



Gestire i dati attraverso ontologie

Maurizio Lenzerini

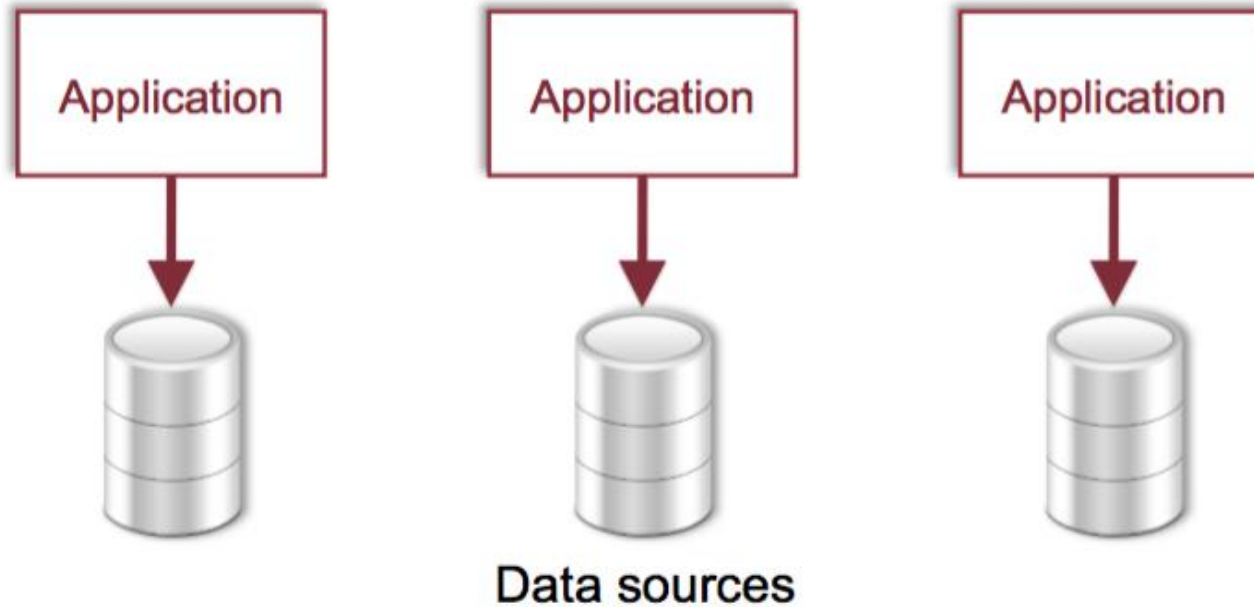
<http://www.diag.uniroma1.it/~lenzerini>

DASI-lab, Data and Service Integration Laboratory
Dipartimento di Ingegneria Informatica
Automatica e Gestionale “Antonio Ruberti”
Sapienza Università di Roma

OBDA Systems s.r.l.

Ingegneria R&D, Roma, 11 maggio 2018

Tipica architettura di un attuale sistema informativo complesso



- Distributed, redundant, application-dependent, and mutually incoherent data
- Desperate need of a coherent, conceptual, unified view of data



Esempio: frammento di una tabella relazionale in una banca

Valore negativo indica una holding

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT	
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N	
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N	
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S	
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N	
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S	
-472900	10-mag-2001	1-gen-9999	62010	S	N	0 00	N	
130976	7-mag-2001	9-lug-2003	75680					



Esempio: frammento di una tabella relazionale in una banca

S → cliente è leader del gruppo

S → cliente è capo del gruppo

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT	
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N	
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N	
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S	
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N	
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S	
-472900	10-mag-2001	1-gen-9999	62010	S	N	0 00	N	
130976	7-mag-2001	9-lug-2003	75680					



Esempio: frammento di una tabella relazionale in una banca

N → fatturato non significativo

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT	
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N	
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N	
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S	
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N	
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S	
-472900	10-mag-2001	1-gen-9999	62010	S	N	0 00	N	
130976	7-mag-2001	9-lug-2003	75680					



