UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

Dipartimento di Ingegneria – “Enzo Ferrari”

Corso di Laurea Magistrale in Ingegneria Informatica

MOMIS e Open Data:
Integrazione di Dati Aziendali con Sorgenti Dati Pubblici

Relatore:
Chiar.ma Prof. Sonia Bergamaschi

Correlatore:
Dott. Mirko Orsini

Elaborato di:
Giovanni Esposito

Anno Accademico 2012/2013

Agenda

- Presentazione
- OpenData
- Scopo del progetto
- Analisi delle sorgenti
- Analisi degli strumenti per estrazione/integrazione dati
- Soluzione adottata
- Conclusioni

Come nasce il progetto…

- ***DataRiver s.r.l** è una spin-off dell'università di Modena e Reggio Emilia, fondata nel giugno 2009 per distribuire la versione Open Source del sistema di Data Integration MOMIS, sviluppato dal gruppo di ricerca DBGroup. DataRiver offre soluzioni nei campi della Data Integration, Information Management, Business Intelligence e Clinical Data Management.*

- *Lo stage è durato 6 mesi*

- *Il progetto è stato realizzato all'interno dell'azienda*

DataRiver
open source data integration

Presentazione del progetto

Il progetto prevede:

1) Integrazione di *Open Data*: dati provenienti da sorgenti pubbliche con dati presenti nei sistemi informativi aziendali (es. ERP, CRM).

2) Utilizzo di *MOMIS* per l'Integrazione dei Dati.

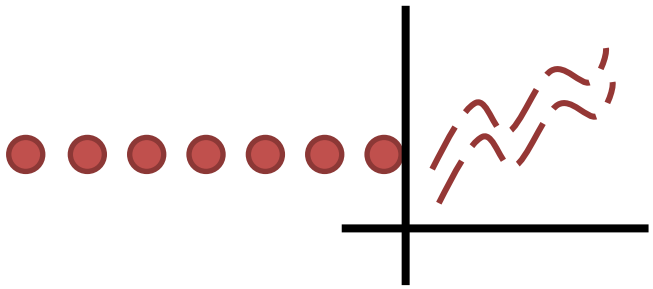
3) Analisi dell'integrazione con la componente di *MOMIS*: *MOMIS dashboard*.

Dati Aziendali

Open Data

INTEGRAZIONE

ANALISI INTEGRAZIONE



Cosa sono gli Open Data?

- L' Open Data è un movimento internazionale ed un insieme di pratiche, secondo cui, l'accesso ai **dati pubblici** deve essere libero e gratuito per tutti, senza restrizioni e brevetti che ne limitino la riproduzione e l'accesso.

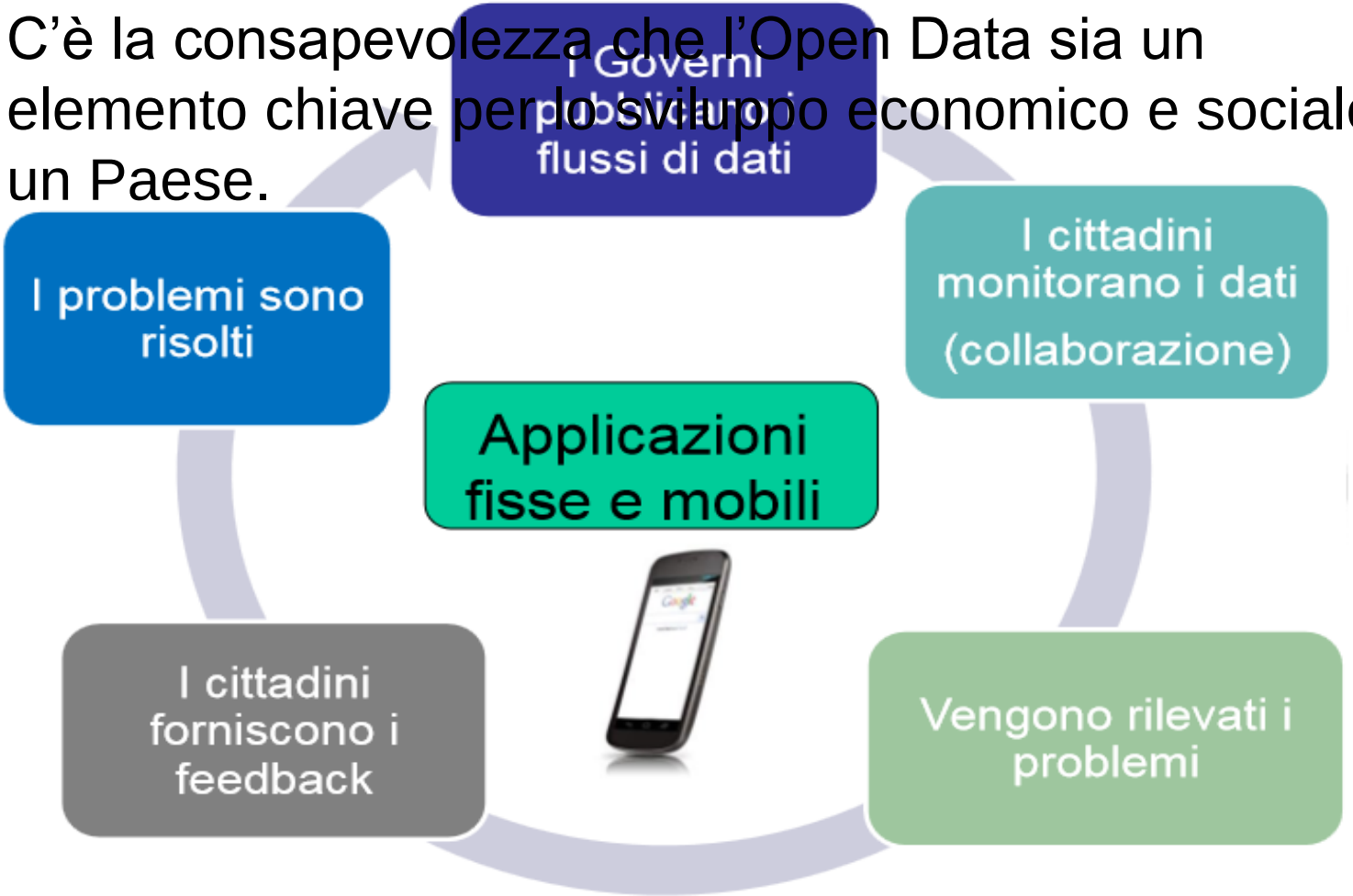
- L‘Open Data si richiama alla più ampia disciplina **dell’open government**, cioè un modello di amministrazione aperto e trasparente da parte degli enti e delle istituzioni pubbliche verso i cittadini.

