http://www.monnet-project.eu

http://twitter.com/monnetproject

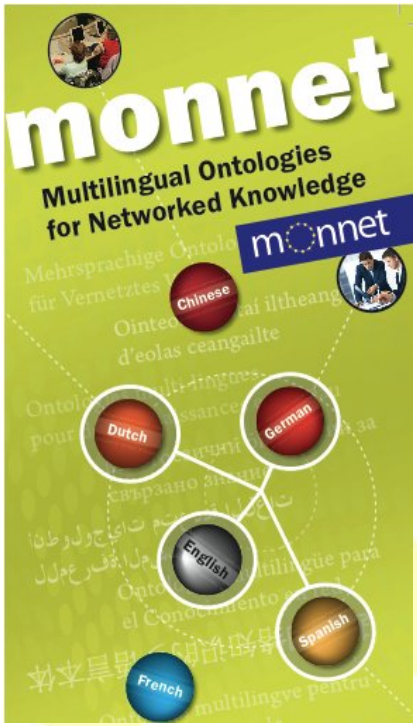
W3C Workshop:
The Multilingual Web - Where Are We?
26-27 October 2010, Madrid

lemon: An Ontology-Lexicon model for the Multilingual Semantic Web

MONNET Project Consortium, represented by
Thierry Declerck (DFKI GmbH), Paul Buitelaar & Tobias Wunner (DERI), John McCrae (CITEC), Elena Montiel-Ponsoda & Guadalupe Aguado de Cea (UPM)

Presentation Overview

- Monnet Project
  - Partners
  - Objectives
  - Financial Use Case
- Introduction to lemon
  - Why lemon?
  - Design of the lemon model
- Relations to other projects and to standardisation bodies
- Conclusion



Project Partners

- Industrial Organizations
  - SAP, Germany (http://www.sap.com/index.epx)
  - BeInformed, The Netherlands (/http://www.beinformed.nl)
  - XBRL-EUROPE, Belgium (http://www.xbrl.org/eu/)
- Research Institutes
  - DERI, Ireland (http://www.deri.ie/), coordinator
  - DFKI, Germany (www.dfki.de)
  - UPM, Spain (http://www.upm.es/institucional)
  - UniBi-CITEC, Germany (http://www.uni-bielefeld.de)

Monnet: Main Objectives

- Monnet addresses some of the topics identified by the "MultilingualWeb" project, http://www.multilingualweb.eu/:

  „The MultilingualWeb project is exploring standards and best practices that support the creation, localization and use of multilingual web-based information".

- With: Multilingual Ontologies for Networked Knowledge
  - Linguistically enriched knowledge representation
  - Multilingual access to structured/networked knowledge: ontologies, knowledge bases, linked data
- How: Handling Information at the Semantic Level
  - Ontology-Based Cross-lingual information Extraction, Storage and Presentation
  - Localization/Translation of information abstracting from language & layout

Monnet Financial Use Case

- Objective: multilingual access to **financial reports** for sharing business knowledge across Europe
  - Semantic-level analysis of business reports of companies
  - Search for financial information and financial report creation in the language of choice of the users
- Outcome: A prototype that will allow a financial analyst to search for information by filling in structured search forms localized to his/her own language
  - The results will be presented in terms of charts, diagrams, results lists etc. localized to the preferred language of the user