Soluzione possibile: gestire i dati attraverso una ontologia di dominio

DASI-lab, Data and Service Integration Laboratory
Dipartimento di Ingegneria Informatica Automatica e
Gestionale “Antonio Ruberti”, Sapienza Università di Roma



www.obdasystems.com

Startup di



SAPIENZA
UNIVERSITÀ DI ROMA

due attori

La squadra



Prof. Maurizio Lenzerini
Co-Fondatore & Presidente



Valerio Santarelli
Co-Fondatore & CEO



Marco Ruzzi
Co-Fondatore & CTO

Prof. Riccardo Rosati
Co-Fondatore & Consulente Scientifico

Prof. Giuseppe De Giacomo
Co-Fondatore & Consulente Scientifico

Prof. Domenico Lembo
Co-Fondatore & Project Manager

Dr. Antonella Poggi
Co-Fondatore & Project Manager

Antonella Poggi
Co-Fondatore & Project Manager

Domenico Fabio Savo
Co-Fondatore & Project Manager

Lorenzo Lepore
Co-Fondatore & R&D

Federico Croce
R&D - Mastro

Giacomo Ronconi
R&D - Mastro

Federico Scafoglieri
R&D – Mastro Studio

Alessandro Bartolucci
Co-Fondatore di Studiare SRL

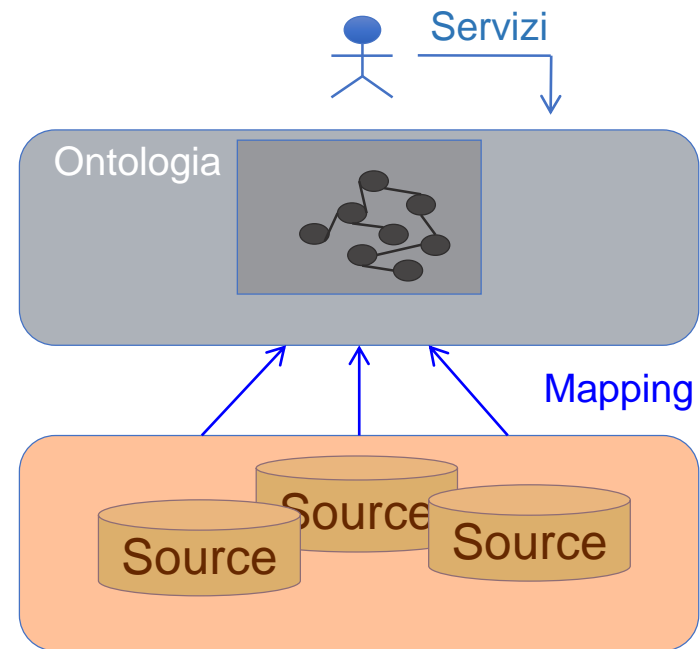
Paolo Naggari
Co-Fondatore di Studiare SRL

(Docenti Sapienza
ideatori dell'OBDM)

Ontology-based data management

L'ontology-based data management (OBDM) è un paradigma per la **gestione** dei dati di una organizzazione basato sulla semantica

- **Ontologia**: descrizione formale e condivisa del dominio di interesse
- **Mapping**: relazione tra i dati nelle sorgenti ed i concetti nell'ontologia
- **Sorgenti dati**: repository di dati gestiti dai sistemi esistenti
- **Servizi**: l'utente finale interagisce con l'ontologia al fine di usufruire di servizi che si basano su una vista unificata e concettuale del patrimonio informativo



Ontology-based data management

- Il paradigma è stato introdotto dal DASI-LAB della Sapienza, le cui ricerche hanno definito
 - Metodologia
 - Suite di strumenti
- Le ricerche hanno condotto alla formazione di una start-up Sapienza: **ODBA Systems s.r.l.** che ha ingegnerizzato i tools