- La diffusione dei dati consente di creare una conoscenza condivisa e un servizio aggiuntivo.

- Il problema è l’estrema disomogeneità dei formati: es. xml, excel, csv, ecc.

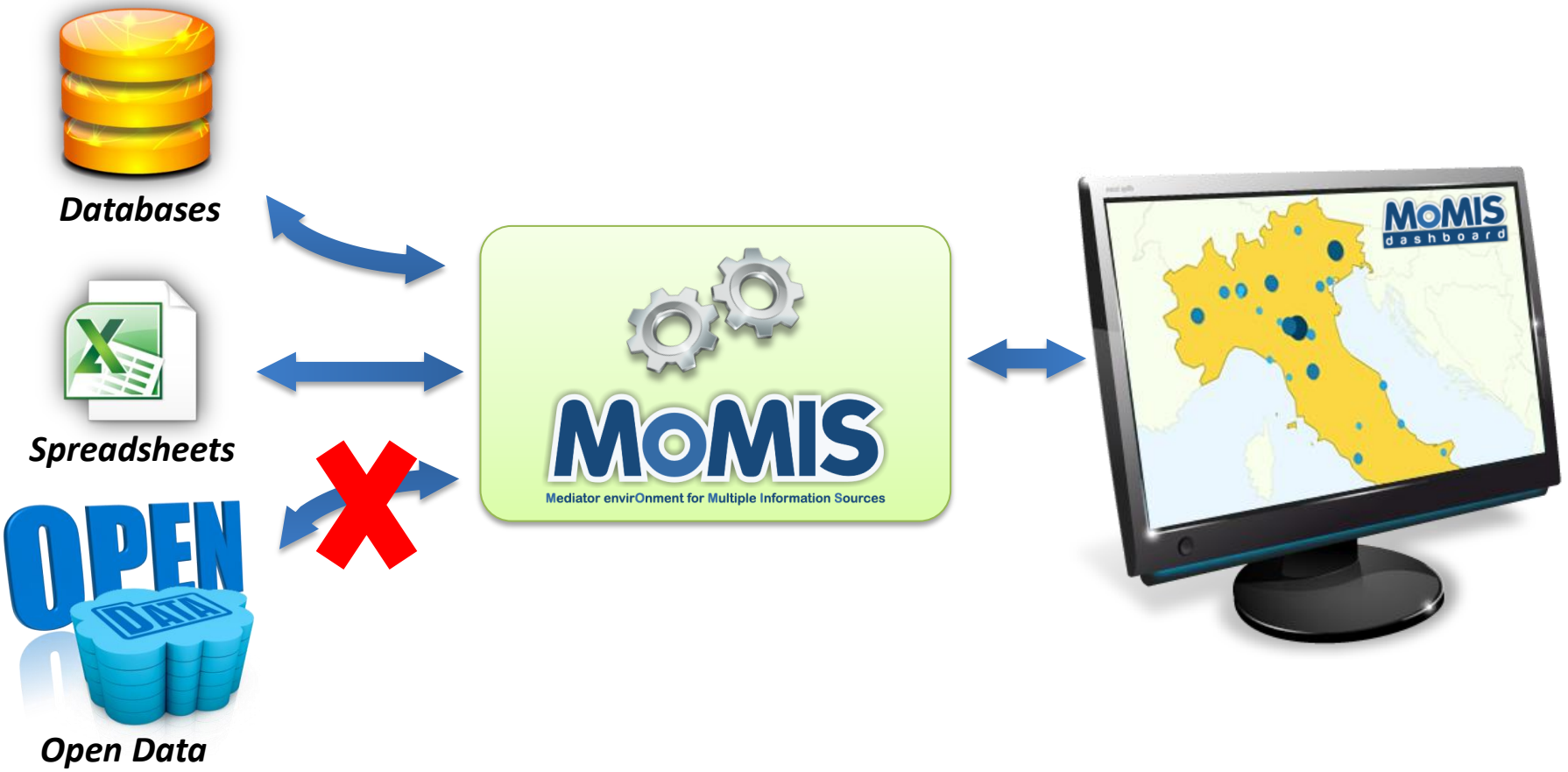
Un esempio di miglioramento della PA

C’è la consapevolezza che l’Open Data sia un elemento chiave per lo sviluppo economico e sociale di un Paese.

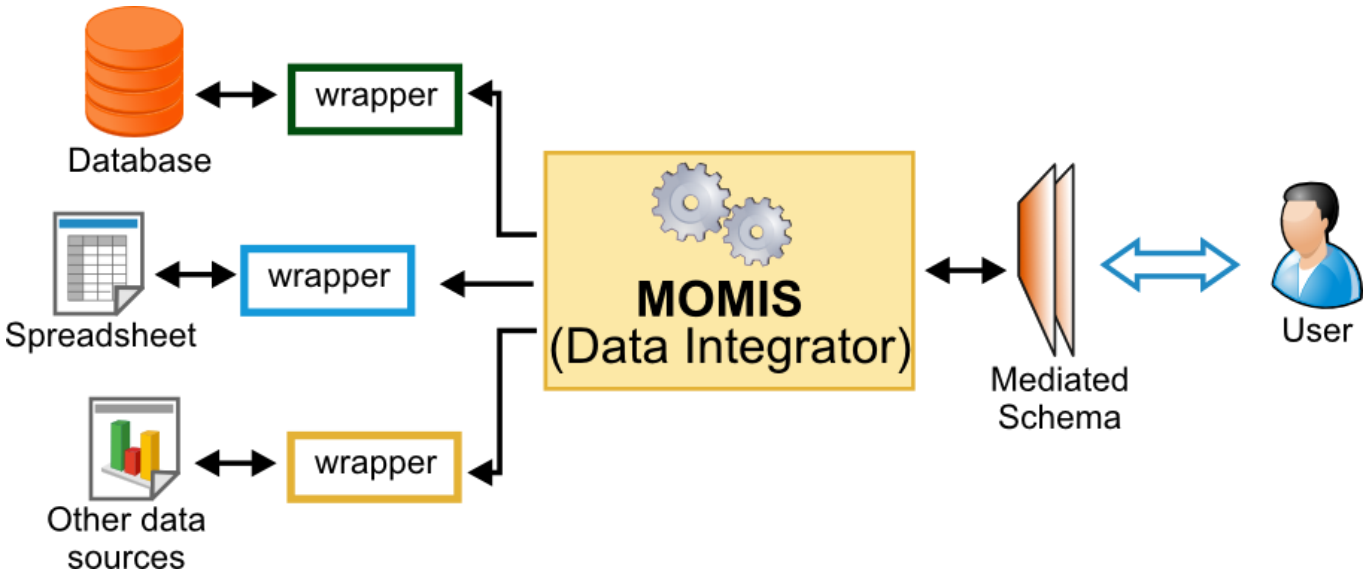


Il diluvio di dati

Per questo ho pensato di usare *MOMIS*, che possiede connettori per elaborare sistemi relazionali, spreadsheet e non per l'eterogeneità degli *Open Data*

