Next Generation
Business Intelligence

Combining XBRL and Semantic Web Data

Christian Leibold

Contributions from the MUSING consortium:
Alessandro Tomassi, Paolo Lombardi,
Monika Jungemann-Dorner, Marcus Spies,
Horacio Saggion, Thierry Declerck

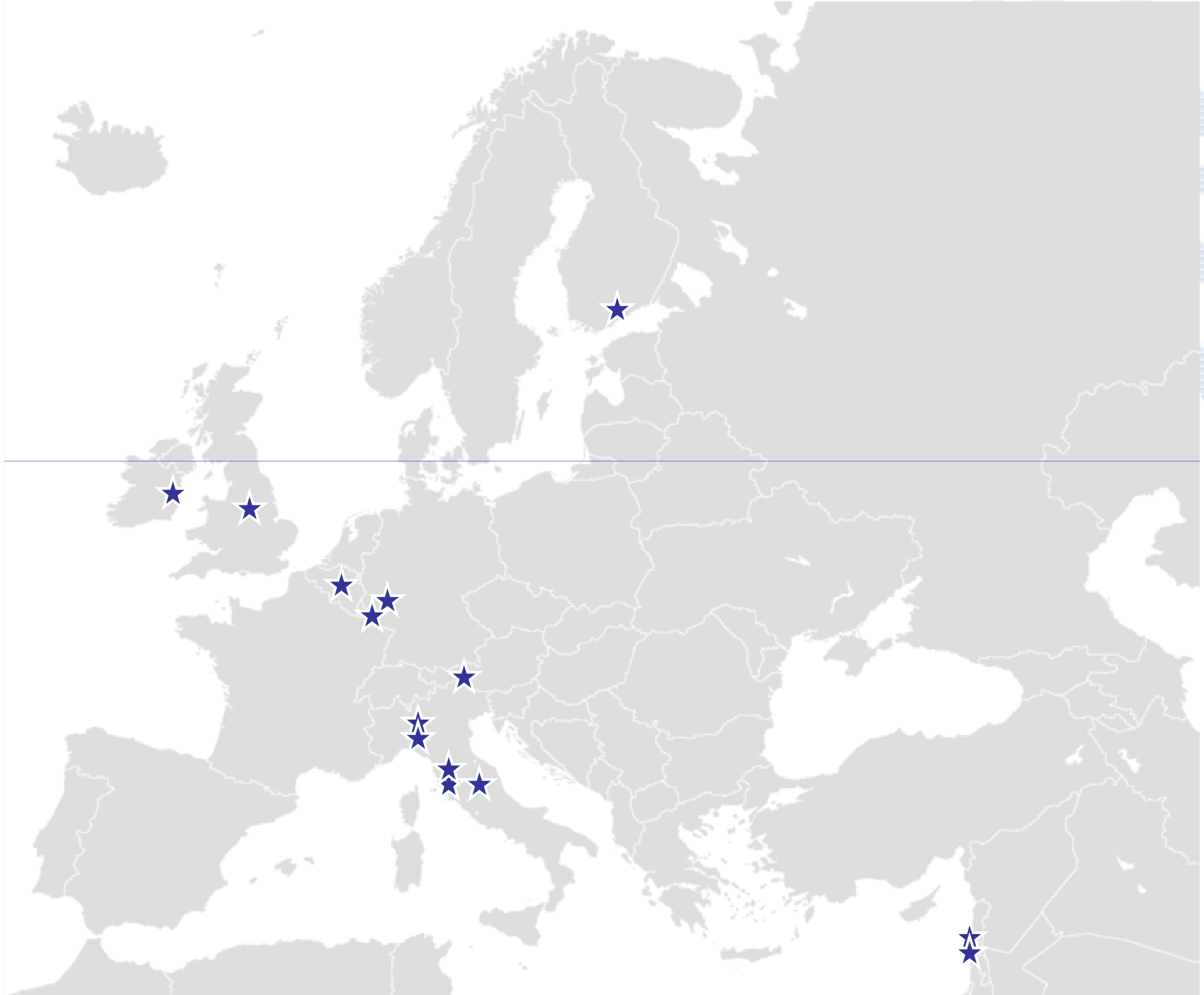
Summary

- The MUSING initiative www.musing.eu
- MUSING ontology motivation and overview
- Integrating XBRL in the MUSING ontology family
- Accessing the information – an example from credit risk management field test
- First conclusions & outlook

The MUSING initiative

- MUlti-industry, Semantic-based next generation business INtelliGence
- Business Intelligence (BI) tools and modules based on semantic-based knowledge and content systems
- Integration of Semantic, Web and Human Language technologies
- Combination of declarative rule-based methods and statistical approaches for knowledge acquisition and reasoning in BI applications.
- Multi-industry impact with focus on three vertical domains:
  - **Finance** (Basel II and beyond) with particular reference to **Credit Risk Management**;
  - **Internationalization**, (i.e., evolve enterprises' business from a local to an international dimension, hereby expressly focusing on the information acquisition work concerning international partnerships, contracts, investments)
  - **Operational Risk Management**, measurement and mitigation tools, with particular reference to operational risks faced by **IT-intensive organizations**.

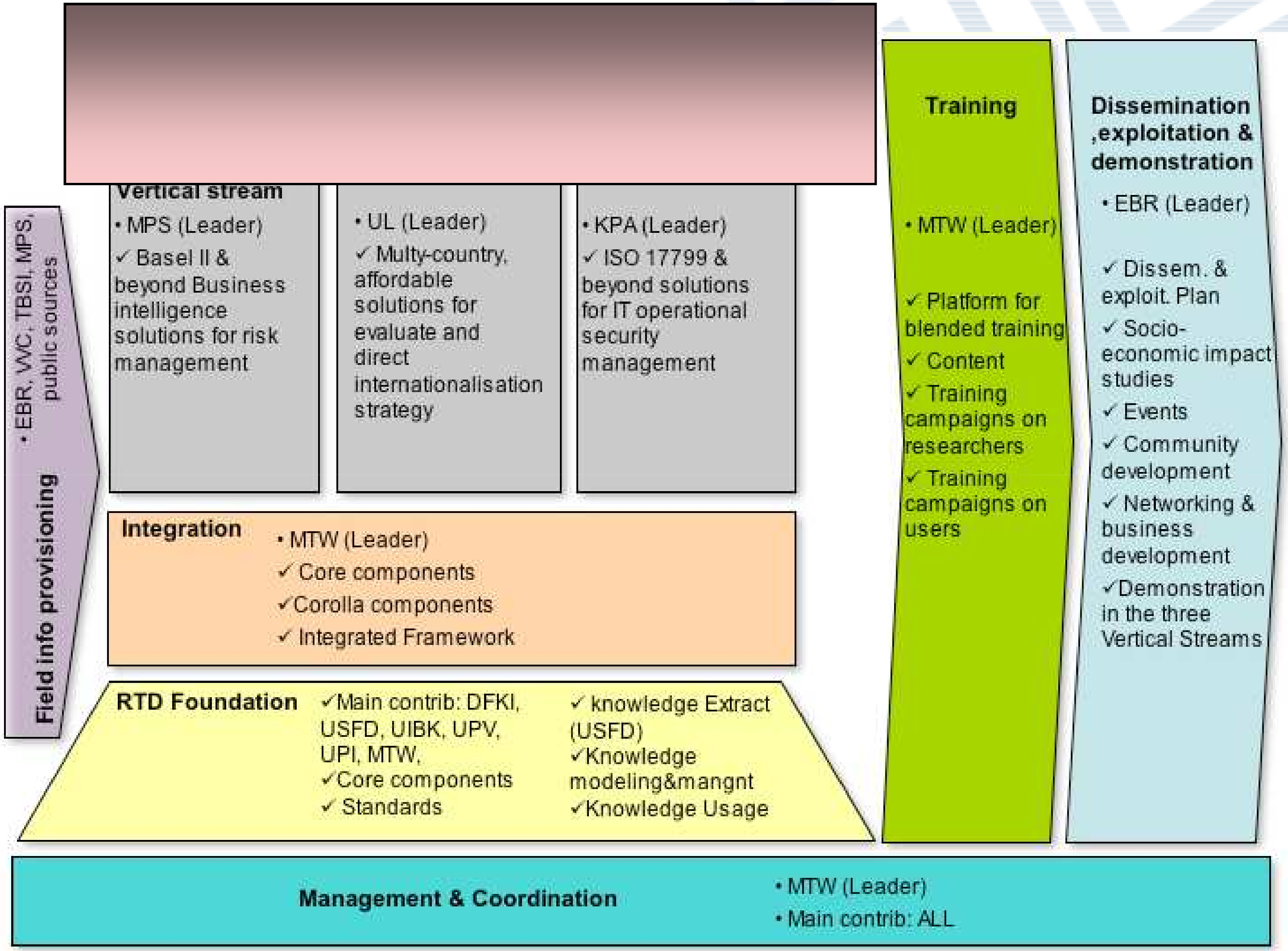
Consortium



- Metaware (CO)
- CI Consultancy Ltd.
- Verband der Vereine Creditreform
- Deutsches Forschungszentrum für Künstliche Intelligenz
- European Business Register
- KPA Ltd.
- MPSnet / Banca Monte dei Paschi di Siena
- Numerica
- University of Innsbruck
- University of Limerick
- University of Pavia
- University of Pisa
- Tadiran Telecom Communication Services
- TBSI
- University of Sheffield