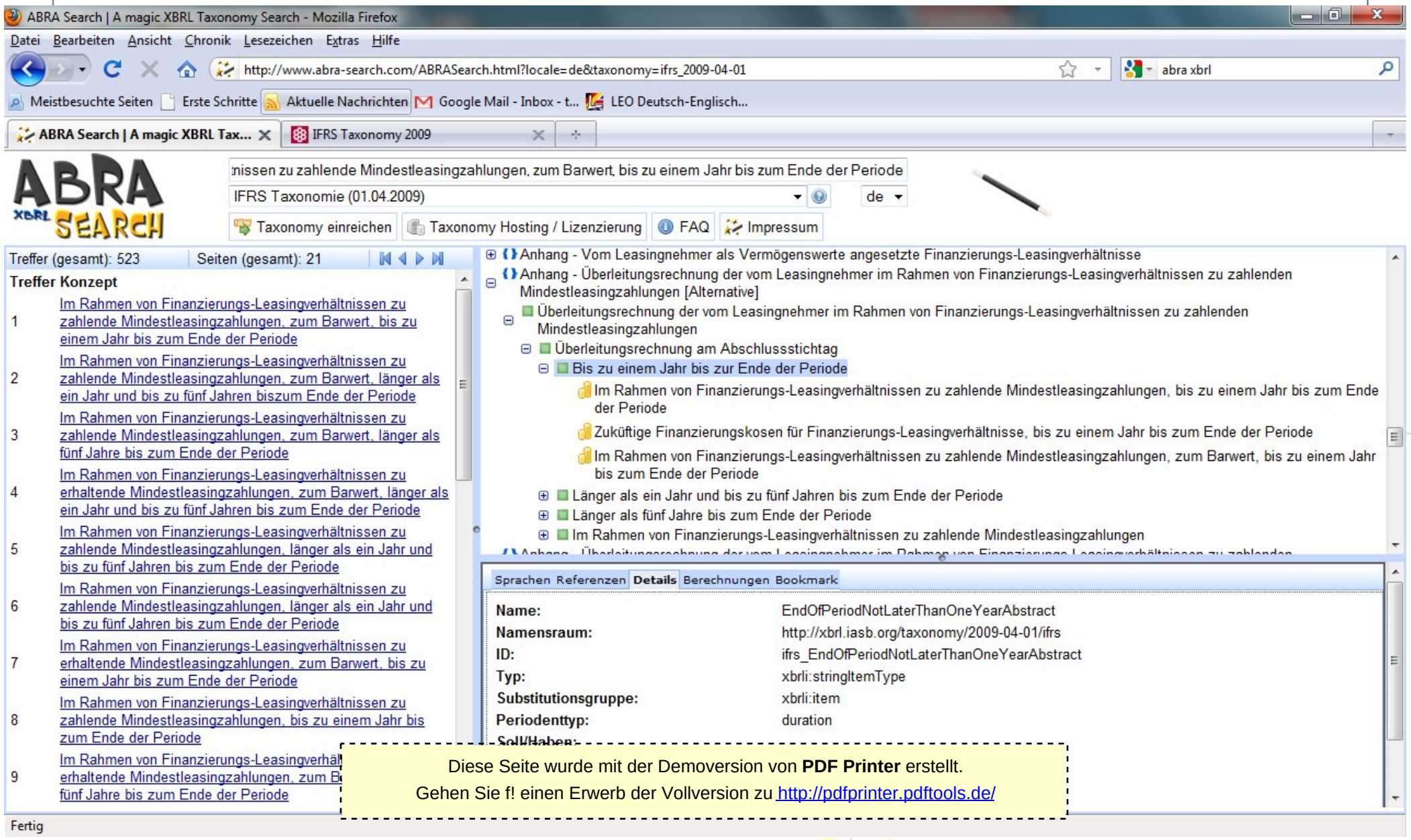
Knowledge Sources

- **XBRL** (eXtensible Business Reporting Language): an XML-based mark-up language for the exchange of business information, including financial reporting. Available taxonomies ecoded in XBRL:
  - **IFRS** (International Financial Reporting Standards): a standard developed by the International Accounting Standards Board (IASB). IFRS Terminology is **translated** in many languages.
  - **GAAP** (Generally Accepted Accounting Principles), taxonomies reflecting national and regional legislations on financial reporting. Terms in GAAP taxonomies are available in the national/regional languages, and maybe in English.