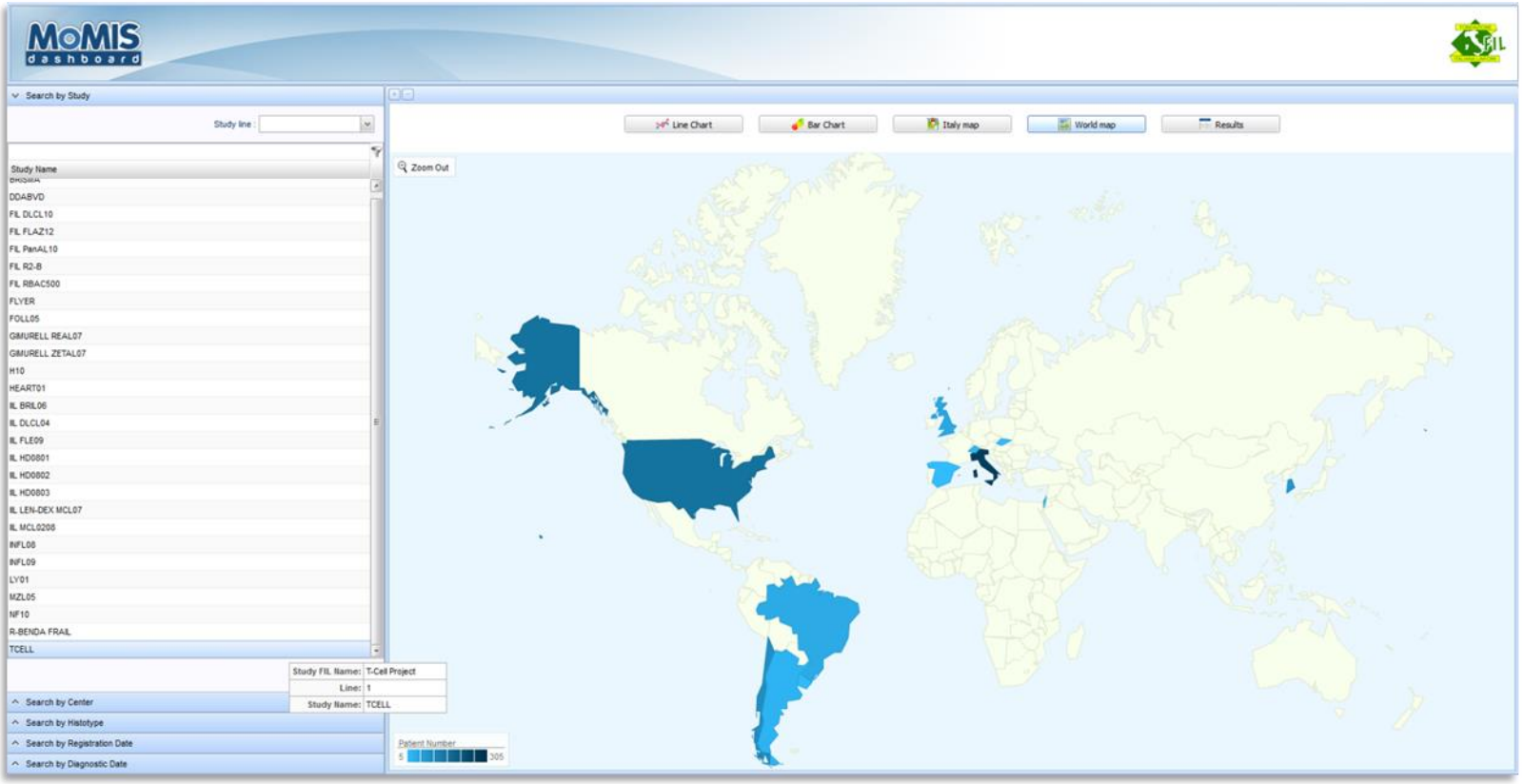


Perché MOMIS?