+ Plancenter Finland
+ Il Sole 24 ore

Organisational setup (since 2006)



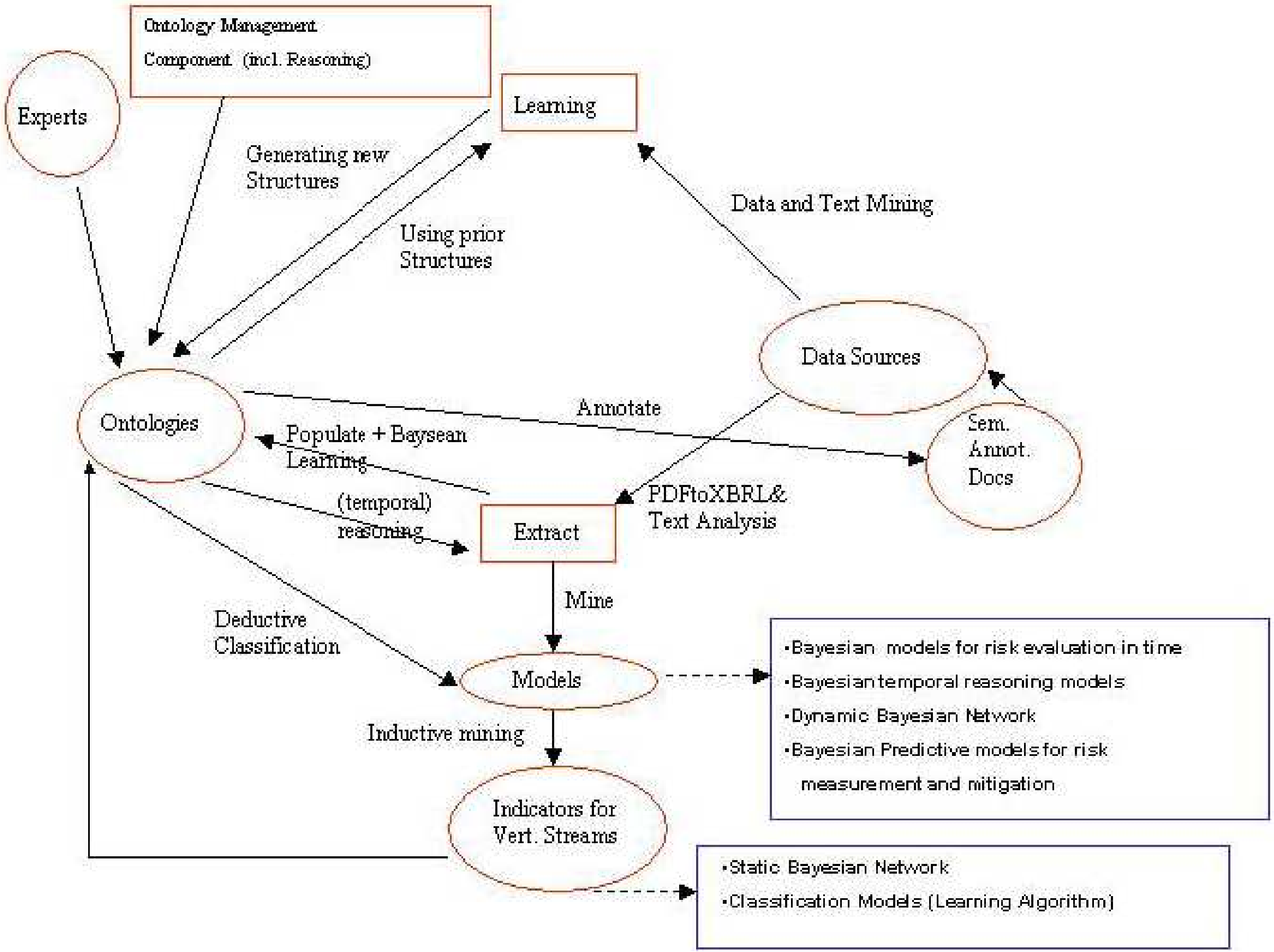
- 48 month

- until
  April 2010
  and beyond

Semantic-based approach

- Semantic-web and human language technologies to support the “next generation BI”:
  - Automatic annotation
  - Reasoning
  - Multi-linguality
  - Ontologies (temporally evolving)
  - Pervasive usage of XBRL
- The need for R&D → No off-the-shelf solutions
- The main benefits:
  - Automation in human-intensive analysis processes
  - Impact to a large user base
  - Basel II compliant services for the financial industries
  - Knowledge building

General overview of semantic technologies in MUSING



Business Intelligence requirements - benefits of and challenges for ontologies - Knowledge support

- ontologies can integrate multiple models qualified by extensible metadata
  - basic structure of entities and relationships
  - population of these structures for specific purposes
- ontologies are suitable as model repositories for access by business applications
  - Visualization
  - DB capabilities
- ontologies can accommodate knowledge structures that are dynamically updated or statistically optimized
- time dimension critical for data warehouse related services
- serve as reference for semantic text annotation
- support query languages (SPARQL) for knowledge reuse

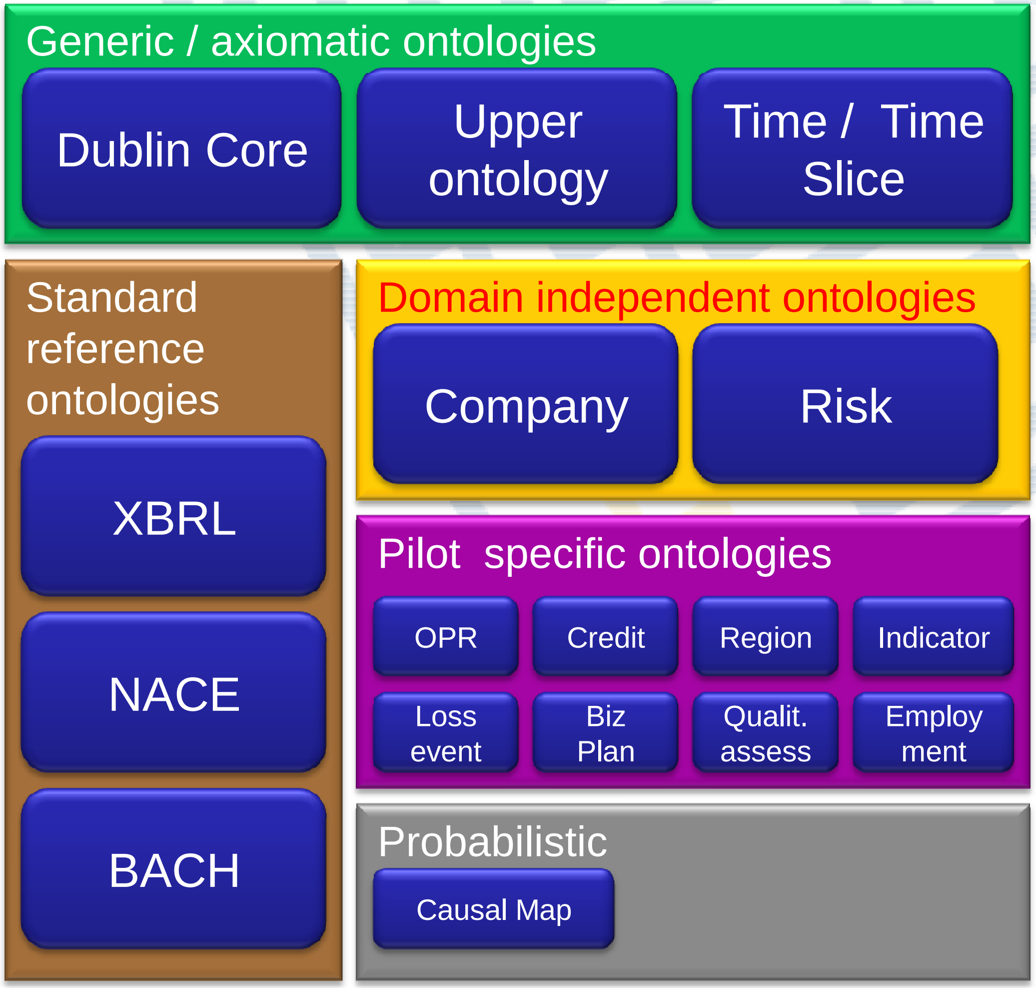
Business Intelligence requirements for ontologies - Challenges (II) – Standard compliance

- MUSING ontologies must enable compliance with reporting and classification standards (regulatory compliance).
  - XBRL accounting principles
    - taxonomy-like structure of balance sheet entries that can be composed to yield analytic quantities
  - NACE codes (Nomenclature of economic activities)
    - taxonomy without explicit classification criteria
  - BACH database information (Bank for the Accounts of Companies Harmonised)
    - coarse version of XBRL-like taxonomies
  - Basel II loss event classification
    - usable in very different modeling contexts

→ Ontologies are ideal means for knowledge models and management in MUSING applications

MUSING ontologies – conceptual model

- **layered** structure comprising
  - **general** level for „upper“ ontologies
    - Time, Meta ontologies
  - **standards** level for adapting industry standards to MUSING
    - NACE, XBRL, BACH
  - **domain** level for ontologies relevant to one or more vertical streams (company, risk)
  - **pilot** level for classes and relationships specific or adapted to specific application needs



