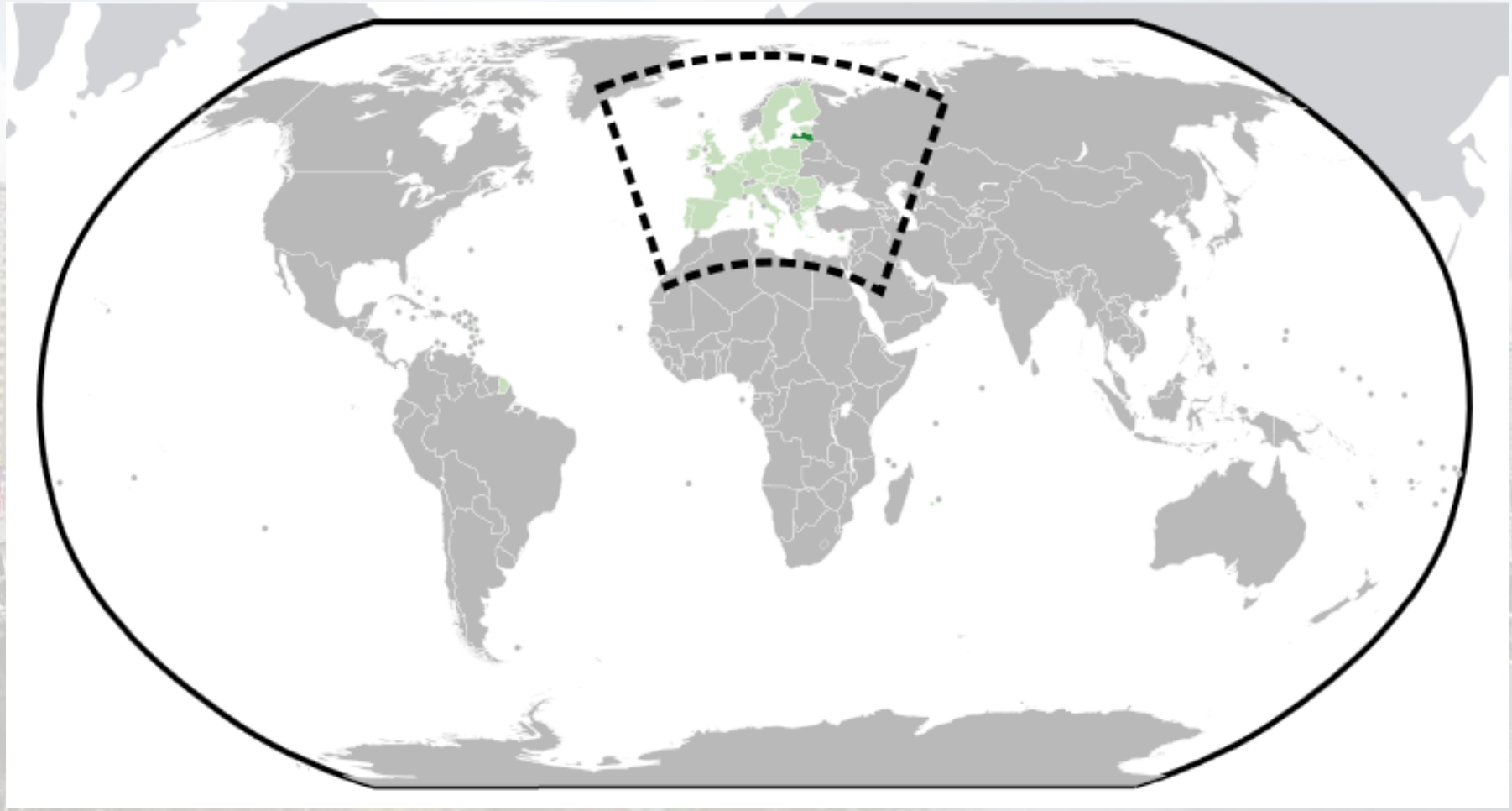
# Extending the Use of Web-Based Terminology Services

Tatiana Gornostay

Tilde, Latvia

Multilingual Web Workshop, Dublin, Ireland
June 11, 2012

# LATVIA



# TILDE tilde.com

- Translation and Localization services
  - Latvian, Lithuanian, Estonian
- Terminology development and management
  - EuroTermBank: >2 mil terms, >25 languages
- Language Technologies and Resources
  - Small languages
- 3 offices
  - Riga (Latvia, headquarters)
  - Vilnius (Lithuania)
  - Tallinn (Estonia)
- >100 employees
  - 4 PhDs and 3 PhD candidates

# European cooperation

# Terminology

- Terminology is everywhere
  - visiting a doctor
  - building a house
  - buying a car
- We come across with terms every day

# Terminology

- Terminology matters
  - efficient and precise communication
    - academia
    - industry
    - government

Society

# Terminology

- Terminology is a language
  - Language for Specific (professional) Purposes (LSP)
  - multilingual consolidated and harmonized terminology is already being utilized as data by human users
    - language workers
      - translators, terminologists, technical writers, editors, etc.
  - now it is being developed as a web-based service for machines as users
    - systems
      - machine translation, indexing, search, annotation, etc.