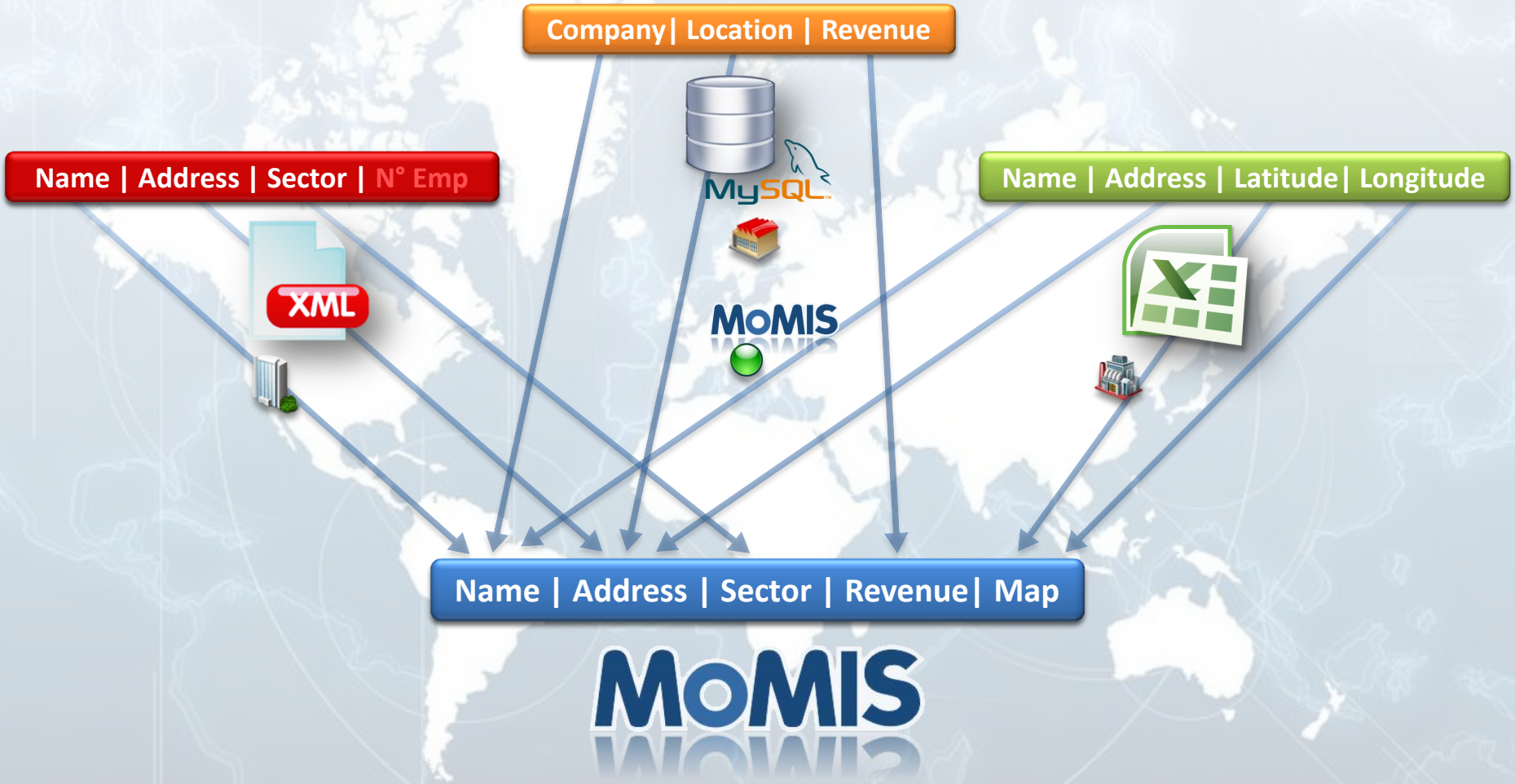


Uno ***Schema Mediato*** fornisce una vista virtuale ed integrata delle sorgenti dati locali coinvolte nell'integrazione. Non viene creata un copia centralizzata dei dati contenuti nelle sorgenti dati, la query posta dall'utente sullo schema mediato viene trasformata in un insieme di query sulle sorgenti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

MOMIS Dashboard è un'applicazione web che consente di visualizzare in formato grafico, tabellare o su mappa i dati eterogenei e distribuiti integrati in modo virtuale attraverso il sistema open source **MOMIS**.

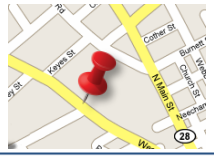


INTEGRATING DATA SOURCES



DATA FUSION

MOMIS

Name	Address	Sector	Revenue	Map
Software Inc.	Nimitz Fwy, Newark, US	Information Technology	€ 6.000 mln	
Fashion Inc.	Via Savona, Cuneo, IT	Textile	€ 930 mln	