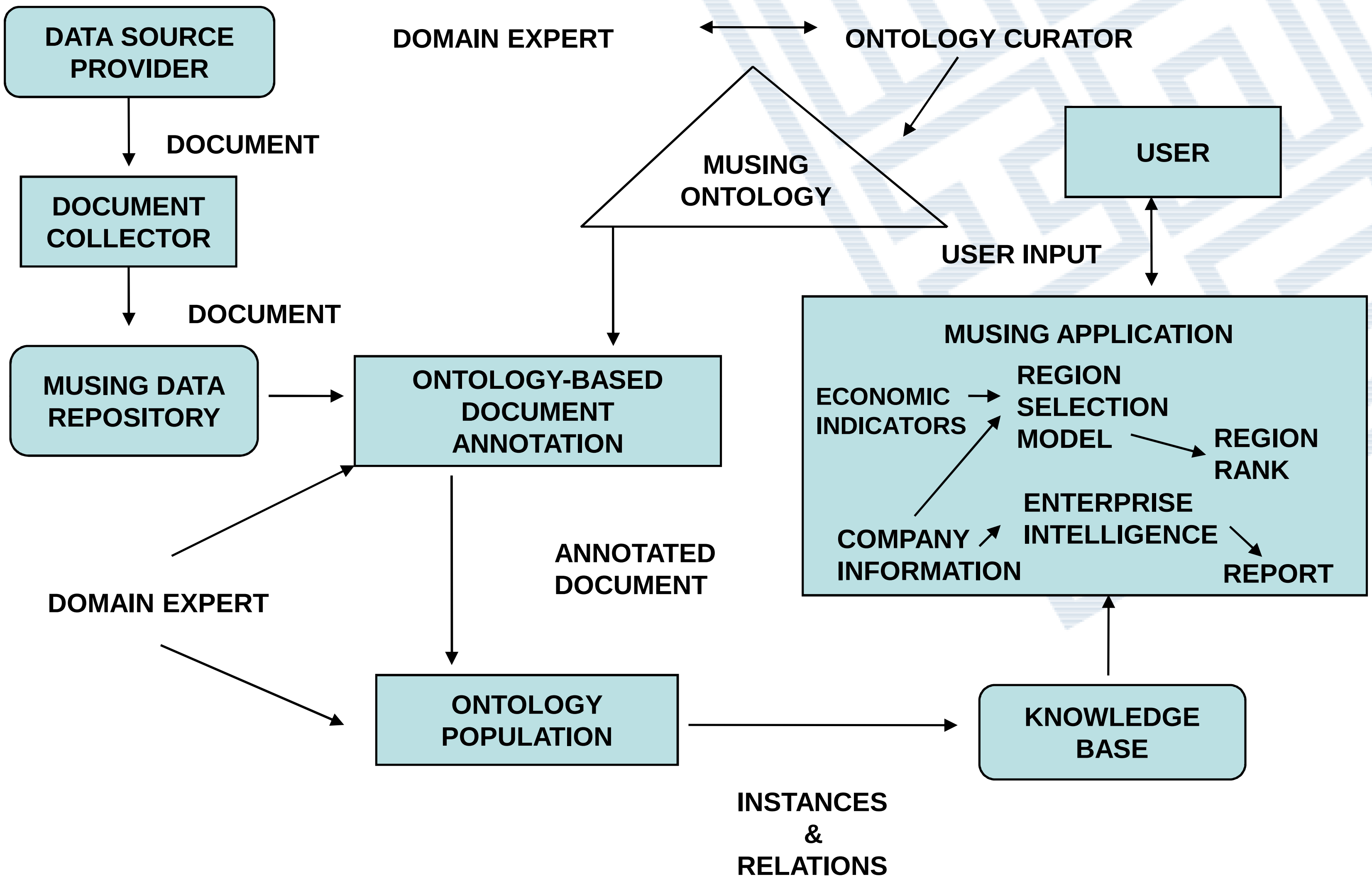
Processing Structured & Unstructured Data

- Ontology-driven analysis of both structured and unstructured textual data
  - Structured Data
    - Profit & Loss tables (which are structured but not *normalized:* extracting from the tables the data (terms, values, dates, currency, etc.) and map them into a normalized representation like XBRL,
    - Company Profiles and International Reports, which give detailled information about company (name, address, trade register, share holders, management, number of employees etc.)
  - Unstructured Data
    - Annexes to Annual Reports, On-Line financial articles, questionnaire to credit institutions etc.
- The Challenge: Merging data and information extracted from various types of documents (structured and unstructured), using a combination of Ontologies/Knowledge Bases, linguistic analysis and statistical models

Information Extraction

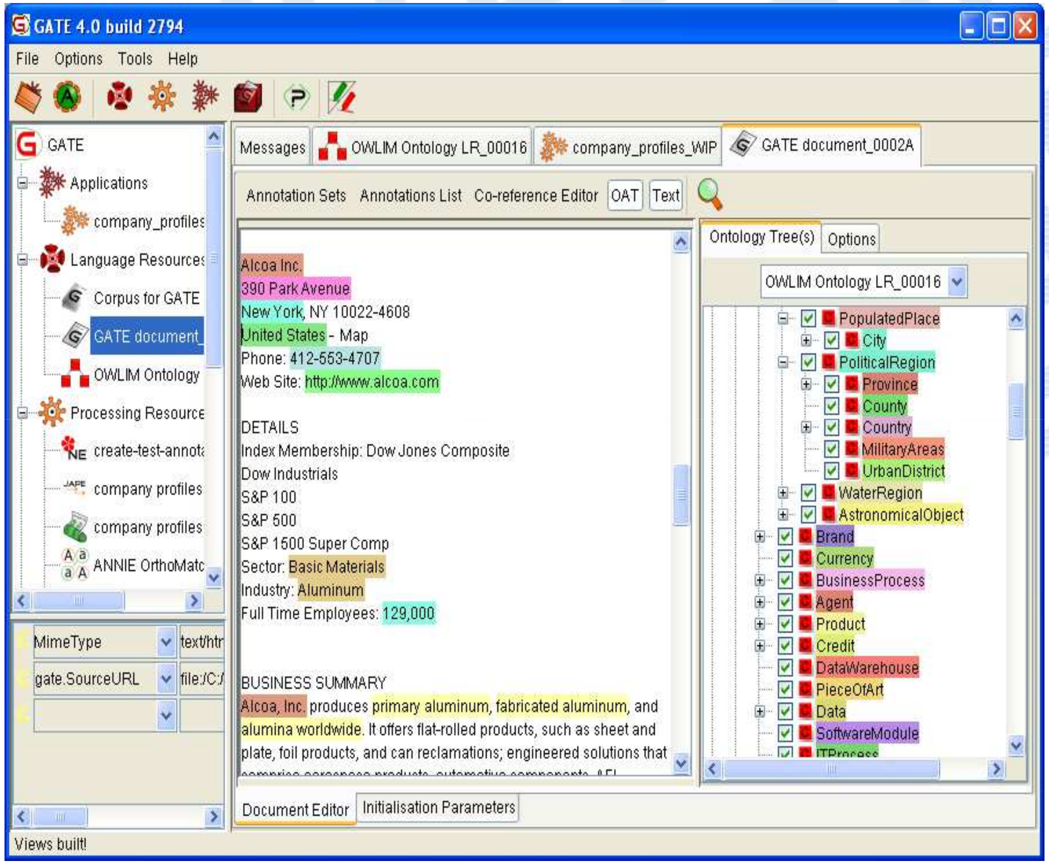
- Information extraction (IE) is a technology which extracts key pieces of information from text
  - generic: identify specific name mentions in text (person names, location names, money, etc.)
  - specific: populate a structured representation (e.g. template) with “strings” from text (e.g., full information on a joint venture)
  - Information extraction has been applied in Business applications in the past: identify management succession events; identify ship sinking events; etc,
  - Message Understanding Conferences & Automatic Content Extraction evaluation frameworks
- Ontology Based Information Extraction (OBIE) is the process of finding in text and other sources concepts, instances, and relations expressed in an Ontology

OBIE in MUSING



Extracting Company Information

- Extracting information about a company requires for example identify the Company Name; Company Address; Parent Organization; Shareholders; etc.
- These associated pieces of information should be asserted as properties values of the company instance
- Statements for populating the ontology need to be created (“Alcoa Inc” hasAlias “Alcoa”; “Alcoa Inc” hasWebPage “http://www.alcoa.com”, etc.)