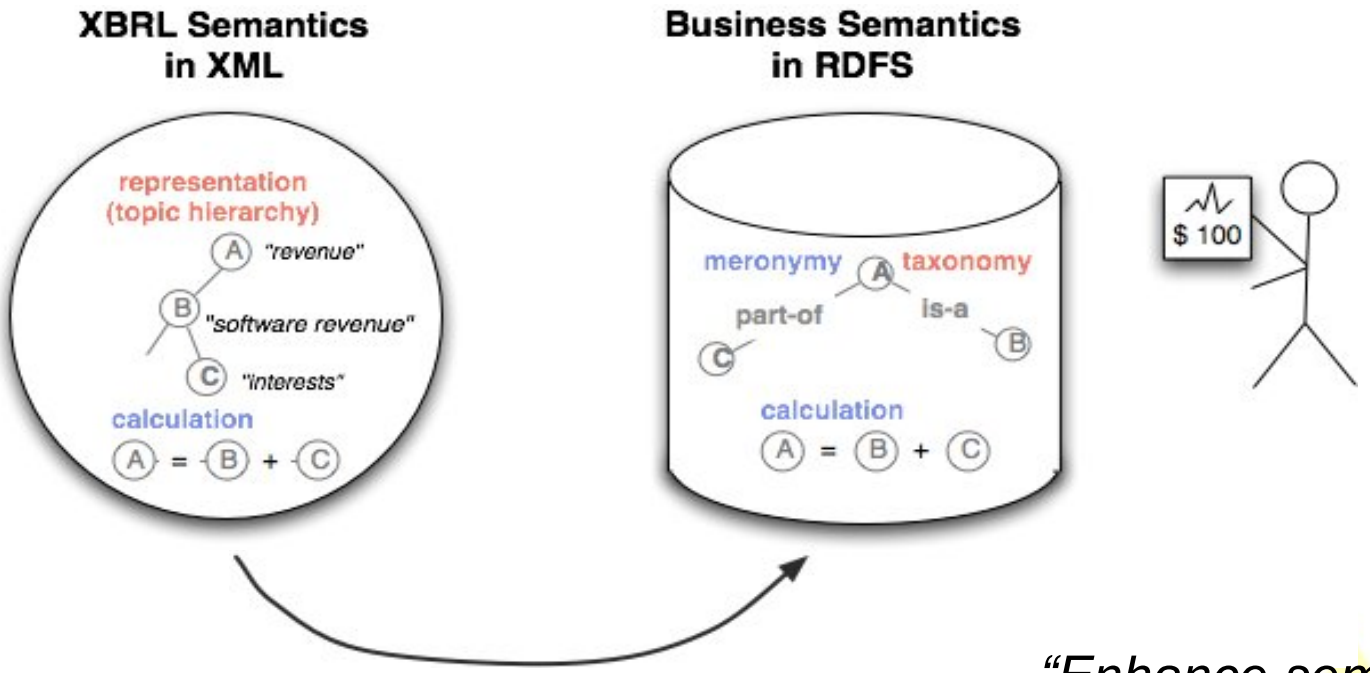
Textual Data Sources

- Structured sources:
  - pubicly available balance sheets (e.g., in German at http://www.bundesanzeiger.de/);
  - short company profiles (e.g from Business Registers, Stock Exchange web pages, etc.);
  - Wikipedia Infoboxes; and
  - XBRL instance documents (the Belgian National Bank has published on-line all the XBRL reports of Belgian companies).
- Semi-structured:
  - longer company profiles;
  - imprint information on company web pages;
  - running tickers on company information.
- Unstructured:
  - annexes to balance sheets in annual reports of companies;
  - newspapers; Specialized web pages etc.

IFRS-XBRL Example



Upgrading XBRL into an Ontology



Semantic Transformation

xbrl presentation:parent-Child	→	rdf:subClassOf
xbrl calculation: summation-Item	→	skos:PartOf

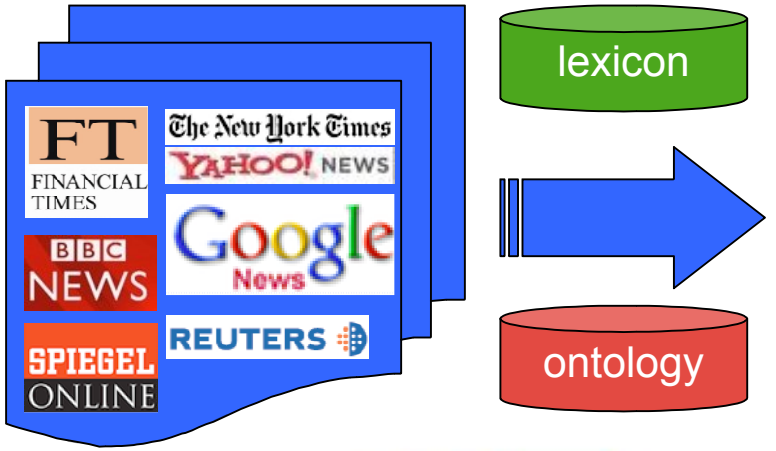
"Enhance semantics to facilitate label translation and information extraction."

XBRL-Ontology as the basis for Localization and OBIE

- **Localization of domain terms**

  ifrs:MinimumFinanceLeasePaymentsPayable
  ifrs:ProfitLossBeforeTax
  ifrs:Revenue