VIRTUAL INTEGRATION

Data stored in Local sources

XML

Name	Address	Sector	N° Emp.
Fashion Inc.	Via Savona, Cuneo, IT	Textile	8000
Software Inc.	Nimitz Fwy, Newark, US	Information Technology	600

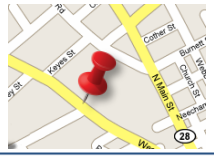
Company	Location	Revenue
Software Inc.	Nimitz Fwy, Newark, US	€ 6.000 mln
Fashion Inc.	Via Libertà, Cuneo, IT	€ 930 mln

Name	Address	Latitude	Longitude
Software Inc.	Nimitz Fwy, Newark, US	37'44 N	122'13 W

DATA CONFLICT

ALWAYS UP TO DATE

MOMIS

Name	Address	Sector	Revenue	Map
Software Inc.	Nimitz Fwy, Newark, US	Information Technology	€ 6.000 mln	
Fashion Inc.	Via Savona, Cuneo, IT	Textile	€ 1.200 mln	

VIRTUAL INTEGRATION

Data stored in Local sources

XML

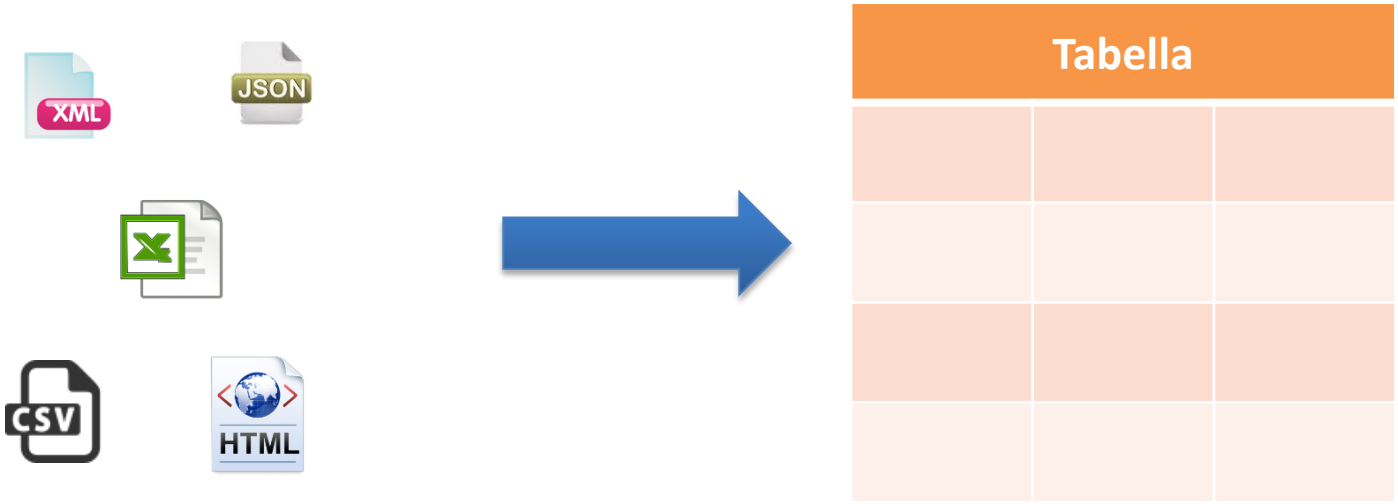
Name	Address	Sector	N° Emp.
Fashion Inc.	Via Savona, Cuneo, IT	Textile	8000
Software Inc.	Nimitz Fwy, Newark, US	Information Technology	600

