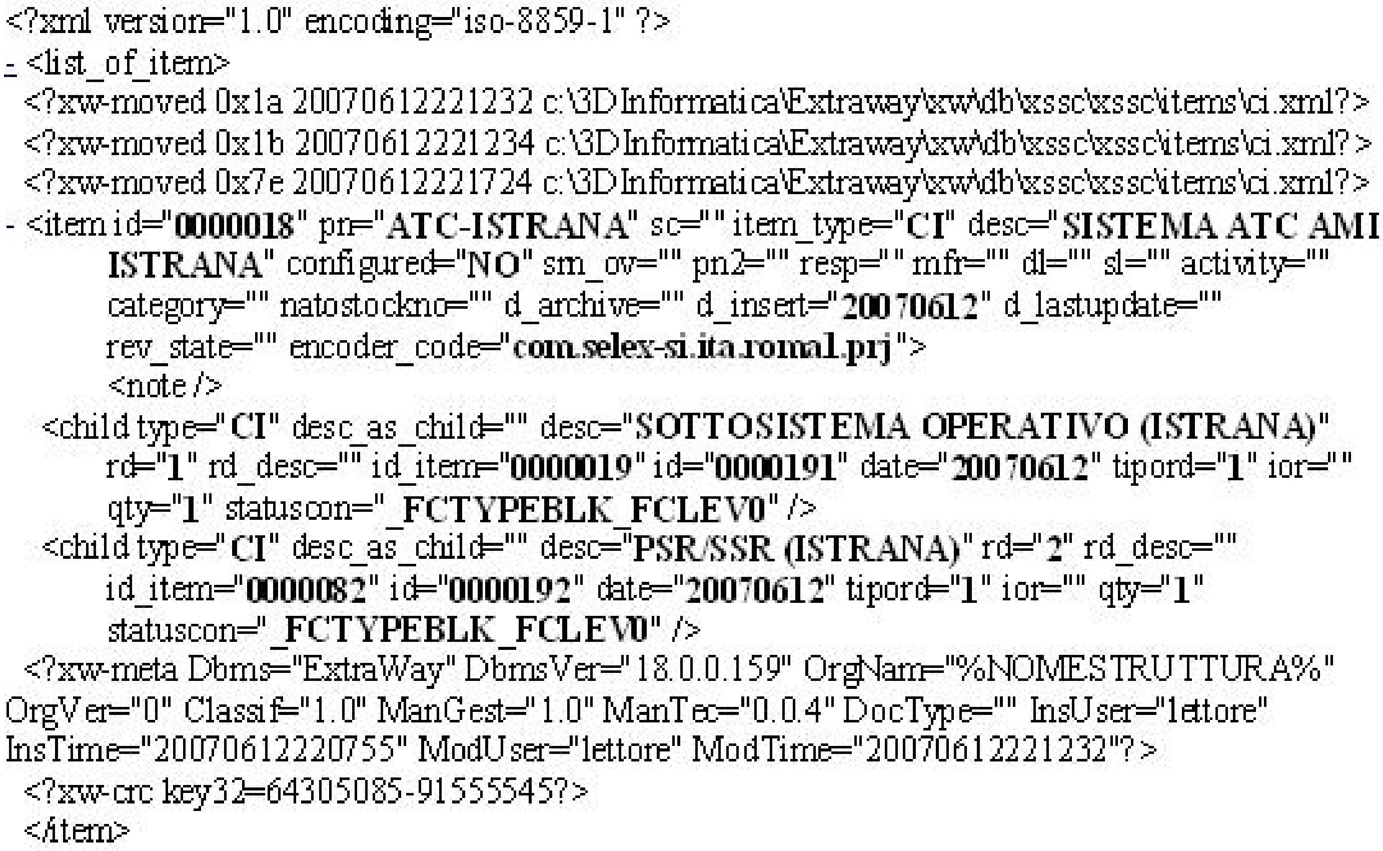
SAP

Edea

O_db

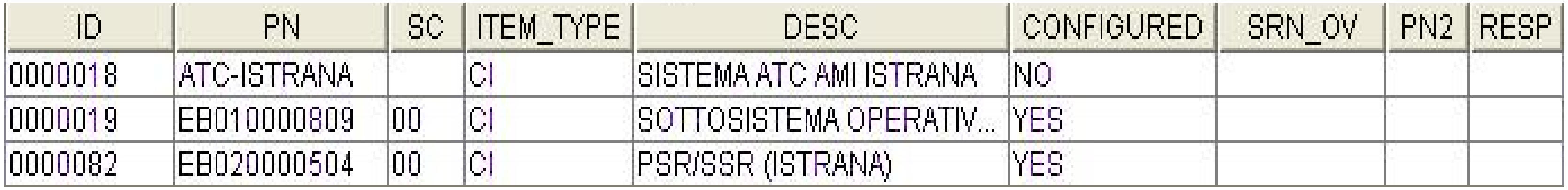
Data Socurces

Example: wrapping Edea



```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <list_of_item>
  <?xw-moved 0x1a 20070612221232 c:\3D Informatica\Extraway\xw\db\xssc\xssc\items\ci.xml?>
  <?xw-moved 0x1b 20070612221234 c:\3D Informatica\Extraway\xw\db\xssc\xssc\items\ci.xml?>
  <?xw-moved 0x7e 20070612221724 c:\3D Informatica\Extraway\xw\db\xssc\xssc\items\ci.xml?>
- <item id="0000018" pn="ATC-ISTRANA" sc="" item_type="CI" desc="SISTEMA ATC AMI
        ISTRANA" configured="NO" srn_ov="" pn2="" resp="" mfr="" dl="" sl="" activity=""
        category="" natostockno="" d_archive="" d_insert="20070612" d_lastupdate=""
        rev_state="" encoder_code="com.selex-si.ita.roma1.prj">
        <note />
   <child type="CI" desc_as_child="" desc="SOTTOSISTEMA OPERATIVO (ISTRANA)"
        rd="1" rd_desc="" id_item="0000019" id="0000191" date="20070612" tipord="1" ior=""
        qty="1" statuscon="_FCTYPEBLK_FCLEV0" />
   <child type="CI" desc_as_child="" desc="PSR/SSR (ISTRANA)" rd="2" rd_desc=""
        id_item="0000082" id="0000192" date="20070612" tipord="1" ior="" qty="1"
        statuscon="_FCTYPEBLK_FCLEV0" />
  <?xw-meta Dbms="ExtraWay" DbmsVer="18.0.0.159" OrgNam="%NOMESTRUTTURA%"
OrgVer="0" Classif="1.0" ManGest="1.0" ManTec="0.0.4" DocType="" InsUser="lettore"
InsTime="20070612220755" ModUser="lettore" ModTime="20070612221232"?>
  <?xw-crc key32=64305085-91555545?>
  </item>
```

Result of the wrapping



ID	PN	SC	ITEM_TYPE	DESC	CONFIGURED	SRN_OV	PN2	RESP
0000018	ATC-ISTRANA		CI	SISTEMA ATC AMI ISTRANA	NO			
0000019	EB010000809	00	CI	SOTTOSISTEMA OPERATIV...	YES			
0000082	EB020000504	00	CI	PSR/SSR (ISTRANA)	YES			

Query 2

Return the obsolescence status (date, state, classification, action) as it is indicated in OdB for the components that are obsolete according to SAP R3

q(pn,dt,st,cl,act) :- Obs_sap_r3(X), Part_N(X,pn), specAssociata(X,Y), Data(Y,dt), Stato(Y,st), Classif(Y,cl), azione(Y,act)