- **Ontology-based Information Extraction**



lexicon

ontology

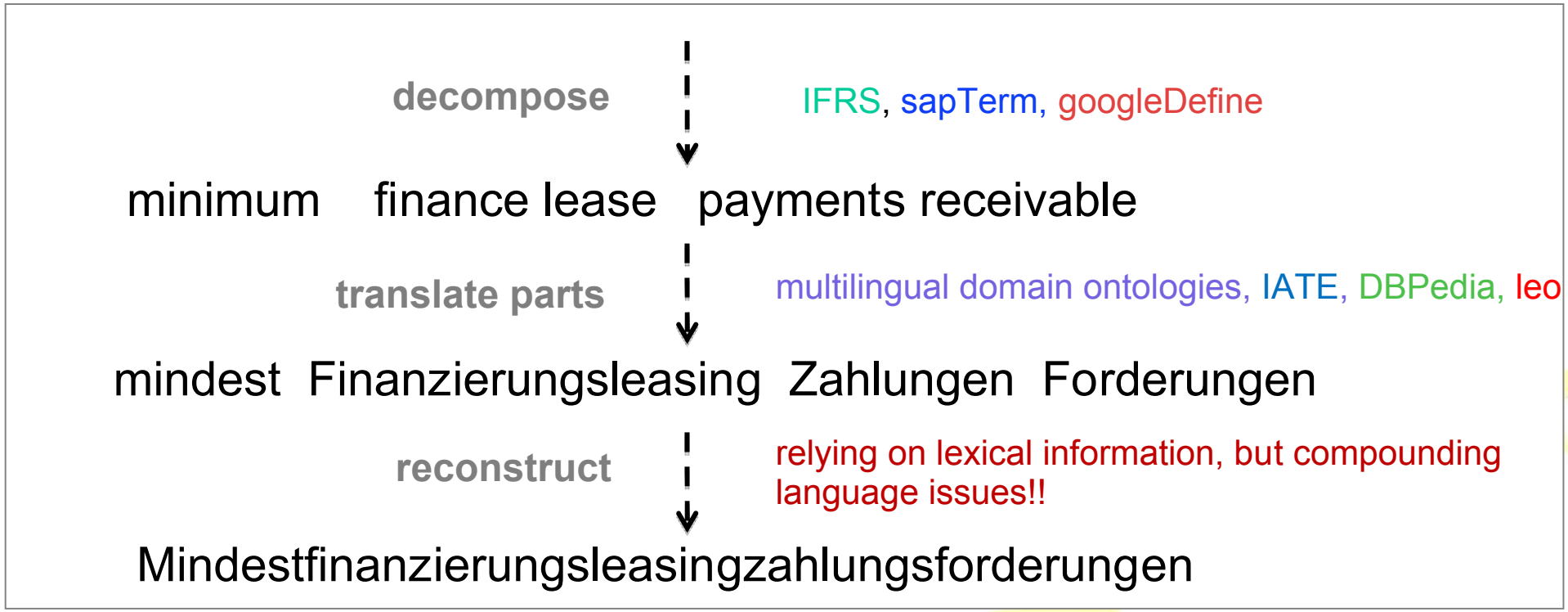
ifrs:Revenue(SAP,2894)
ifrs:Revenue(Tesco,56910)
ifrs:ProfitLossBeforeTax(SAP,676)
ifrs:ProfitLossBeforeTax(Tesco,3176)

beinformed! CITEC SAP XBRL monnet

Domain Terms Localization: “minimum finance lease payments receivable”

Google-Translate: Minimum Leasingzahlungen Forderungen

Official IFRS: Im Rahmen von Finanzierungs-Leasingverhältnissen zu erhaltende Mindestzahlungen



Google-Translate: Minimum finance lease payments receivable

Application in OBIE

semantically lifted

```
:MinimumFinanceLeasePaymentsReceivable
  rdfs:subClassOf xbrli:monetaryItemType ;
  rdfs:label "Minimum finance lease payments receivable"@en .
```

term analysis

[Minimum] [finance lease] [payments] [receivable]

linguistic analysis

payments received

receivables

SAP Annual Report 2008

..As at December 31, 2008, the future minimum lease payments expected to be received was **€16** *million…*

