

Databases and Natural Language Processing

Challenges and contributions

Information Technologies for Business Intelligence (IT4BI)
Erasmus Mundus Master's Program
Computer Science Department
Faculty of Science and Technology
Université François Rabelais Tours, Blois, France

May, 2013

Research team behind IT4BI in Blois/Tours

People

- 4 **full professors**: Jean-Yves Antoine, Thomas Devogele, Arnaud Giacometti, Denis Maurel
- 9 **associate professors**: Béatrice Bouchou Markhoff, Nathalie Friburger, Haoyuan Li, Patrick Marcel, Nizar Messai, Verónica Peralta, Yacine Sam, Agata Savary, Arnaud Soulet
- 2 **doctors**: Samir Sebahi, Wissam Khalil
- 4 **PhD students**: Julien Aligon, Mouhamadou Saliou Dialo, Anaïs Lefeuvre, Cheikh Niang

Research team behind IT4BI in Blois/Tours

People

- 4 **full professors**: Jean-Yves Antoine, Thomas Devogele, Arnaud Giacometti, Denis Maurel
- 9 **associate professors**: Béatrice Bouchou Markhoff, Nathalie Friburger, Haoyuan Li, Patrick Marcel, Nizar Messai, Verónica Peralta, Yacine Sam, Agata Savary, Arnaud Soulet
- 2 **doctors**: Samir Sebahi, Wissam Khalil
- 4 **PhD students**: Julien Aligon, Mouhamadou Saliou Dialo, Anaïs Lefeuvre, Cheikh Niang

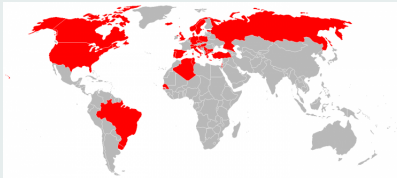
Sites

- Faculty of Science, Computer Science Department, **Blois/Tours**
- University Institute of Technology, **Blois/Tours**

Collaborations

International

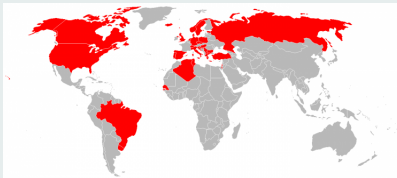
23 European countries,
Algeria, Brazil, Canada,
Israel, Russia, Senegal,
Tunisia, Uruguay, USA



Collaborations

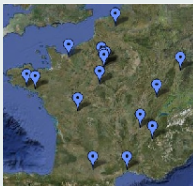
International

23 European countries,
Algeria, Brazil, Canada,
Israel, Russia, Senegal,
Tunisia, Uruguay, USA



National

universities,
research agencies,
enterprises,
associations



Scientific Domains: Bioinformatics and Big Data

Scientific Domains: Bioinformatics and Big Data

Unstructured data

- natural language processing, language resources, named entity recognition, multi-word expressions, coreference resolution, human-computer interaction, text mining

Scientific Domains: Bioinformatics and Big Data

Unstructured data

- natural language processing, language resources, named entity recognition, multi-word expressions, coreference resolution, human-computer interaction, text mining

Semi-structured data

- XML processing, semantic web, schema and document evolution, web services

Scientific Domains: Bioinformatics and Big Data

Unstructured data

- natural language processing, language resources, named entity recognition, multi-word expressions, coreference resolution, human-computer interaction, text mining

Semi-structured data

- XML processing, semantic web, schema and document evolution, web services

Structured data

- data warehouses, data mining, geographical information systems, decision support

Science and Technology Responding to Societal Needs

- IT support for **impaired** people,
- improving **health care** by decision support in medicine,
- **multilingual language technology** tools which pay greater attention to language phenomena,
- efficient use of heterogeneous and dynamic **web data**,
- facilitate the use of **decision support** techniques, in particular to **non experts**.