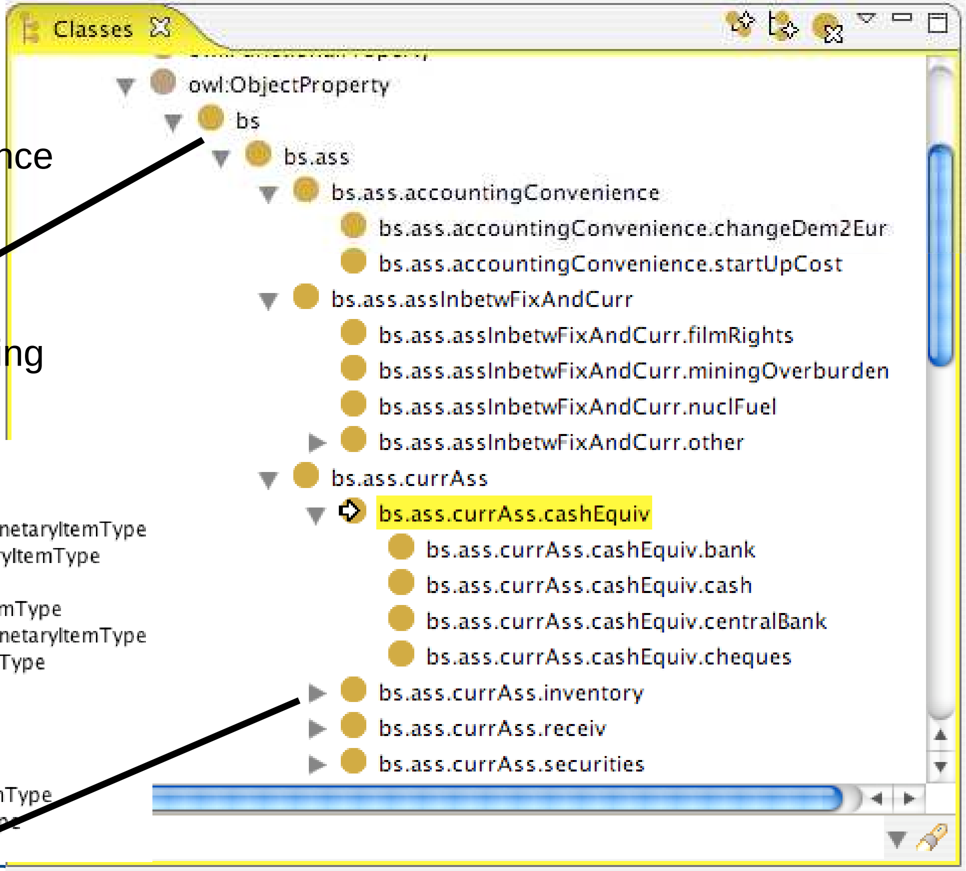


Example of processing of Structured data sources

- The PDFtoXBRL tools
  - Extract financial tables from PDF documents (Annual reports of companies)
  - Reconstruct a tabellar representation of the information contained in the tables (dates, amount, financial terms etc.) and annotate those with the corresponding semantics
  - Tables are structured but not normalized: Results are valid ontology concept and XBRL instances (for example de-GAAP).
  - Good quality so far: depending on the qualitiy of the processable input document: 75% up to 95% F-Measure.

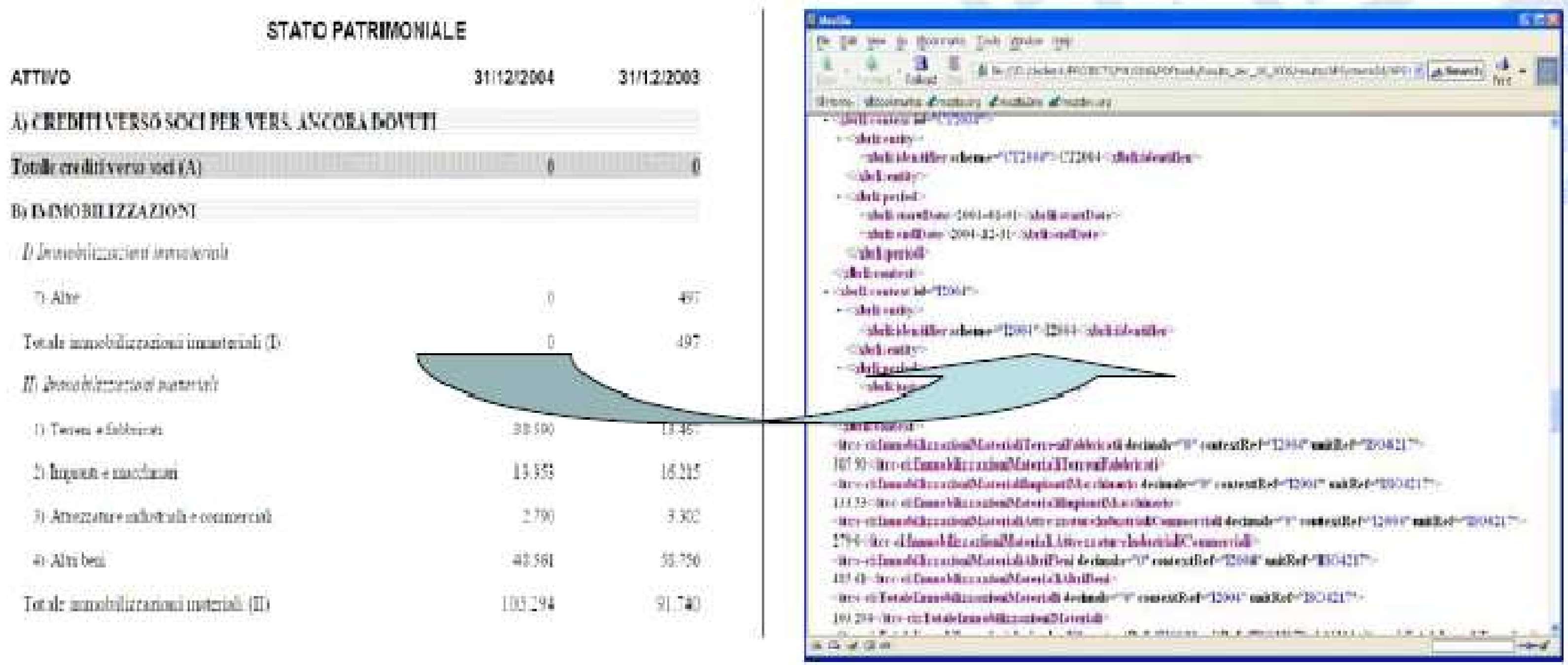
XBRL taxonomy representation for extraction

- Linking taxonomies to ontologies: exploiting the full XBRL potential
- MUSING ontology architecture combines ontologies with relational tables
  - XBRL represented as property taxonomy
  - This is analogous to a relational table for balance sheets BUT
    - we make the XBRL hierarchy explicit by subclassing
    - we gain with the flexibility of labeling, linking metadata etc.



Structured Data in the Scenario

- Profit & Loss tables etc. are structured but not *normalized.*
  - First processing step consists in automatically extracting from the balance tables the data (terms, values, dates, currency, etc.) and map them into a XBRL representation (the MUSING PDF2XBRL tools)

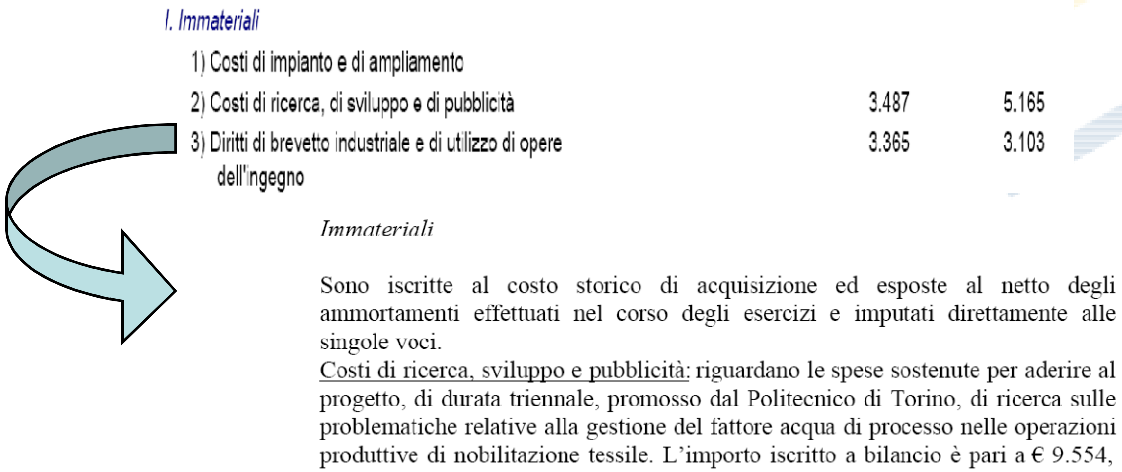


10/6/2009

Unstructured Documents

Aligning the normalized quantitative information in the financial tables with the relevant text parts in the annex documents.

- normally availble only in unstructured forms (free text)
- Linguistic and semantic analysis of such textual documents results in Semantic metadata that enrich the original document,
- towards a XBRL normalization of the unstructured text,
- Making information available for reuse.