Company	Location	Revenue
Software Inc.	Nimitz Fwy, Newark, US	€ 6.000 mln
Fashion Inc.	Via Libertà, Cuneo, IT	€ 1.200 mln

Name	Address	Latitude	Longitude
Software Inc.	Nimitz Fwy, Newark, US	37'44 N	122'13 W

Estrazione Open Data

I dati sono disponibili/forniti in formati eterogenei:



L'obiettivo è quello di trasformare questi formati eterogenei in un formato comune (modello relazionale). In modo da poterne fare l'analisi con opportuni strumenti statistici.

Sorgente	Descrizione	Interfaccia Sorgente	Formato	Frequenza Aggiornamento
THE WORLD BANK Working for a World Free of Poverty	Indici sullo sviluppo mondiale	WebService RESTFul	XML JSON	Year/Month
OpenWeatherMapAPI	Condizioni metereologiche	WebService RESTFul	XML JSON	Daily
YAHOO!	Andamento titoli borsa	WebService RESTFul	CSV	Daily
FOREX INTERNATIONAL TRADING GROUP INC	lista dei cambi attuali	WebService RESTFul	CSV	Daily
OpenSpending	Italian Spending per Region and Function	WebService RESTFul	XML JSON	Year
Formula 1	Risultati della formula 1	WebService RESTFul / SPARQL	XML	Weekly

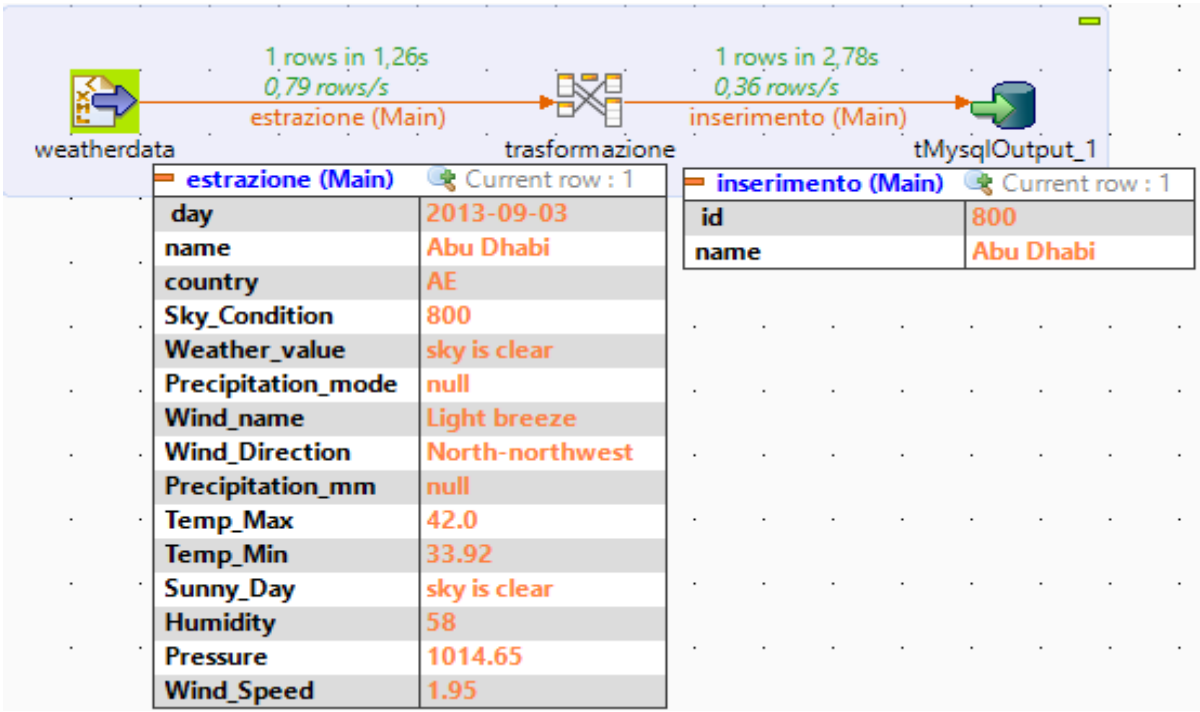
Esempio di risposta dalla sorgente World Bank

```
− <wb:data page="1" pages="1" per_page="50" total="36">
  − <wb:data>
      <wb:indicator id="DPANUSIFS">Exchange rate (IFS), LCU per USD, period average</wb:indicator>
      <wb:country id="BR">Brazil</wb:country>
      <wb:date>2009M10</wb:date>
      <wb:value/>
      <wb:decimal>0</wb:decimal>
    </wb:data>
  − <wb:data>
      <wb:indicator id="DPANUSIFS">Exchange rate (IFS), LCU per USD, period average</wb:indicator>
      <wb:country id="BR">Brazil</wb:country>
      <wb:date>2009M09</wb:date>
      <wb:value/>
      <wb:decimal>0</wb:decimal>
    </wb:data>
  − <wb:data>
      <wb:indicator id="DPANUSIFS">Exchange rate (IFS), LCU per USD, period average</wb:indicator>
      <wb:country id="BR">Brazil</wb:country>
      <wb:date>2009M08</wb:date>
      <wb:value/>
      <wb:decimal>0</wb:decimal>
    </wb:data>
  − <wb:data>
      <wb:indicator id="DPANUSIFS">Exchange rate (IFS), LCU per USD, period average</wb:indicator>
```