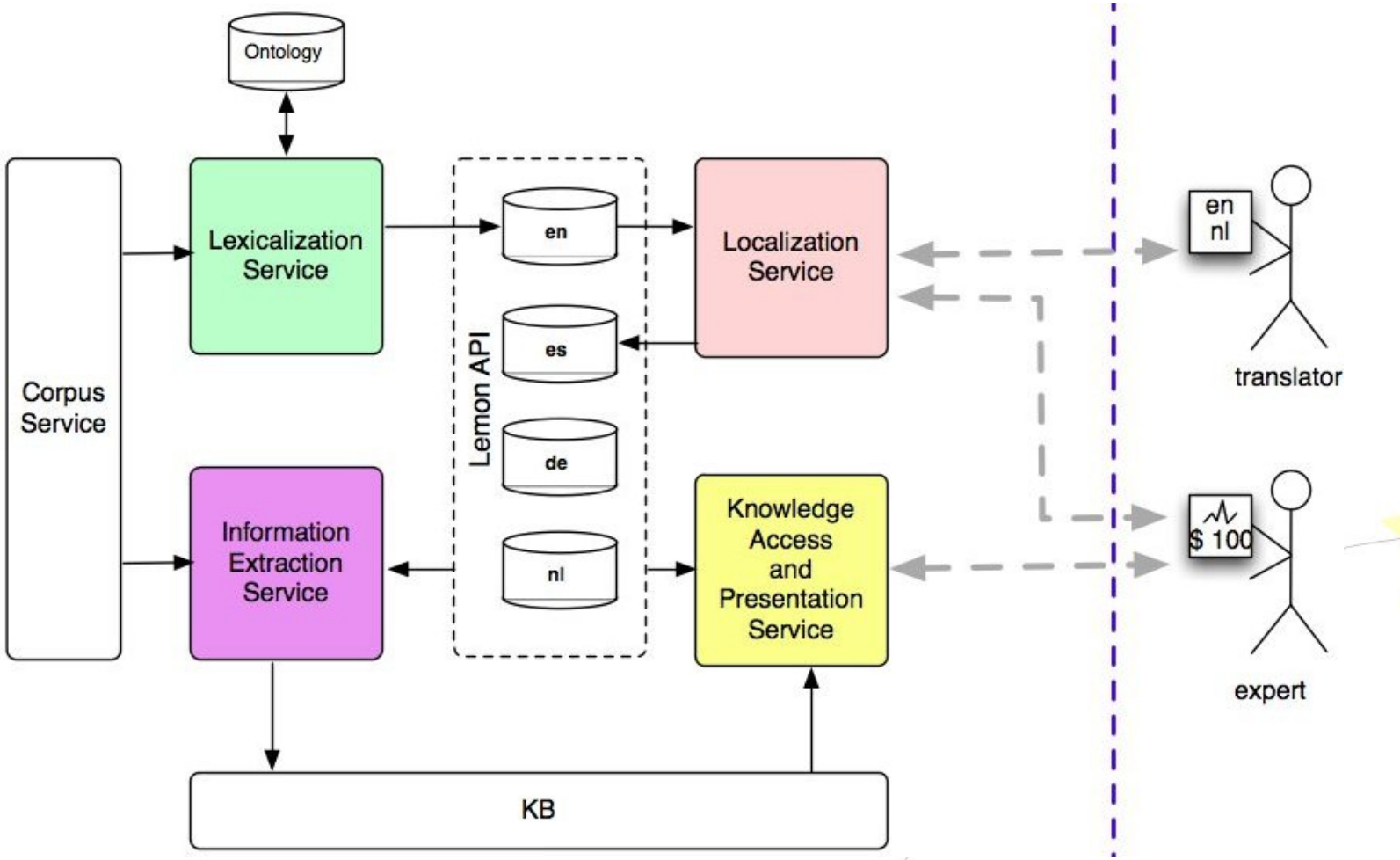
Tesco's Annual Report 2009

…The fair value of the Group's finance lease receivables at 23 February 2008 was **£5m***…*

Monnet Components to address the issues

- An *Ontology Lexicalization Service* to extract lexical expressions used in labels of ontologies, and to create an ontology lexicon automatically enriched with linguistic information
- An *Ontology Localization Tool* to create an ontology lexicon in a target language from an ontology lexicon in a source language, semi-automatically
- A *Cross-Lingual Ontology-based Information Extraction System* (CLOBIE) to leverage mulitlingual ontology lexicons to extract information from text and populate ontologies
- A *Cross-Lingual Query & Presentation System*, which uses multilingual ontology lexicons to enable quick customization of knowledge access systems to many natural languages. The knowledge being here stored in populated ontologies

Monnet Components to address the issues



Lexicalization Component

- Taxonomies and ontologies have labels that contain NL expressions (terms), which need to be linguistically enriched for supporting both, the localization of labels and the multilingual OBIE task
- For this purpose, we define **lemon** as a ontology-lexicon model
  - RDF(S)
  - Minimalist (i.e., a hub for connecting lexical and terminological description elements with ontologies)
  - Not prescriptive (i.e., uses data categories)
  - Relational semantics (i.e., uses ontologies)
  - Modular and extensible

lemon's origins

- Lexical Markup Framework (ISO 24613)
  - Standard for representing lexicons
  - XML/UML
- LexInfo, LIR
  - Represent lexical information relative to an ontology
  - OWL
- SKOS (W3C Standard)
  - Designed for Taxonomy/Vocabulary representation
  - RDF

Why lemon: RDF(S)

- RDF models are labelled directed graphs
  - (Allows for smarter representation)
- Each entry has a URI
  - Query-able on the web using standards
  - Clear ownership of data
- Linking possible between different lexica and representation standards
  - Reuse of (lexicon) available data
- Some induction possible (sub-properties, classes etc.)