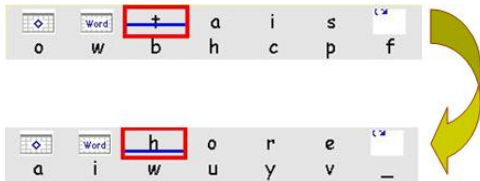
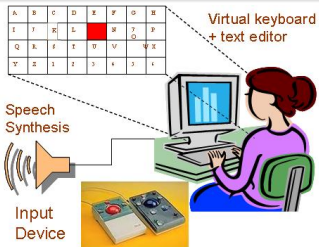
Support for the Impaired (JYA)

Societal need

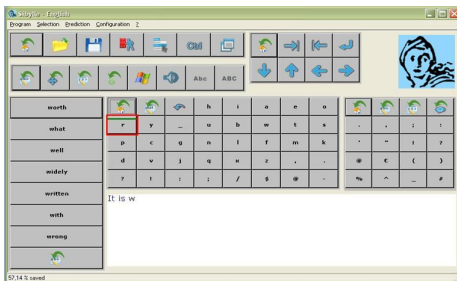
Support people with severe speech and motion impairments by **virtual keyboards**

Scientific challenge

Speeding up and facilitating message composition on a virtual keyboard



SIBYLLE system




Novelty

- Among the first methods to use semantic context for prediction (Latent Semantic Analysis).
- Improved word prediction (up to 54% keystroke saving). Highly adaptive interface.


SIBYLLE system

Impact

- journal papers: **ACM Transactions on Accessible Computing** (AWCR-pA: 32.5; SRJ h-index: 8; Harzing h-index: 12; 42 citations), **Annals of Physical and Rehabilitation Medicine** (AWCR-pA: 138.6 ; SRJ h-index: 17)
- Daily use by dozens of patients in **Kerpape Rehabilitation Center**

- Integration in the CVK open source Keyboard (**Garches hospital**)

SIBYLLE system

Impact

- journal papers: **ACM Transactions on Accessible Computing** (AWCR-pA: 32.5; SRJ h-index: 8; Harzing h-index: 12; 42 citations), **Annals of Physical and Rehabilitation Medicine** (AWCR-pA: 138.6 ; SRJ h-index: 17)
- Daily use by dozens of patients in **Kerpape Rehabilitation Center**

- Integration in the CVK open source Keyboard (**Garches hospital**)

Perspectives

- Commercializing the tool.
- Integrating the word prediction in large domotics (smart home). Semantics could stem from sensors, remote control devices, etc.

French anti-cancer campaign support (NM, ArSou)

OneDoc2[®] Prise en charge thérapeutique des cancers du sein
Référentiel CancerEst (Protocole AP-NP K 07663)

Hôpital Tenon
4, rue de la Chine, 75020 Paris

Version 5.5.3 Intégrale - 26 Jun 2009

RCP - Participants - Patients | Référentiel | Critères - Traitements - Essais

mode automatique debug

NIP: 12345678 Nom: Dupont Prénom: Céilia DDN: 21/12/1956 Âge: 53 Responsable: Dr Martin Dossier Fermer

Décisions: gauche Mastectomie à gauche + Ganglion sentinelle à gauche. droite Enregistrer

Traitement du cancer du sein non métastatique. (v2.19)

Tableau clinique

- Cancer avec tumeur mammaire = Oui
- Type de la lésion mammaire = Carcinome invasif
- Foyer invasif unique = Oui
- Présence d'un foyer in situ = Non
- Traitement néo-adijuvant déjà réalisé = Non
- Intervention chirurgicale déjà réalisée = Non
- Tumeur accessible à un traitement chirurgical = Oui
- Classe N supérieure ou égale à 2 = Non
- Récidive locale = Non
- Patient opérable = Oui
- Contre-indication à la tumorectomie = Non
- Taille de la lésion invasive = Inférieure ou égale à 2 cm
- Contre-indication au ganglion sentinelle = Non

Résumé clinique :
 Patiente de 53 ans. Carcinome invasif. Lésion invasive de moins de 2 cm.

Recommandations thérapeutiques du référentiel CancerEst pour le sein gauche

- Tumorectomie à gauche + Ganglion sentinelle à gauche.

Décision de RCP: Mastectomie à gauche + Ganglion sentinelle à gauche.