I pilastri di OBDM

1. **Condivisione** della conoscenza del dominio
2. **Integrazione** dei dati (superamento dell'architettura "a silos")
3. **Servizi** di querying e data analytics sul patrimonio informativo dell'organizzazione
4. **Qualità** dei dati e data governance
5. **Semantic** open data preparation and publishing

Altre caratteristiche:

- ristrutturazione delle sorgenti
- progettazione di nuove sorgenti
- modellazione semantica di servizi e processi
-

Condivisione della conoscenza

- L'ontologia è un **asset fondamentale** dell'organizzazione
- Padroneggiare concetti, relazioni e regole è anche importante rispetto allo scopo di **collaborare** e **comunicare** con altre organizzazioni
- Il livello concettuale in cui è espressa l'ontologia **libera** dalla “schiavitù” delle applicazioni

Integrazione dei dati

- L'ontologia è il mezzo per **integrare** i dati dispersi in diverse sorgenti, interni o esterni all'organizzazione
- Il sistema complessivo offre una visione concettuale e unificata del **patrimonio informativo** dell'organizzazione
- L'integrazione favorita da OBDM non è **procedurale** (task-based), ma **dichiarativa** ("general purpose") e può seguire un approccio "**pay-as-you-go**"

Querying e data analytics

- Le esigenze informative si esprimono in termini degli elementi dell'ontologia
- È il sistema che si preoccupa di **tradurre** la richiesta in processi che accedono alle sorgenti
- La complessità complessiva del query answering è influenzata dalla “distanza” tra la struttura delle sorgenti e l'ontologia.
- OBDM è una risposta al problema della “**data preparation**” in data analytics e data mining (data scientists spend at least 60% of their time in data organization and preparation – CrowdFlower 2017)

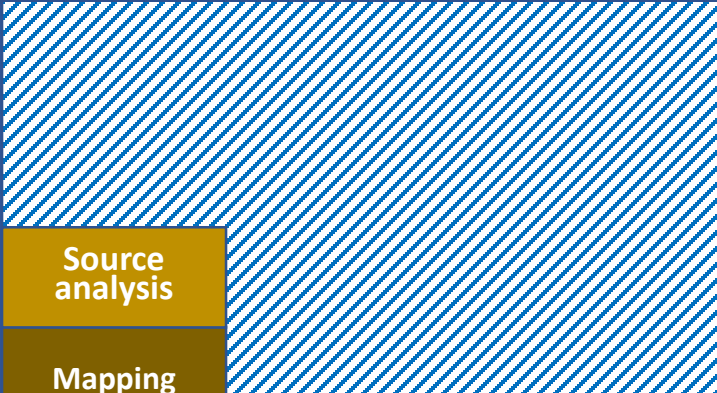

Qualità dei dati e data governance

- L'ontologia diventa il mezzo per **valutare** la qualità dei dati, secondo le classiche dimensioni (correttezza, completezza, ...)
- La qualità dei dati si **misura** rispetto alle regole di business dell'organizzazione, ed una volta disegnati l'ontologia e i mapping, ogni **discrepanza** tra questi e i dati testimonia problemi di qualità
- Oltre alla valutazione della qualità, OBDM promuove nuove modalità di **data governance** (source profiling, gestione di metadati azioni di miglioramento della qualità, ecc.)