Strumenti ETL Open Source analizzati:

- Talend Open Studio
- Pentaho (Kettle)

Esempio di un processo di estrazione dalla sorgente OpenWeather:



Confronto di strumenti ETL Open Source per estrazione/integrazione dati*

	talend*	pentaho (Kettle)
User frendly	+ + +	+
Connettori per le sorgenti	+ + +	+ +
Conoscenze necessarie per l'utilizzo	+ +	+
Funzionalità per ETL (Estrazione e trasformazione)	+ + +	+
Facilità nello sviluppo di nuovi connettori	+ + +	+
Documentazione	+ + +	+ +
Forum di Supporto	+ + +	+

***NOTA**: Talend e Kettle non fanno integrazione dati ma fanno solo il mapping mentre MOMIS fa data integration

Accoppiamento di uno strumento ETL in MOMIS

Fasi di integrazione:

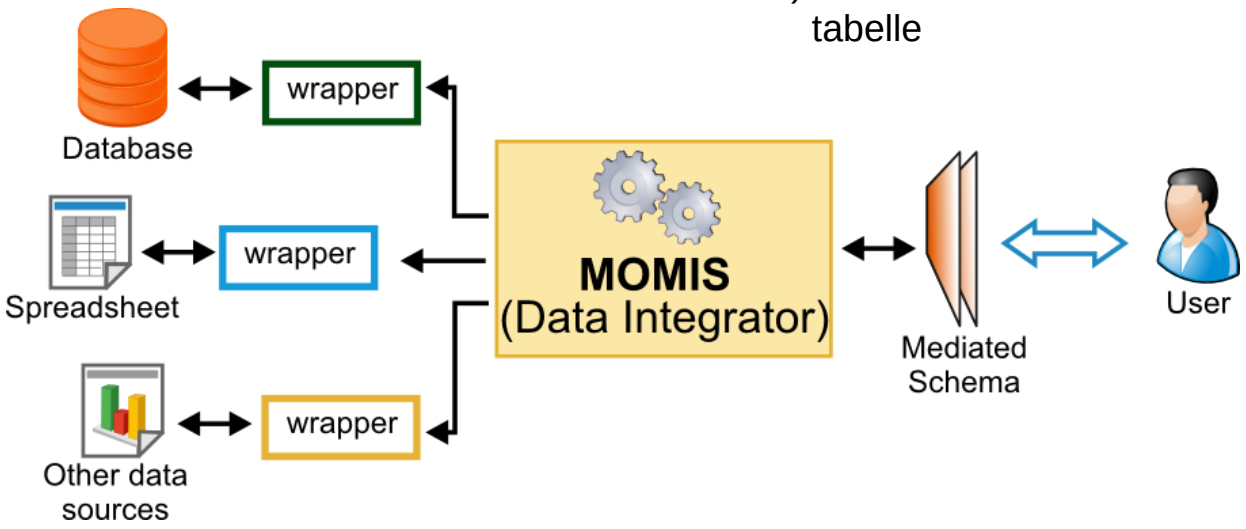
1) Creazione del processo di estrazione con *Talend*

2) Viene prodotto un .jar, che viene integrato nell'ambiente di sviluppo di *MOMIS*