Il existe une condition spécifique non prise en compte dans la caractérisation du profil de la patiente (p. ex. femme enceinte, mutation constitutionnelle délétère identifiée) qui justifie que la décision de RCP ne suive pas le référentiel local.

French anti-cancer campaign support

Societal need

- Improve patient health care by identifying **lacks in clinical guidelines**.
- Increase the efficiency of **decision support in medicine**.

Scientific challenge

Characterizing clinicians' **wrong decisions**

French anti-cancer campaign support

Societal need

- Improve patient health care by identifying **lacks in clinical guidelines**.
- Increase the efficiency of **decision support in medicine**.

Scientific challenge

Characterizing clinicians' **wrong decisions**

Techniques

General-purpose mining approaches adapted to clinical data:

- FCA
- contrast mining

Customized data mining

Novelty

Better assessment of the **unexpectedness** of clinician's decisions

Customized data mining

Novelty

Better assessment of the **unexpectedness** of clinician's decisions

Impact

- conference papers: **AIME** (CORE: A), **AMIA** (h-index: 47), **MEDINFO** (CORE: B),
- results used by **Assistance Publique – Hôpitaux de Paris** (federation of Parisian hospitals) for improving clinical guidelines and recommendation systems.

Customized data mining

Novelty

Better assessment of the **unexpectedness** of clinician's decisions

Impact

- conference papers: **AIME** (CORE: A), **AMIA** (h-index: 47), **MEDINFO** (CORE: B),
- results used by **Assistance Publique – Hôpitaux de Paris** (federation of Parisian hospitals) for improving clinical guidelines and recommendation systems.

Perspectives

- Taking the semantics of data into account.

Linguistic Precision and Multilingualism (JYA, NF, DM, AgSa, ArSou)

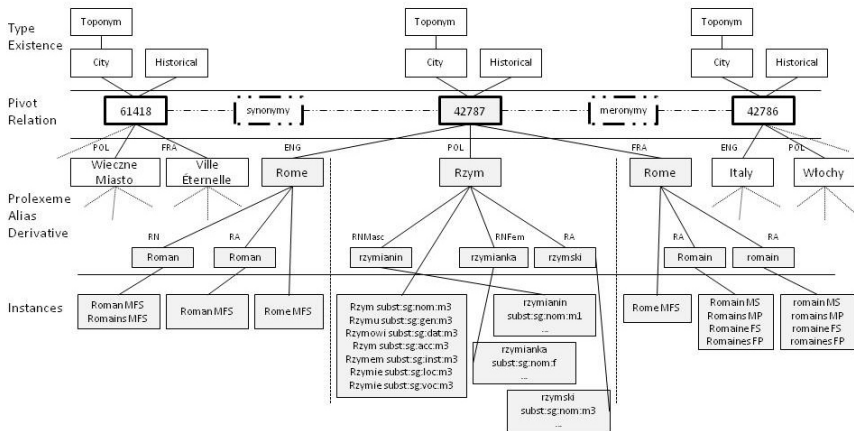
Societal needs

Provide language technology tools which pay greater attention to **language phenomena** (variation, composition etc.). Support **multilingualism**.

Scientific challenges

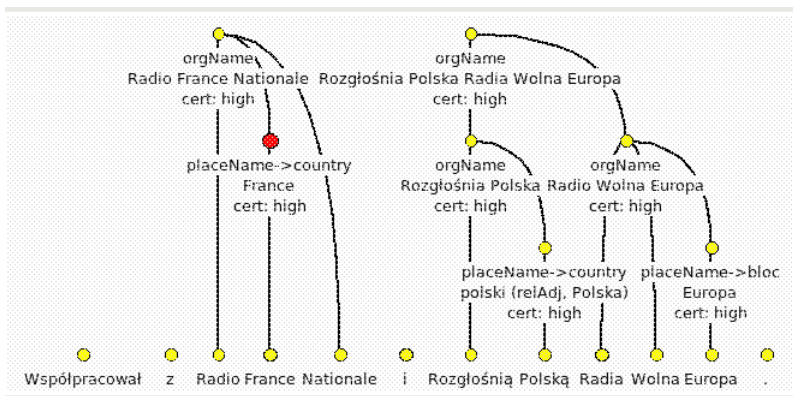
- Linguistic precision in modeling **Named Entities** (NEs) and **Multi-Word Expressions** (MWEs), in particular in morphologically-rich languages.
- Named Entity Recognition (NER) with a **fine-grained typology** and **nested structures**, possibly in a **noisy input** (dialogues)