Why lemon: Minimalism

- Small models (i.e., fewer links, fewer kB)
- Easier to understand
- "Open-world": Not necessary to state all facts
  - Multiple points of view

Why lemon: Semantics

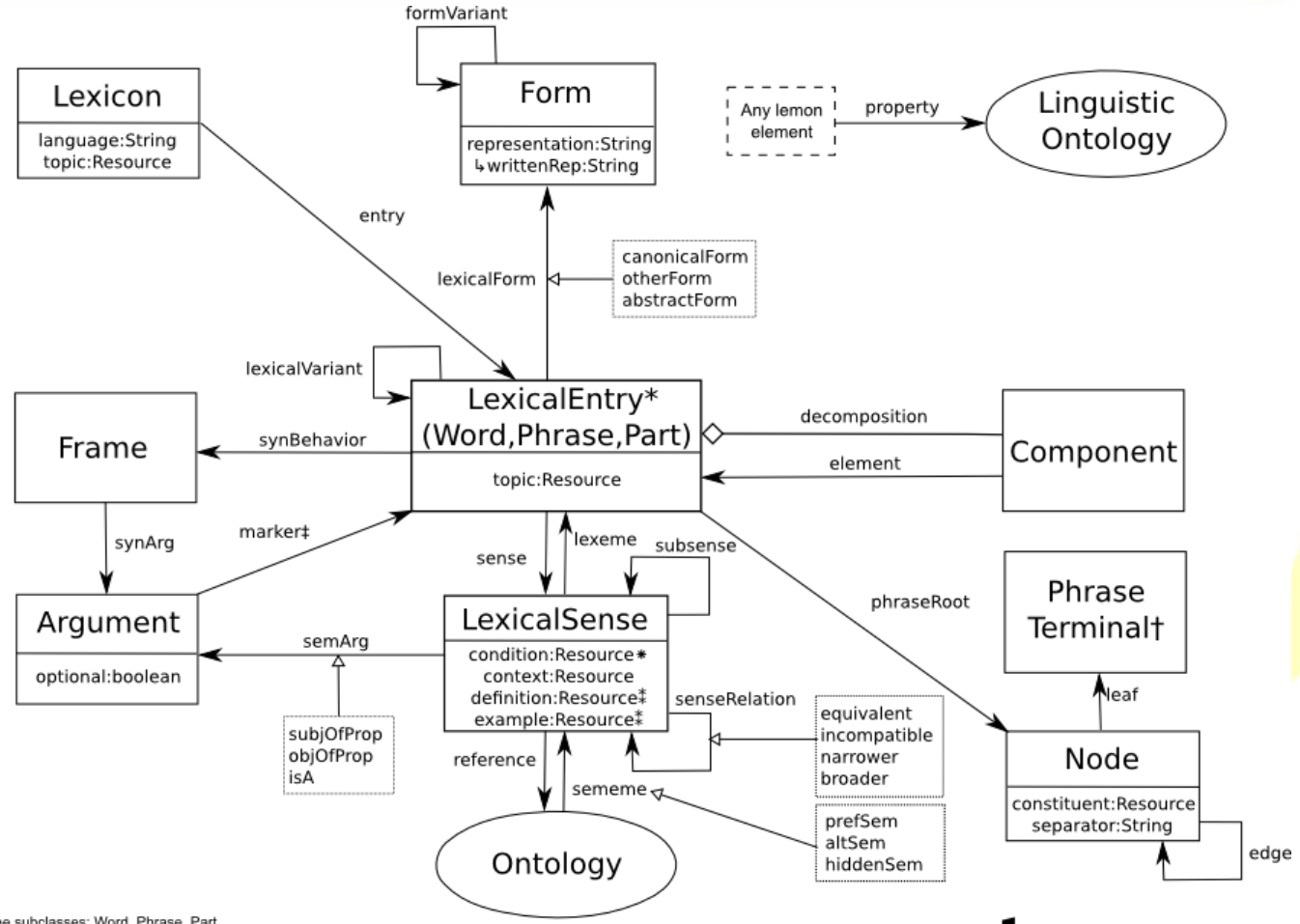
- Meaning of a word given by reference
- Reference (generally an ontology) capable of representing more complex semantic information
- Disambiguation is performed relative to the ontology
- No (traditional) word senses
  - No clashing of word senses in cross-lingual mappings

Why lemon: Modularity

- RDF(S) extensibility allows representation of
  - Subtle differences
  - Unexpected data categories
- Modularity
  - Different modules for different user requirements
  - New modules can be added later without affecting core

The model

The model provides a principled chain between the semantic representation and its linguistic realization