3) Il .jar viene lanciato da *MOMIS* quando interroga i dati

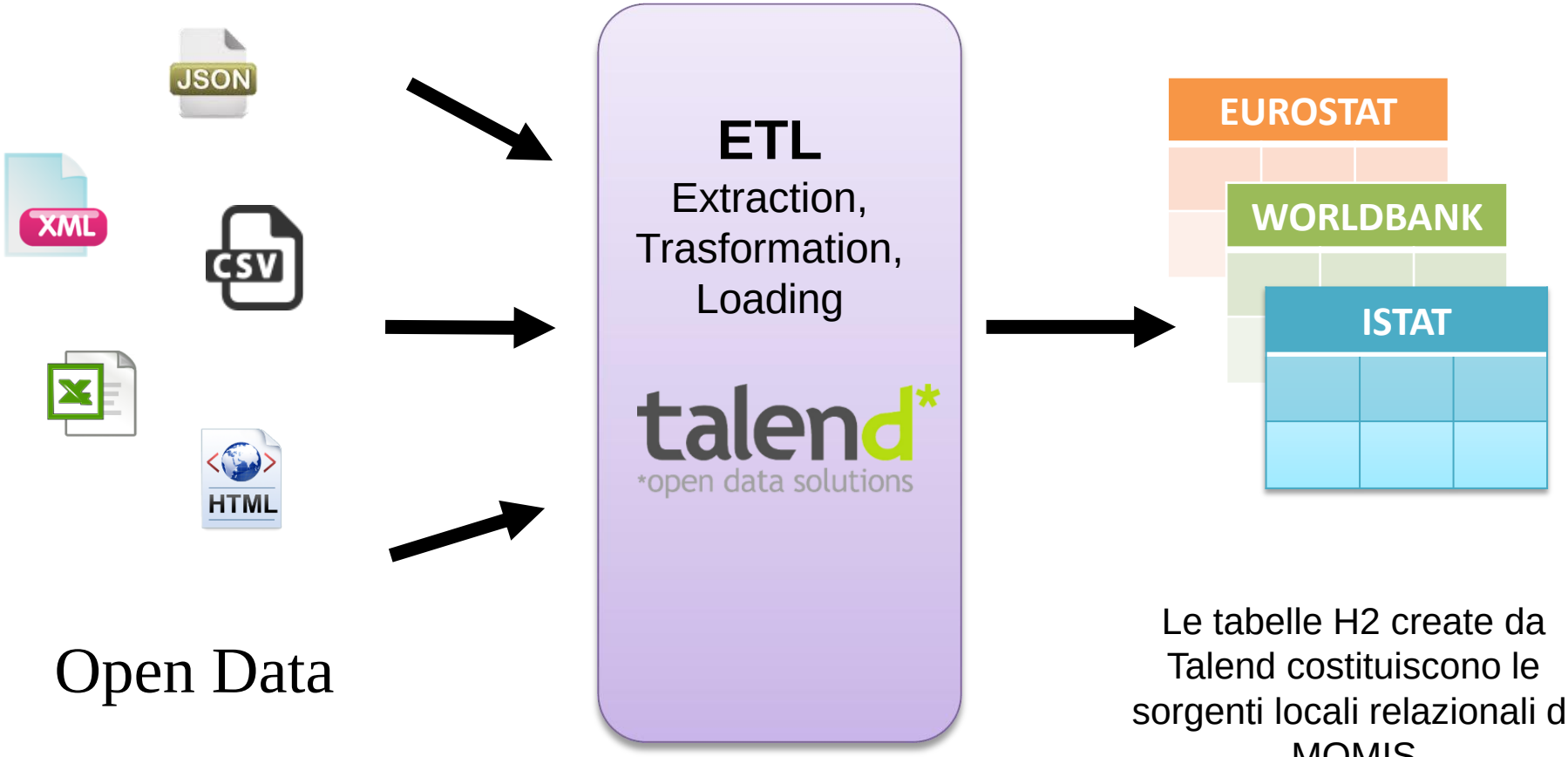
4) Il risultato dell'esecuzione del jar è inserito in tabelle H2 (hsql)

5) *MOMIS* elabora i dati all'interno delle tabelle



Prima fase

Estrazione, Trasformazioni e caricamento dei dati pubblici con *Talend*

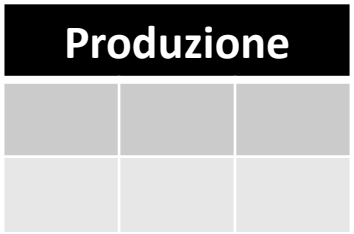


Open Data

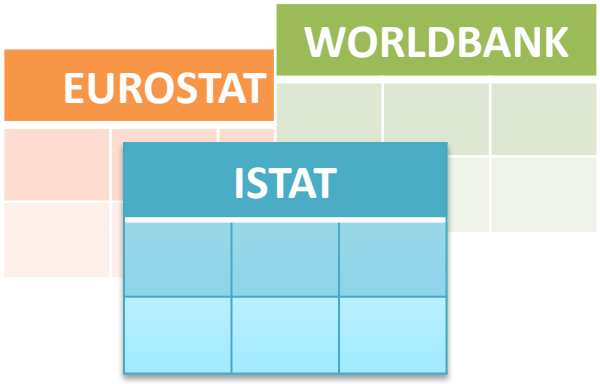
Le tabelle H2 create da Talend costituiscono le sorgenti locali relazionali di MOMIS

Seconda fase

Integrazione dei dati pubblici (individuati in fase di analisi) con i dati presenti nei sistemi informativi aziendali



Dati contenuti nei sistemi aziendali



Sorgenti locali relazionali di MOMIS

Motore semantico semi-automatico di MOMIS che effettua l'integrazione dei dati

Obbiettivi raggiunti

- Analisi sorgenti *Open Data* e i vari formati eterogenei.

- Analisi e confronto strumenti ETL *Open Source*, *Talend* e *kettle*

- Estrazione dei dati con strumenti ETL Open Source, *Talend* e *Kettle*

- Accoppiamento di uno strumento ETL, *Talend Open Studio,* in MOMIS

- Ho affrontato e risolto le problematiche di estrazione e integrazione di *Open Data*

FINE

Grazie per l’attenzione