ProlexBase – multilingual ontology of proper names



200,000 proper names in FR, PL and EN; morphological variants; manual validation

Nested Named Entity Recognition (CassEN, MXS, Nerf)



Linguistic Precision and Multilingualism

Novelty

- Modeling proper names as **ontology concepts**
- Large, fine-grained, **manually validated** language resources
- **Conflating** morpho-syntactic **variants** within the same framework
- First NER system recognizing separately the left and the right frontiers of NEs (**MXS**).
- First Polish NER system recognizing nested structures (**Nerf**).

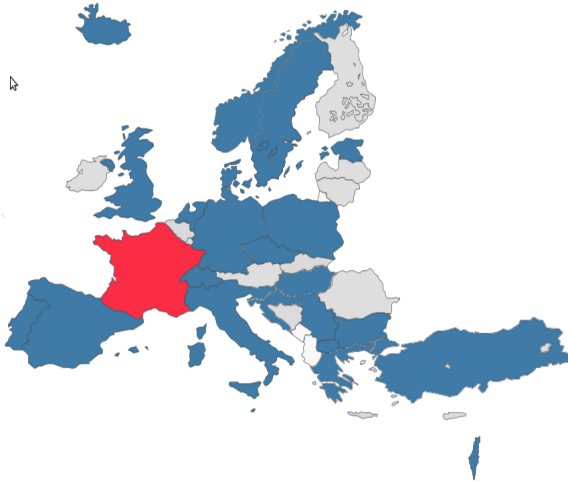
Techniques

- Ontology population from open sources (Wikipedia, GeoNames),
- Finite-state transducers with unification, transducer cascades,
- Conditional Random Fields.

NLP Impact

- journal papers: **Theoretical Computer Science**; H-index: 63, SJR: 1,175, **Control & Cybernetics**; h-index: 22; SJR: 0.35, IF: 0.38,
- **editor** of **TAL** journal special issue on named entities;
- integration of tools and resources in corpus and dictionary processors: **Unitex**, **Leximir**, **Toposław**; users in **France**, **Greece**, **Poland**, and **Serbia**,
- international conference **CIAA/FSMNLP 2011** organized in Blois,
- **European expert** for the FP7-SME-2013 call,
- French and Polish national (**ANR EPAC**, **NKJP**) and regional projects (**Variling**, **ANCOR**, **Renom**, **ERDF NEKST**),
- coordinator of the **FP7-COST-IC1207 PARSEME** action.

IC1207 PARSEME: Parsing and Multi-Word Expressions



24 countries,
80 members,
23 languages,
experts from
Brazil and USA
meetings,
missions (<6 months),
workshops,
training schools,
impact on ESRs

NLP Perspectives

- Integrating MWEs in **parsing**
- Integrating fine-grained language data into **Linked Data** (DBPedia, YAGO...)
- NER with **semantic grounding** (attaching NER mentions to ontology nodes)

Web Data Support (BBM, DM, ChN, YS, AgSa)

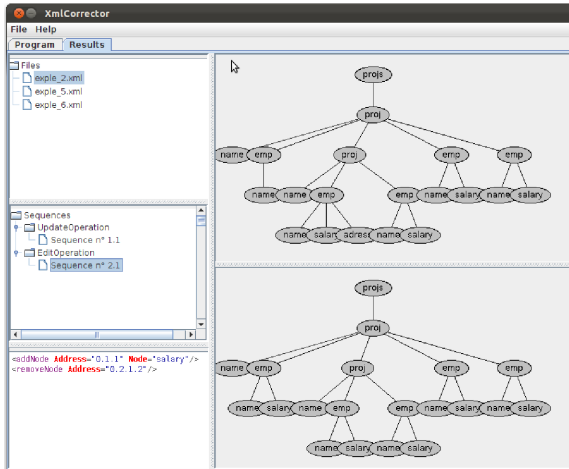
Societal needs

Efficient use of **heterogeneous** and **dynamic** web data.