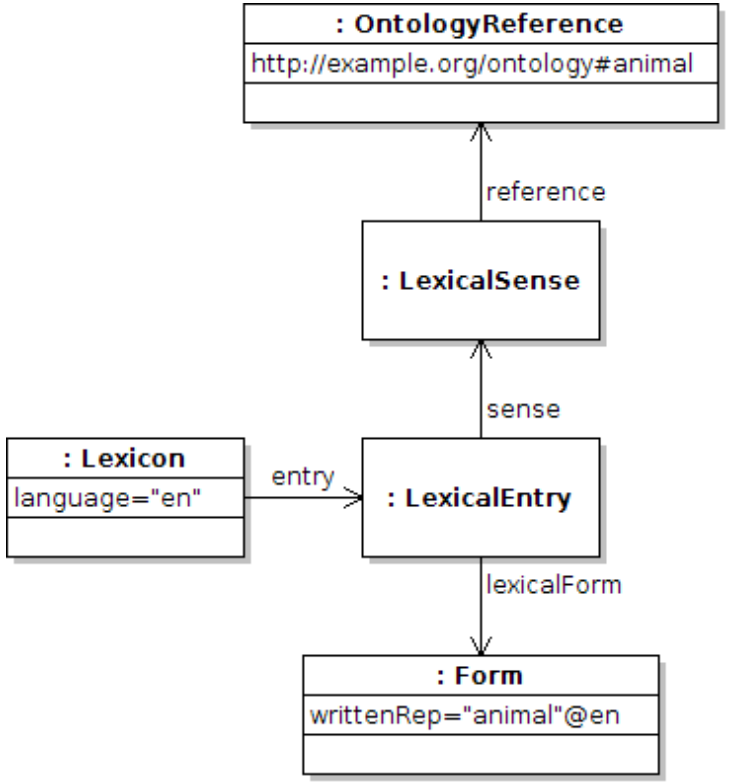


* LexicalEntry has three subclasses: Word, Phrase, Part
‡ definition and example are stated as nodes with a value
* condition has subproperties propertyDomain and propertyRange
† PhraseTerminals are Arguments or Components
‡ marker can also refer to linguistic ontology

lemon

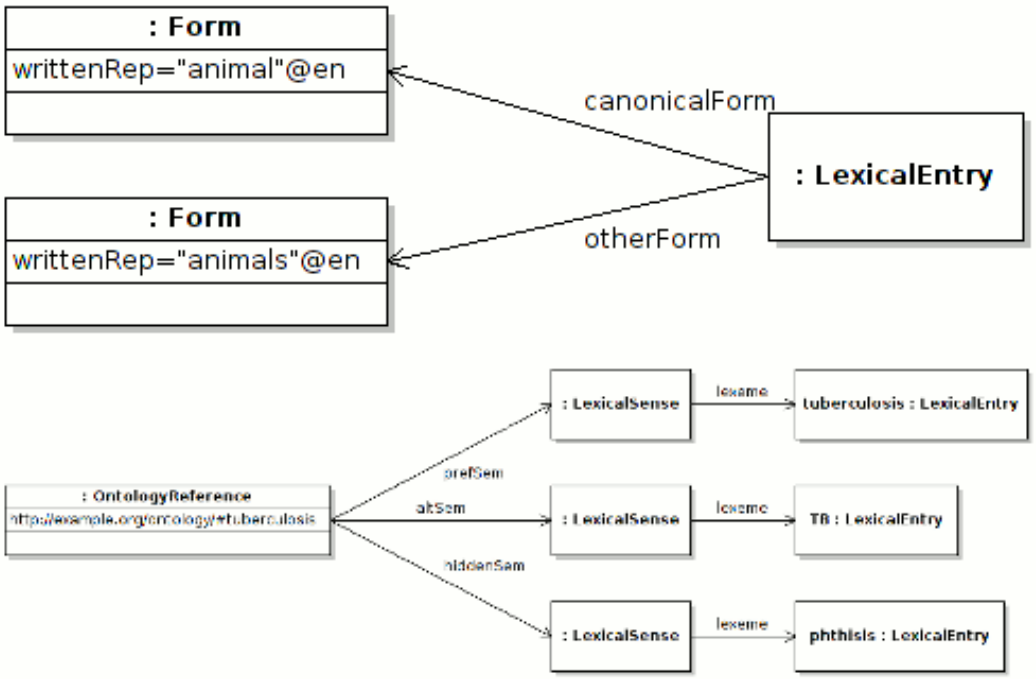
monnet

The core path



- **Lexicon:** Represents the lexicon. Marked with a single language tag (ISO-639)
- **Lexical Entry:** An entry in a lexicon. Syntax-invariant
- **Lexical Sense:** The relationship between the entry and its ontology reference.
- **Reference:** The ontology entity
- **Form:** A form of an entry. Orthography-invariant
- **Representation:** The string. IETF lang-tagged