I. Immateriali

1) Costi di impianto e di ampliamento		
2) Costi di ricerca, di sviluppo e di pubblicità	3.487	5.165
3) Diritti di brevetto industriale e di utilizzo di opere dell'ingegno	3.365	3.103

Immateriali

Sono iscritte al costo storico di acquisizione ed esposte al netto degli ammortamenti effettuati nel corso degli esercizi e imputati direttamente alle singole voci.
<u>Costi di ricerca, sviluppo e pubblicità</u>: riguardano le spese sostenute per aderire al progetto, di durata triennale, promosso dal Politecnico di Torino, di ricerca sulle problematiche relative alla gestione del fattore acqua di processo nelle operazioni produttive di nobilitazione tessile. L'importo iscritto a bilancio è pari a € 9.554,

Linguistic Structuring

- Dr. „Ernst Mustermann“ ist Mitglied des Aufsichtsrats seit dem 7. März 2005.
  - Using both „Constituency“ (red marks below) and „Dependency“ (blue marks below)

- [$_{\text{NP-PERS}}$ Dr. „Ernst Mustermann“ $_{\textit{SUBJ}}$][$_{\text{VG}}$ ist $_{\textit{pred-sein}}$][$_{\text{NP}}$ Mitglied $_{\textit{head}}$ [$_{\text{NP}}$ des Aufsichtsrates $_{\textit{MOD}}$] $_{\textit{PRED-OBJ}}$][$_{\text{NP-DATE}}$ seit dem 2005-03-0 $_{\textit{DATE-MOD}}$]

- We can do that in Multi-lingual scenario

Example of a XBRL Taxonomy for a Specific Legislation: BNB - Multilingualism

- <label xlink:label="WithdrawalFromAllocatedFunds_lab" xlink:type="resource" xlink:role="http://www.xbrl.org/2003/role/documentation" xml:lang="fr">Prélèvements sur les fonds affectés</label>
- <label xlink:label="WithdrawalFromAllocatedFunds_lab" xlink:type="resource" xlink:role="http://www.xbrl.org/2003/role/verboseLabel" xml:lang="nl">Onttrekking aan de bestemde fondsen</label>
- Further „semantic“ specification of a term:
  – <element name="WithdrawalFromAllocatedFunds" type="pfs-dt:nonNegativeMonetary14D2ItemType" abstract="false" substitutionGroup="xbrli:item" nillable="false" id="pfs_WithdrawalFromAllocatedFunds" xbrli:balance="credit" xbrli:periodType="duration"/>

Temporal Information

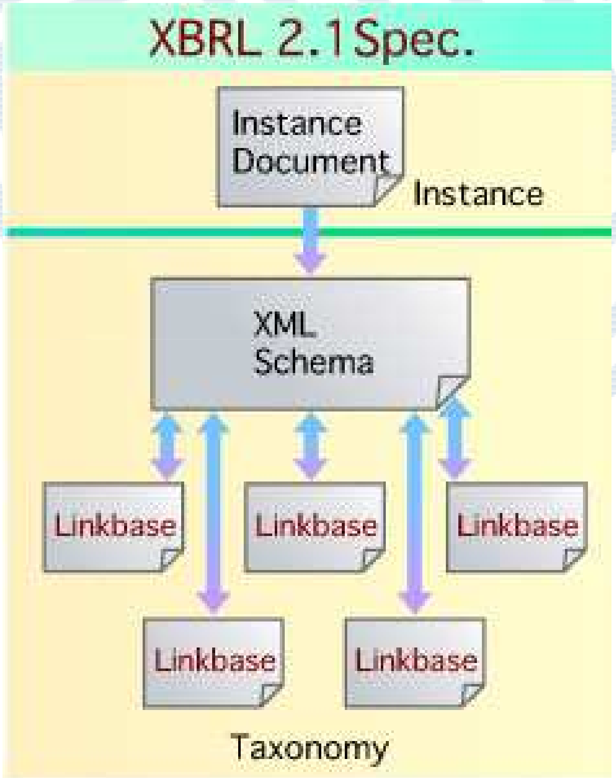
- In the example last slide, we notice that publication dates and validaty date of financial reports are not enough in order to gain information from Annual reports. Need to encode event dates (beyond XBRL taxonomy)
- Need to encode temporal dependencies. So functional attributes (like CEO_of) can in fact have more than one value in the reports, but the temporal information allows to „justify“ the information.
- In MUSING we developed a temporal representation framework which can is integrated within the (OWL) ontologies of MUSING (Perdurant / Time Slice)

Information fusion

- Not only linking, but Merging/Fusion of data from various sources, using an ontologized version of XBRL
- Combination of several Taxonomies, Ontologies and Knowledge Bases (XBRL, OWL) with deep linguistic analysis for ontology population (enriching the MUSING specific knowledge base)

Combining XBRL – structural recap

- XBRL recap
  - taxonomy-like structure of balance sheet entries that can be composed to yield analytic quantities
- XBRL class has nine associated properties
  - Four relevant for computation of information contained in XBRL instance:
    - Item, Context, Tuple, Unit
  - Five that make up the taxonomy and make up the XBRL Linkbase namespace documents
    - ArcRoleRef, FootnoteLink, LinkbaseRef, RoleRef, SchemaRef,



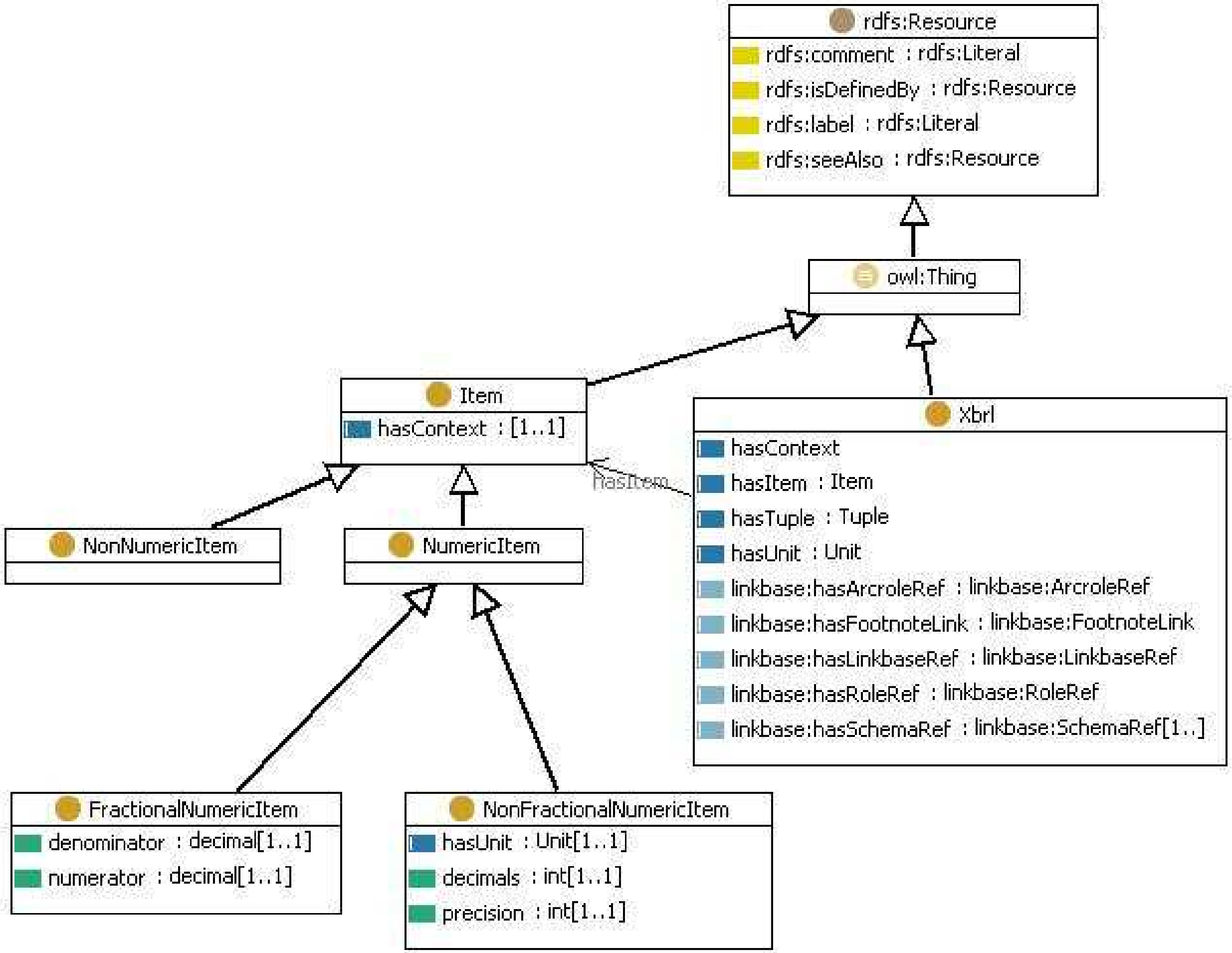
Deriving ontologies from XBRL structures

- Only a small part of the ontology is general across XBRL taxonomies (GAAP, IFRS // DE, IT, US, ...)