Scientific challenges

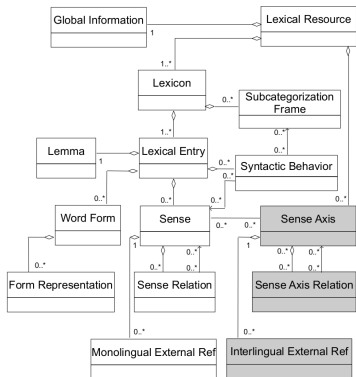
- Management of **dynamic XML documents and schemas**,
- Making heterogeneous ontologies **interoperable** with a minimum human intervention/expertise.

XMLCorrector – correcting an XML doc. wrt. a DTD



Outputs all correction trees whose distance from the initial tree is no higher than a given threshold

ProLMF: a multilingual data exchange XML standard for proper names

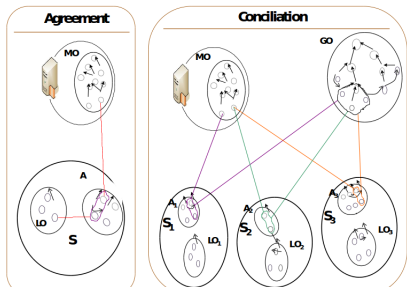
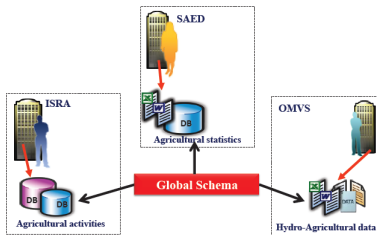


```

<?xml version='1.0' encoding='UTF-8' ?>
<LexicalResource>
  <GlobalInformation languageCoding="ISO 639" scriptCoding="ISO
15924" characterCoding="UTF-8" entrySource="Prolexbase"
resourceName="Prolmf" version="1.2"/>
  <Lexicon languageIdentifier="fra" script="latn">
    <LexicalEntry partOfSpeech="noun">
      <Lemma>Italie</Lemma>
      <WordForm grammaticalGender="feminine"
grammaticalNumber="singular">Italie</WordForm>
      <Sense idSense="P42786" refSenseAxis="42786"
termProvenance="fullForm" frequency="commonlyUsed"
label="properName">
        <SyntacticBehaviour
refSubcategorizationFrame="CO4"/>
        <SyntacticBehaviour
refSubcategorizationFrame="CO7"/>
      </Sense>
    </LexicalEntry>
  </Lexicon>
</LexicalResource>
  
```

[...]

Data Integration



- Algorithm for automatically building a **global ontology** from several local ontologies and a **mediator ontology**
- System for **querying** the global ontology

Web Data Support

Novelty

- The first **full-fledged** algorithm solving the problem of the **document-to-schema** correction
- The first framework for a **grammarware incremental validation** of data **integrity constraints**.
- An international **standard** for representating proper names.
- **Minimizing human intervention** by mediator ontology use.
- **Incremental** data integration (adding a new ontology is easy).

Web Data Support

Novelty

- The first **full-fledged** algorithm solving the problem of the **document-to-schema** correction
- The first framework for a **grammarware incremental validation** of data **integrity constraints**.
- An international **standard** for representating proper names.
- **Minimizing human intervention** by mediator ontology use.
- **Incremental** data integration (adding a new ontology is easy).

Techniques

- Finite state automata, tree automata, dynamic programming, attribute grammar,
- String-to-string & tree-to-language edit distance,

Web Data Support

Impact

- Papers: **The Computer Journal** (since 1962, A+ in CORE, 5-year IF: 0.943), **Transactions on Large-Scale Data and Knowledge-Centered Systems** (since 2011); **IJARAS** (since 2010); chapter in a reference book on **LMF**,
- National project: **ANR CODEX**, collaboration with **Senegal**.