Semantic open data preparation and publishing

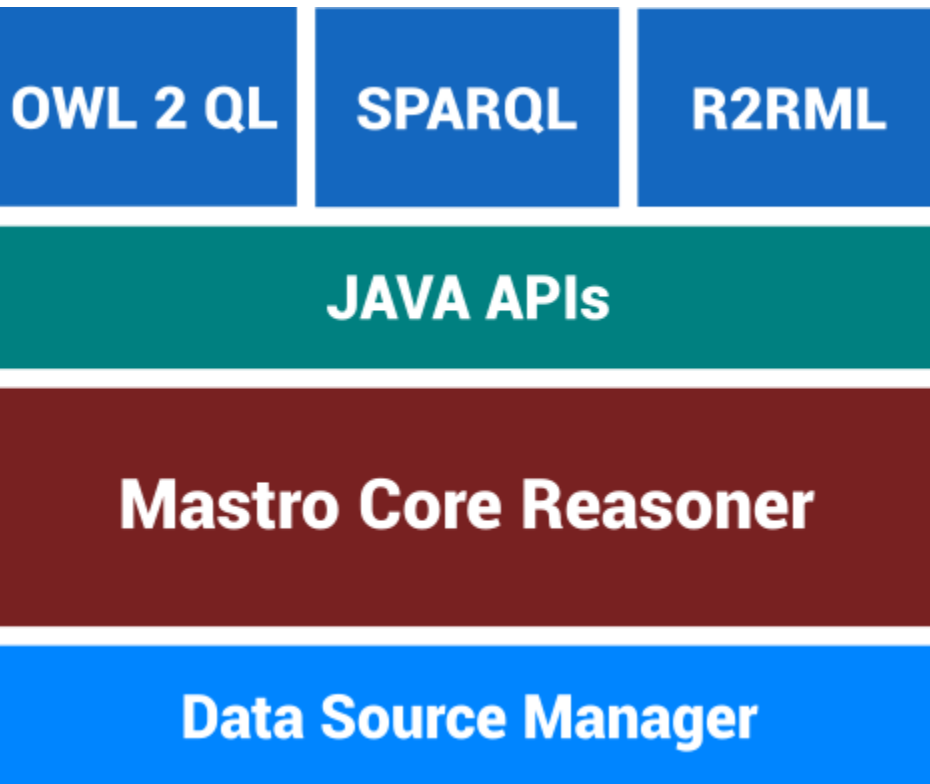
- La produzione dei datasets si realizza mediante il servizio di query answering
- L'ontologia diventa il mezzo per annotare i datasets ed i valori contenuti nei datasets
- Il ragionamento sull'ontologia diventa parte del processo di risposta alle query nello SPARQL endpoint
- La rappresentazione della provenance dei dati viene semplificata dall'utilizzo dei mapping

La suite degli strumenti

Languages	Tools	Phases				Methods	Actors
OWL	EDDY	Ontology development				Ontology design methodology	Domain experts
DL-Lite	PROTÉGÉ						KR experts
Graphol							
OWL	MASTRO	Ontology and mapping refinement	Source analysis			Source profiling	Data source experts
R2RML	PROTÉGÉ (MASTRO plug-in)		Mapping development			Mapping design methodology	Domain experts
			Competency questions	Documentation preparation	Data quality assessment and data governance	KR experts	
SPARQL	MASTRO-STUDIO		Usage (queries, quality check, publishing)	Documentation analysis	Open data publishing	Final users	



mastro



Supporto agli standard

Ontologia: OWL 2 (QL)

Mapping: R2RML (+propr.)

Query: SPARQL

DBMS supportati

Oracle, MySQL, Postgres,
SQL Server

Java API proprietarie

Mastro Web Services



eddy

Distribuzione open-source sotto licenza GPL v3

Supporto per Windows, macOS, Linux

Esportazione di ontologie in format OWL 2 (support standard)



protégé plug-in

Protégé è un editor Java-based per ontologie OWL 2

Distribuzione open-source sotto licenza GPL v3

Supporto per Windows, macOS, Linux



mastrostudio

Suite di moduli per DKAN (Drupal-based)

Servizi di Mastro offerti tramite Web Services

Supporto per application server Wildfly, Tomcat



Progetti

- Monte dei Paschi Siena
- Ministero Economia e Finanza
- Telecom
- Statoil
- Bloomberg
- ISTAT
- ACI
-

Attuali sviluppi

- Modellazione ontologica di aggregati (coerente con il “Data Cube vocabulary”)
- Estensione del profilo di SPARQL supportato, in particolare al fine di esprimere funzioni aggregate
- Potenziamento del linguaggio di espressione di vincoli per “data quality assessment”
- Supporto al versioning di ontologia e mapping
- Metodologia e supporto al disegno e ristrutturazione di sorgenti a partire da ontologia