However: It is possible to construct ontologies directly from XBRL schemata and linkbase with tool support

- useful spec ODM (ontology definition metamodel) by OMG
  - allows, e.g., „n-ary relational“ transformation from / to ontologies
  - we use ODM metamodels for integration of XBRL and OWL structures

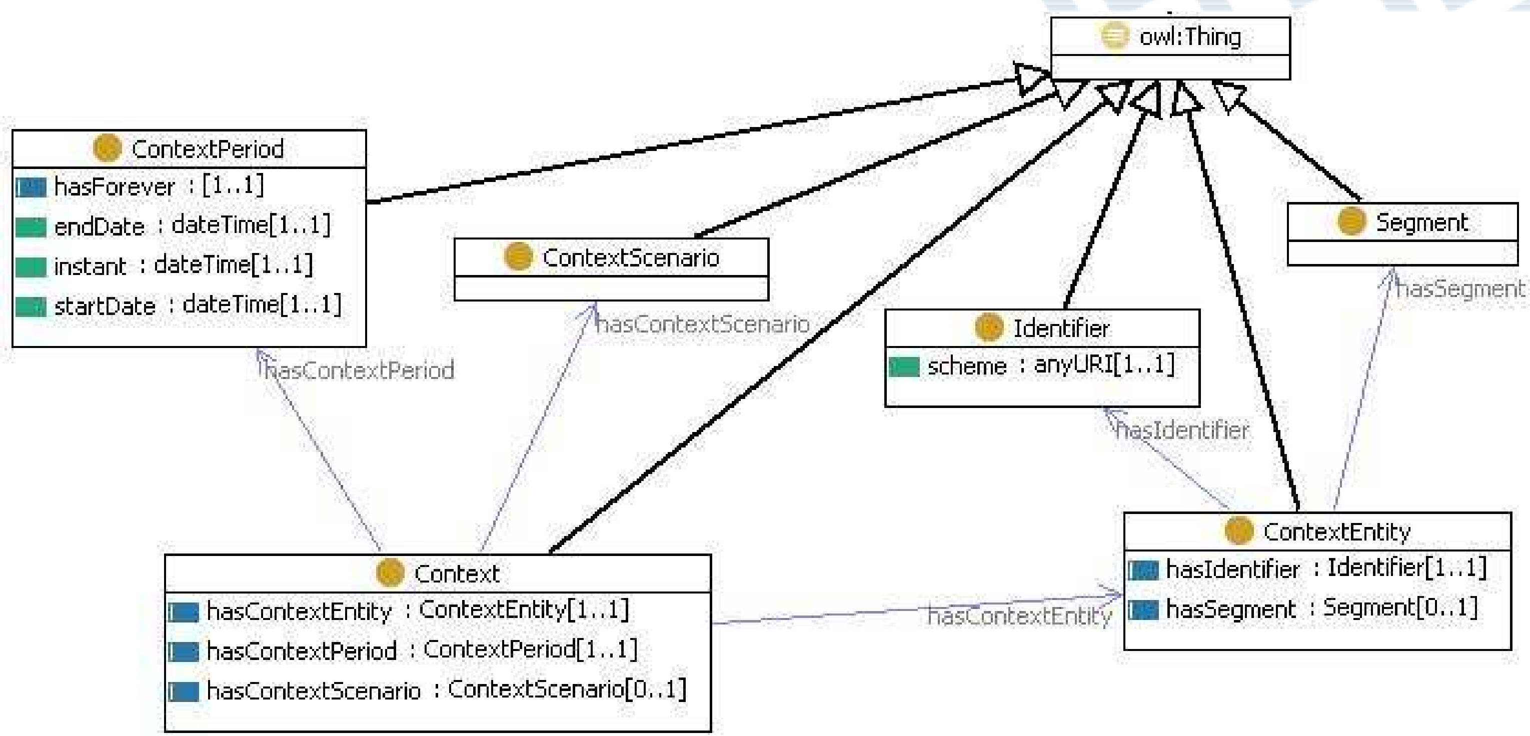
XBRL ontology for integration



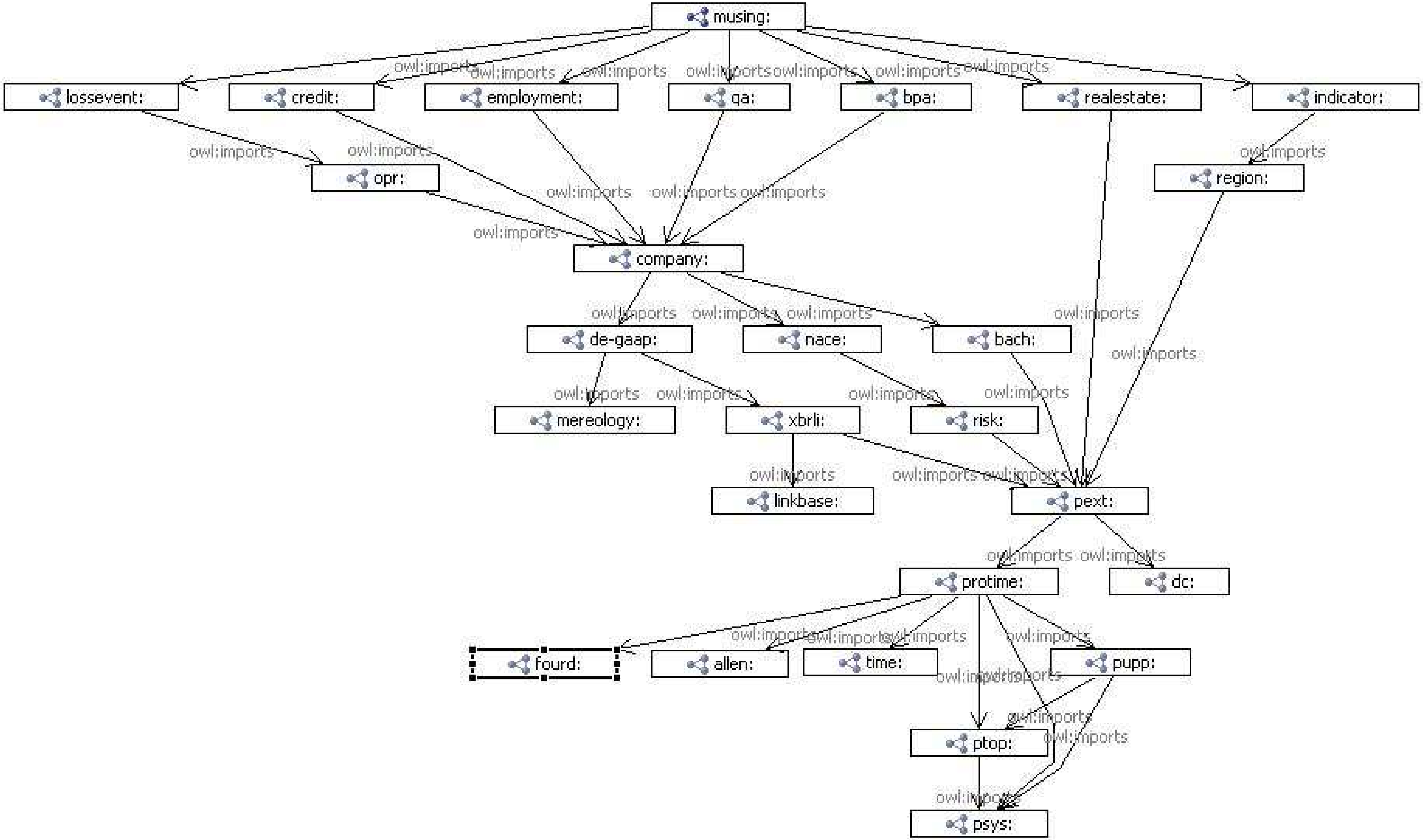
- XBRL relies on conceptual hierarchies for accounting
- XBRL provides rich semantic annotations for reporting data
- proposal to integrate XBRL and domain ontologies
- showcased for

Context concept and domain ontologies

- Concept Context linked as attribute to concept XBRL
- ContextEntity as „owl:*equivalentClass*“ of LegalEntity (e.g. a company) enables connection to domain ontologies



Excerpt from current import diagram

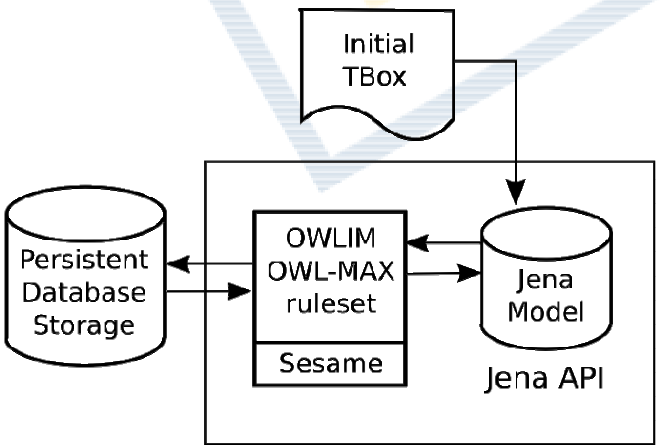


Accessing the ontologies form the web

- Introducing an applied reasoning architecture and accomplishing ontology persistence task

→**A persistent repository for RDF**

- OWLIM forward reasoner does the TBox and ABox reasoning
- Relational database backend is used for the persistency
- Repository is initialized with the MUSING ontology schemata
- Full closure of inferred triples are dumped to DB
- Interfaces for precomputed facts:
  - SPARQL queries (select & update)
  - XML-RPC interface
  - WSDL available
- Integration in MUSING services that are delivered as BPEL processes