Core path preferences



- Lexical Form
  - Canonical - The 'lemma' form of the word
  - Other - Inflected forms
  - Abstract - Forms as a stem or root

Sememe

- Preferred - The default term for given concept
- Alternative - A non-default term for concept
- Hidden - A deprecated of disallowed term for concept

Properties

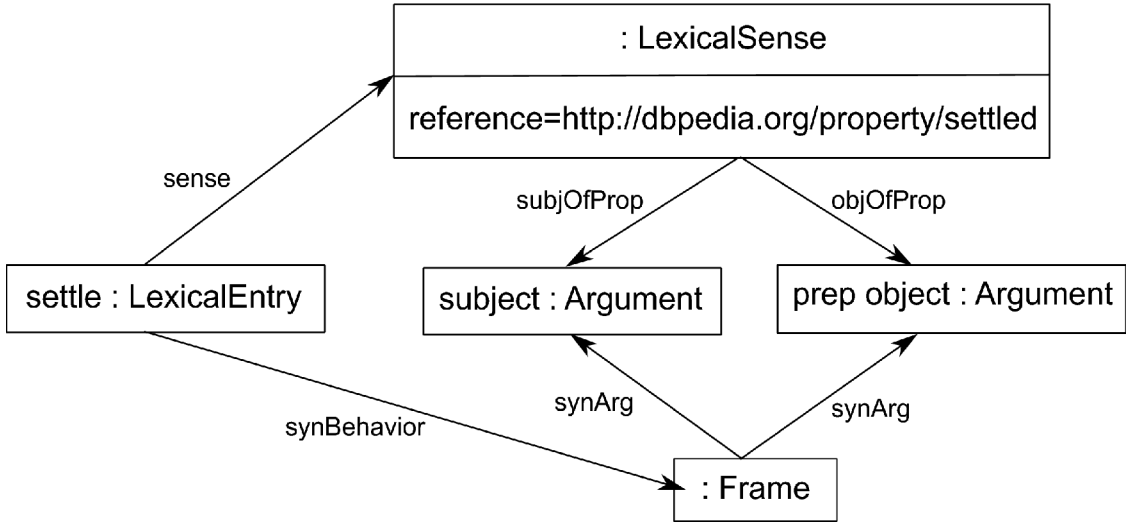
- Any element in the lexicon may have properties
- All propeties are stated as subproperties of lemon's property
- lemon does not have any such properties or values. A separate ontology is required
- e.g., ISOcat, GOLD, LexInfo II

Variation

- Forms, Entries and Senses may be marked as variants
- Again, few lemon properties, mostly use external ontology
- Mark links as sub-properties of `formVariant`,`lexicalVariant`, `senseRelation`
- Sense Relation does have sub-properties `equivalent`, `broader`, `narrower`, `incompatible`
- Set-theoretic semantics
- Sense Relation can be used to model `translationOf`

Syntax/Semantics

- Valency
  - Syntax modelled by subcategorization frames
  - Each frame has a number of arguments
  - Each argument may be optional and have a marker (adposition, particle, case)
- Mapping via sense.



- Sense can be further specified
  - Context
  - Condition: e.g., `propertyDomain/ propertyRange`
  - Definition

Senses link to sub-categorization

Relations to other Projects

- Monnet has a clear link to MultilingualWeb, proposing a semantic approach for extracting, storing and presenting information across languages.
- Monnet has a clear link to META-NET (formal collaboration agreement being discussed).
    - A possible application would be to offer a localization of the terminology in use in the ontology underlying the META-NET portal, and to apply OBIE to documents in the field of Human Language Technology

Relation to Standards

- **lemon** is drawing on existing standards in both the W3C (RDF, OWL, SKOS, ...) and the ISO framework (LMF, TMF, ISOCat, ...).
- **lemon** can contribute also to both:
  - W3C: on how to represent lexicalised ontologies and to put them into relation with other semantic resources on the web
  - ISO: on guidelines for linguistic driven lexicalization of ontologies
    - Adjusting the role played by LMF and other ISO standards on language resource management

Conclusions

- Monnet aims at providing a novel combination of semantic web solutions and localization technologies for semantically structured information
- 1st step in this direction has been the development of an ontology-lexicon model, **lemon**, for supporting the
  - Localization of ontology labels
  - Ontology-based multilingual IE
- Future work will be dedicated to ontology-based, cross-lingual information access and presentation

Thanks for your attention

Contact:

http://www.monnet-project.eu &

http://twitter.com/monnetproject