Web Data Support

Impact

- Papers: **The Computer Journal** (since 1962, A+ in CORE, 5-year IF: 0.943), **Transactions on Large-Scale Data and Knowledge-Centered Systems** (since 2011); **IJARAS** (since 2010); chapter in a reference book on **LMF**,
- National project: **ANR CODEX**, collaboration with **Senegal**.

Perspectives

- **Taxonomy** of the existing XML correction algorithms.
- Applications in the **humanities**:
 - **SIC-Senegal**: geographic data on the Senegal river,
 - **PERSONAE**: prosopography of the Renaissance period,
 - **BIBLIMOS**: ancient manuscripts of the Western Sahara.

Decision Support (JA, TD, MSD, AG, HL, PM, VP, ArSou)

Societal needs

Facilitate the use of Decision Support techniques, in particular to **non experts**.

Decision Support (JA, TD, MSD, AG, HL, PM, VP, ArSou)

Societal needs

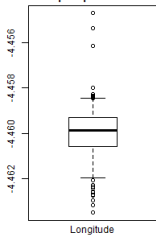
Facilitate the use of Decision Support techniques, in particular to **non experts**.

Scientific challenges

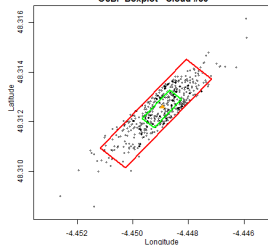
- Improving **outlier detection**, better **data visualization** and analysis,
- **Benchmarking OLAP sessions** for effectiveness (what is a “useful” OLAP session?),
- Providing **user profiles** that are **sound**, **concise**, and **easy to interpret**,
- **High-level modeling** of data mining tasks.

Contributions

Boxplot percentiles



OSBP Boxplot - Cloud #80



- 3-dimensional **BoxPlot**,
- Recommendation approaches for **OLAP session detection and comparison**,
- An approach to build **user profiles from interesting sets of contextual preferences**,
- A declarative **high-level language** for modeling data mining tasks.

Decision Support

Novelty

- The first 3-dimensional **spacio-temporal BoxPlot**.
- **Pioneering work** on personalizing OLAP queries and on exploiting OLAP logs.
- The first method which offers a **user-understandable profile** (others are “black boxes”). The user can modify and complete his/her profile easily.
- The first method enabling **reasoning on queries**.

Decision Support

Novelty

- The first 3-dimensional **spacio-temporal BoxPlot**.
- **Pioneering work** on personalizing OLAP queries and on exploiting OLAP logs.
- The first method which offers a **user-understandable profile** (others are “black boxes”). The user can modify and complete his/her profile easily.
- The first method enabling **reasoning on queries**.

Techniques

- Collaborative filtering,
- Symbolic semi-supervised learning,
- Relational algebra.

Impact

- A chapter in **Mobility Data: Modeling, Management and Understanding**, Cambridge press,
- Member of European **COST IC0903 MOVE** Action,
- Journal and conference papers: **KAIS** (IF=2.225); **DASFAA** (A in CORE); **DAWAK'2012** (B in CORE),
- **Invited speaker** in the **VLDB** (CORE: A) workshop,
- Collaborations: Bologna, Barcelona, Quebec, Uberlandia (Brazil), Saint Louis (Senegal).

Perspectives

- Full **integration of personalization** into OLAP systems,
- **Scalability**: taking into account the characteristics of the user preferences,
- Use of profiles for **collaborative filtering**,
- Extracting **temporal preferences**,
- Implementation the modeling language within **hadoop framework**; cost model.

Call for Collaboration

Research contracts

- **Research internships** in semester 4 (6 months),
- **PhD theses**,
- **European networks** (MOVE, PARSEME),
- **Development** contracts,
- **Funding**: permanent budget, current projects, pending national and European project proposals.

Call for Collaboration

Research contracts

- **Research internships** in semester 4 (6 months),
- **PhD theses**,
- **European networks** (MOVE, PARSEME),
- **Development** contracts,
- **Funding**: permanent budget, current projects, pending national and European project proposals.

Research and R&D collaborations

- International and national **projects** (European, AUF, CAPES-COFECUB, Egide, ANR, ...),
- **Missions** (Eiffel PhD grants, post-doctoral grants, Marie-Curie programs, ...).