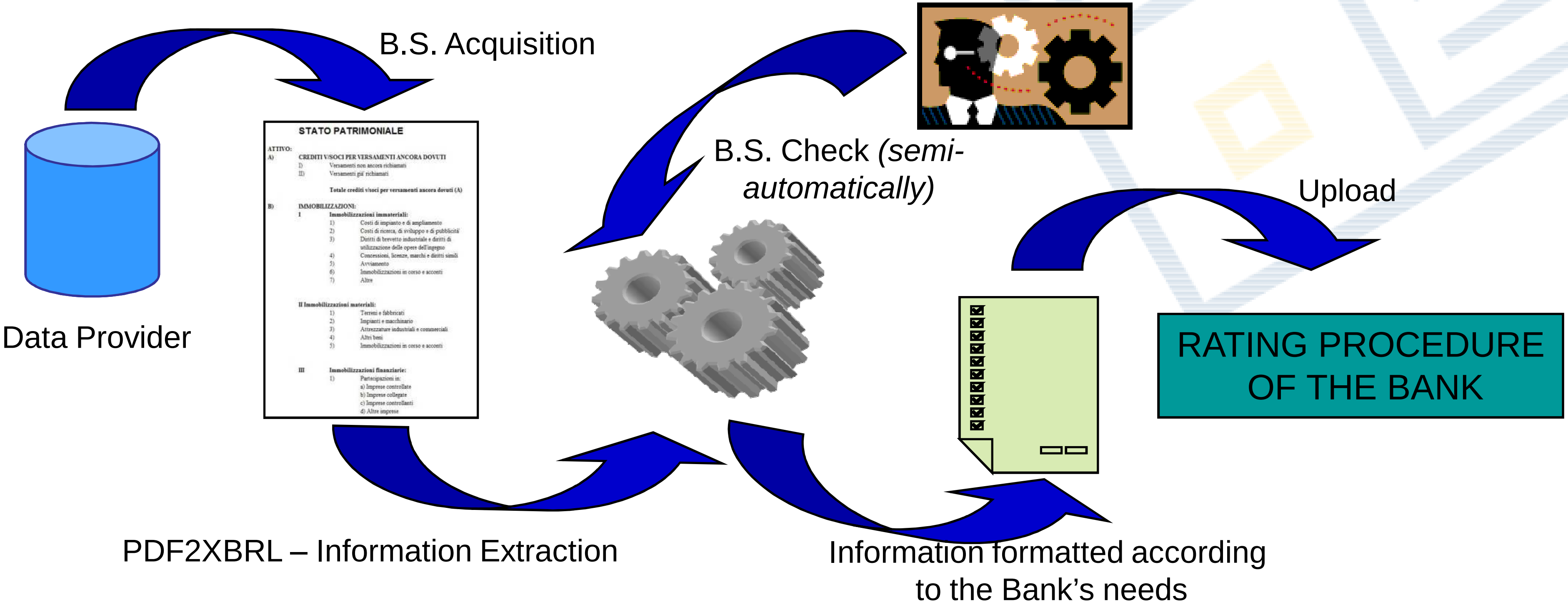
MUSING services

- Financial Risk Management value chain & services

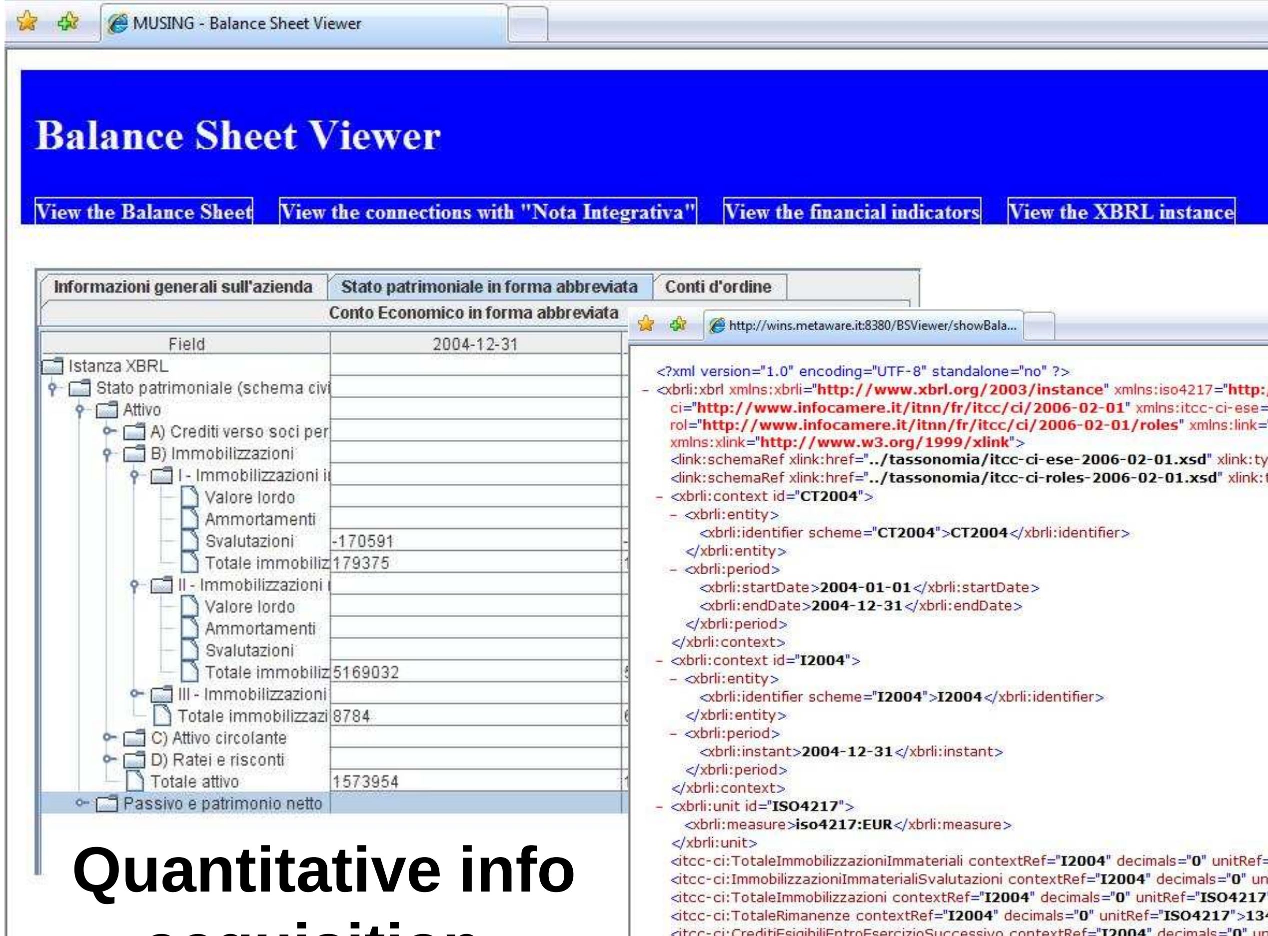
FRM Value Chain	***Data Acquisition***	***Information Processing***	***Risk Measurement***	***Credit Issuing and Communic.***	***Credit Management***	***Delivery***
Process Description	Internal and external data acquisition	Quantitative and Qualitative information processing	Credit scoring techniques Rating	Credit issuing and comunication of information dealing with the rating evaluation	Management of credit issuing, credit monitoring and credit collecting process	BI product packaging and delivery
Services		• Quantitative and qualitative info extraction • BS processing • Business Inspector	• Bayesian quali-quantitative information integration for PD measurement			• EU-harmonised payment pool

Field-tested services:
1. Balance Sheet processor

- Objective: automate upload of enterprises Balance Sheets (e.g., into an IRB system), according to a given XBRL taxonomy (tested: Germany & Italy).
- Current performances: precision>75%, recall ~95%.



Field-tested services:
1. Balance Sheet processor - ctd

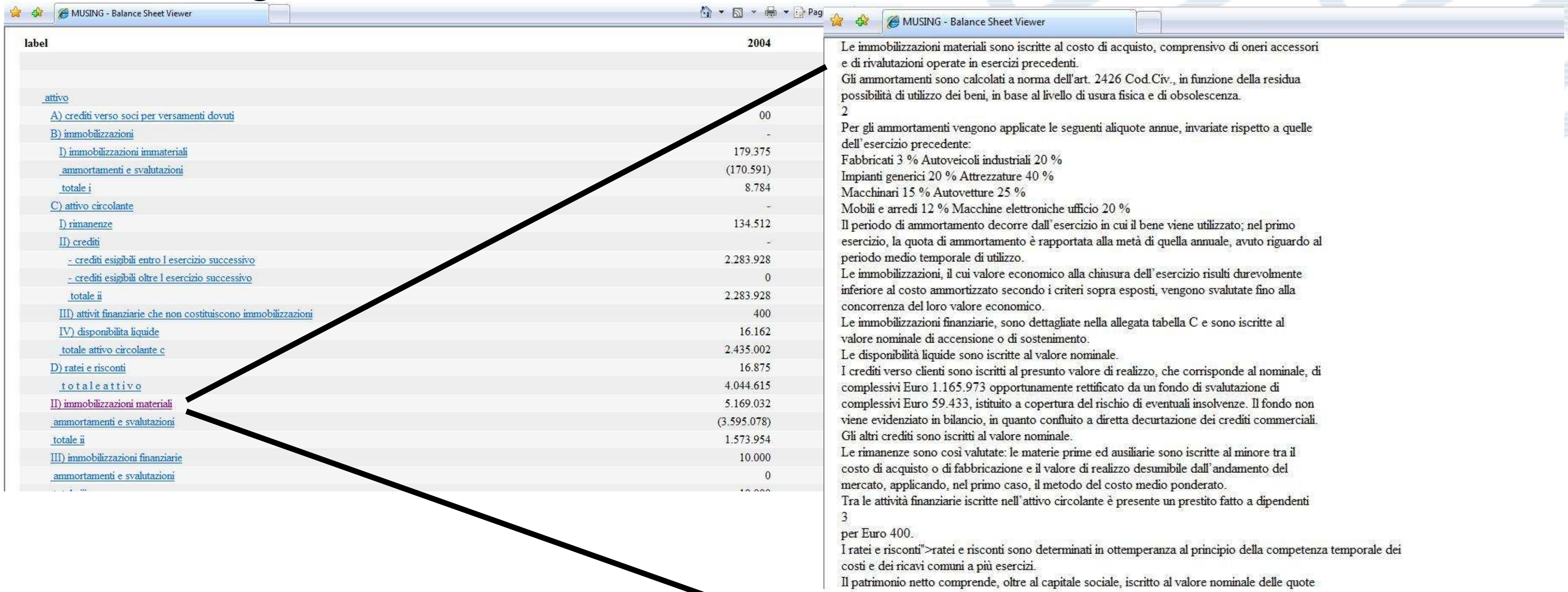


Quantitative info acquisition

XBRL instance

Field-tested services:
2. Qualitative info acquisition

- ❑ Objective: automate acquisition of qualitative information
- ❑ Current performances: questionnaire acquisition, linkages Balance Sheet/unstructured annex



MUSING - Balance Sheet Viewer

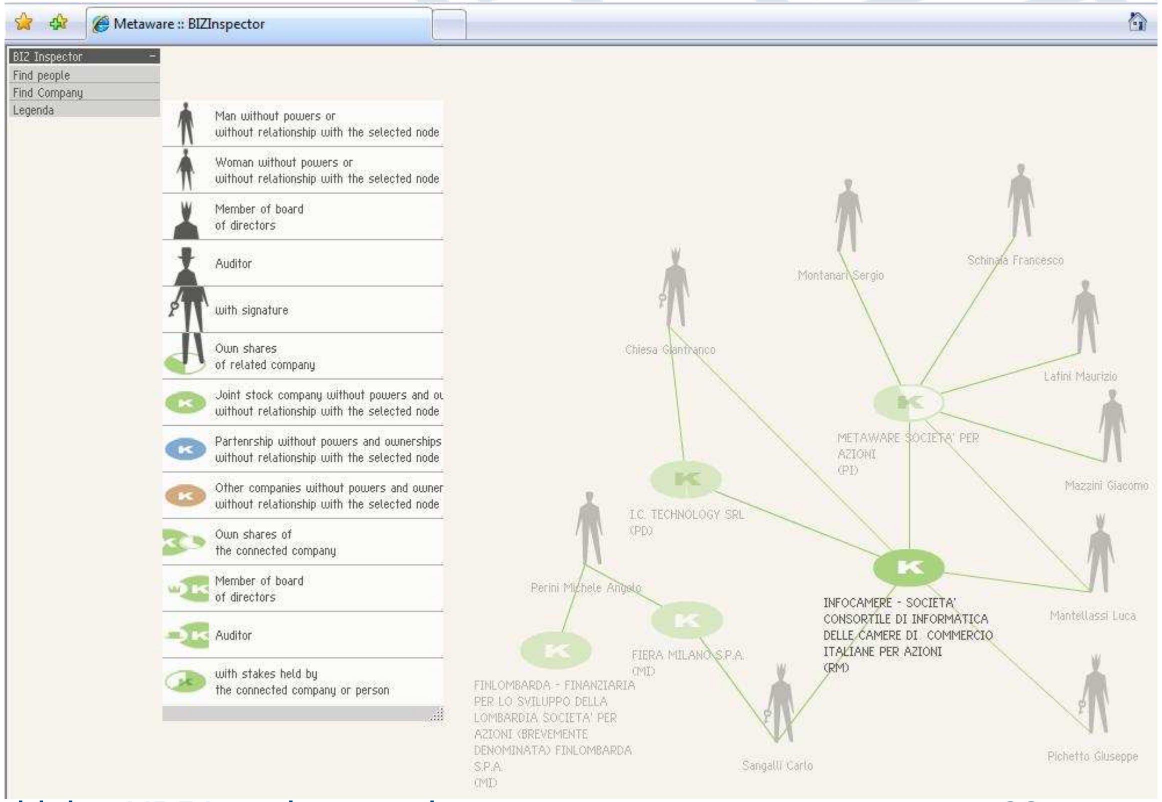
label	2004
attivo	
A) crediti verso soci per versamenti dovuti	00
B) immobilizzazioni	-
I) immobilizzazioni immateriali	179.375
ammortamenti e svalutazioni	(170.591)
totale i	8.784
C) attivo circolante	-
I) rimanenze	134.512
II) crediti	-
- crediti esigibili entro l esercizio successivo	2.283.928
- crediti esigibili oltre l esercizio successivo	0
totale ii	2.283.928
III) attivit finanziarie che non costituiscono immobilizzazioni	400
IV) disponibilita liquide	16.162
totale attivo circolante c	2.435.002
D) ratei e risconti	16.875
t o t a l e a t t i v o	4.044.615
II) immobilizzazioni materiali	5.169.032
ammortamenti e svalutazioni	(3.595.078)
totale ii	1.573.954
III) immobilizzazioni finanziarie	10.000
ammortamenti e svalutazioni	0

MUSING - Balance Sheet Viewer

Le immobilizzazioni materiali sono iscritte al costo di acquisto, comprensivo di oneri accessori e di rivalutazioni operate in esercizi precedenti.
Gli ammortamenti sono calcolati a norma dell'art. 2426 Cod.Civ., in funzione della residua possibilità di utilizzo dei beni, in base al livello di usura fisica e di obsolescenza.
2
Per gli ammortamenti vengono applicate le seguenti aliquote annue, invariate rispetto a quelle dell'esercizio precedente:
Fabbricati 3 % Autoveicoli industriali 20 %
Impianti generici 20 % Attrezzature 40 %
Macchinari 15 % Autovetture 25 %
Mobili e arredi 12 % Macchine elettroniche ufficio 20 %
Il periodo di ammortamento decorre dall'esercizio in cui il bene viene utilizzato; nel primo esercizio, la quota di ammortamento è rapportata alla metà di quella annuale, avuto riguardo al periodo medio temporale di utilizzo.
Le immobilizzazioni, il cui valore economico alla chiusura dell'esercizio risulti durevolmente inferiore al costo ammortizzato secondo i criteri sopra esposti, vengono svalutate fino alla concorrenza del loro valore economico.
Le immobilizzazioni finanziarie, sono dettagliate nella allegata tabella C e sono iscritte al valore nominale di accensione o di sostenimento.
Le disponibilità liquide sono iscritte al valore nominale.
I crediti verso clienti sono iscritti al presunto valore di realizzo, che corrisponde al nominale, di complessivi Euro 1.165.973 opportunamente rettificato da un fondo di svalutazione di complessivi Euro 59.433, istituito a copertura del rischio di eventuali insolvenze. Il fondo non viene evidenziato in bilancio, in quanto confluito a diretta decurtazione dei crediti commerciali.
Gli altri crediti sono iscritti al valore nominale.
Le rimanenze sono così valutate: le materie prime ed ausiliarie sono iscritte al minore tra il costo di acquisto o di fabbricazione e il valore di realizzo desumibile dall'andamento del mercato, applicando, nel primo caso, il metodo del costo medio ponderato.
Tra le attività finanziarie iscritte nell'attivo circolante è presente un prestito fatto a dipendenti
3
per Euro 400.
I ratei e risconti">ratei e risconti sono determinati in ottemperanza al principio della competenza temporale dei costi e dei ricavi comuni a più esercizi.
Il patrimonio netto comprende, oltre al capitale sociale, iscritto al valore nominale delle quote

Field-tested services:
3. Business Inspector

- Objective: explore complex connections among enterprises/individuals (e.g., ownership, board membership, cross-borders dependencies, etc.) and provide navigation-like interaction to the user.
- Current performances: tested on Italian data through EBR:
  - **Real-time**: Participations & Ownership; Partners & Members; Powers & Offices; Sector of Economic Activity.
  - "**asynchronous way**" (5 - 10 min): Annual Accounts; Deeds.



Opportunities for XBRL and semantic technology combination

- Multilingual instance document presentation
- Inclusion of free text data (e.g. unstructured part of SEC forms (8-K, 10-K), annexes to balance sheets, etc)
- detection of relevant information for further plausibility checks etc.
- Mapping of business rules in the ontology (automated linguistic annotation and - if necessary - augmentation of T-Box)

Further notes and conclusions

What makes OWL unique (as compared to RDFS or even XML Schema) is the fact that it can describe resources in more detail and that it comes with a well-defined model-theoretical semantics, inherited from description logic. Integration in BI possible.

Note: The MUSING ontology approach to XBRL is not committed to OWL but can be cast in more general UML modeling elements equivalently.

Publication: Spies, M. (2009). An ontology modeling perspective on business reporting languages. Information Systems, Elsevier, (in press, DOI information: 10.1016/j.is.2008.12.003).

Thanks for your attention

Multi-Industry Semantic-based Business Intelligence

www.musing.eu

Contact - christian.leibold@sti2.at

Acknowledgements **– Support from the European Commission, under the IST Programme, project IST-27097, is gratefully acknowledged**