# Challenges

- creation, consistency, extraction
  - according to recent surveys, 84% professionals select terms from documents manually
  - acquisition
    = term identification in a text
  - recognition
    = term comparison with existing resources
- consolidation & harmonization
- sharing & interoperability

- MT domain adaptation
- concept formalization
- data annotation, indexing and search, etc.

Terminology is on the cusp between

semantic and language technologies

Terminology is bridging the three communities

Linked Open Data

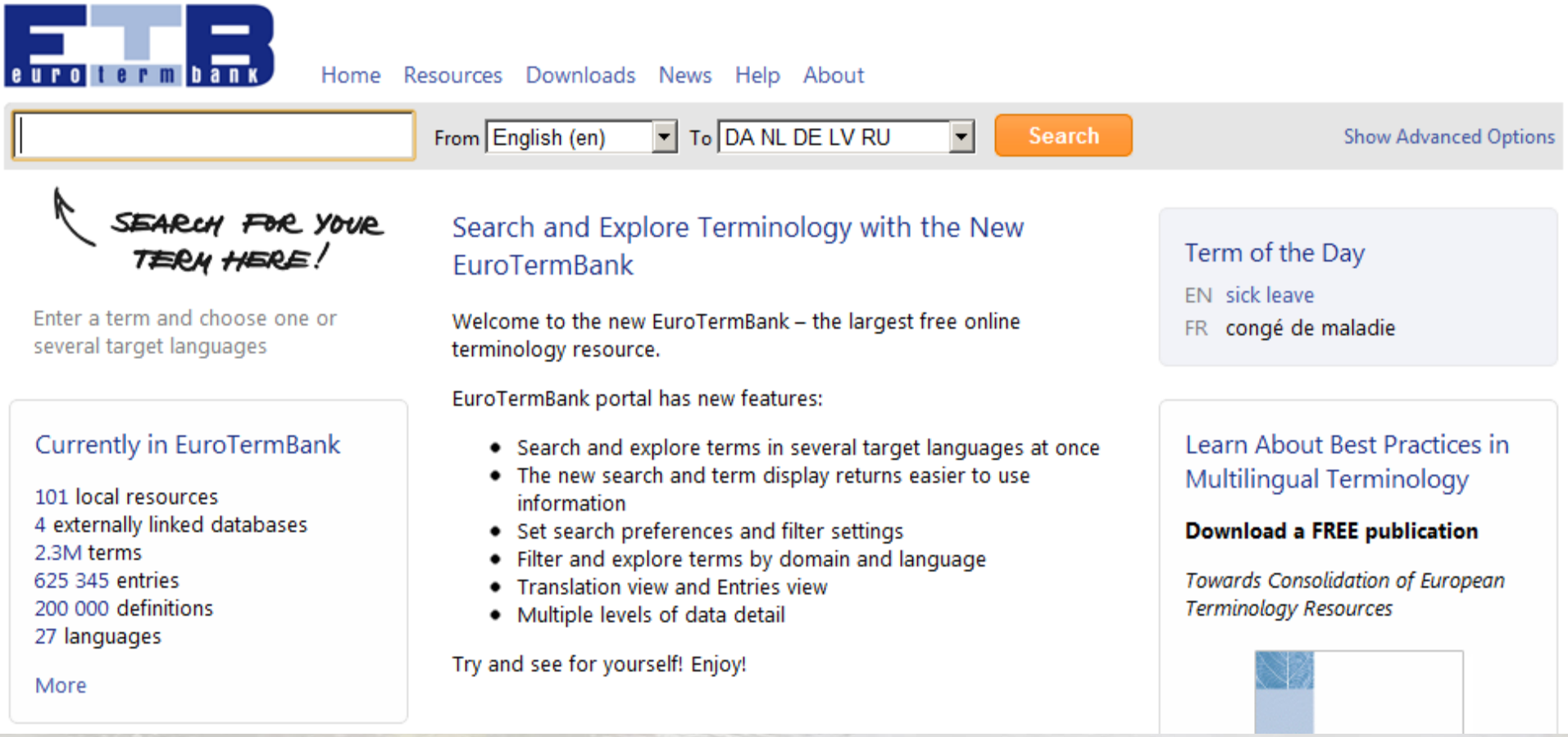
Multilingual Web

Multilingual Language Technologies, i.e. NLP

# Tilde's best practices & use cases

# EuroTermBank

- www.eurotermbank.eu



# EuroTermBank

- www.eurotermbank.eu
  - MS Word
  - memoQ
  - Microsoft multilingual terminology
  - IATE
  - Open Terminology Platform
  - sharing & exchange terminology in META-SHARE
  - will be used in terminology services both for human & machines as users

# ACCURAT & TTC

Analysis and Evaluation

of Comparable Corpora

for Under-Resourced Areas

of Machine Translation

Terminology Extraction

Translation Tools

Comparable Corpora

# ACCURAT & TTC

- Comparable corpora
- Reference term lists and annotated texts
- Rule sets for term variant recognition and mapping
- Toolkit for multi-level alignment and information extraction from comparable
- Neo-classical multi-word term detection program
- TTC TermSuite

# TaaS

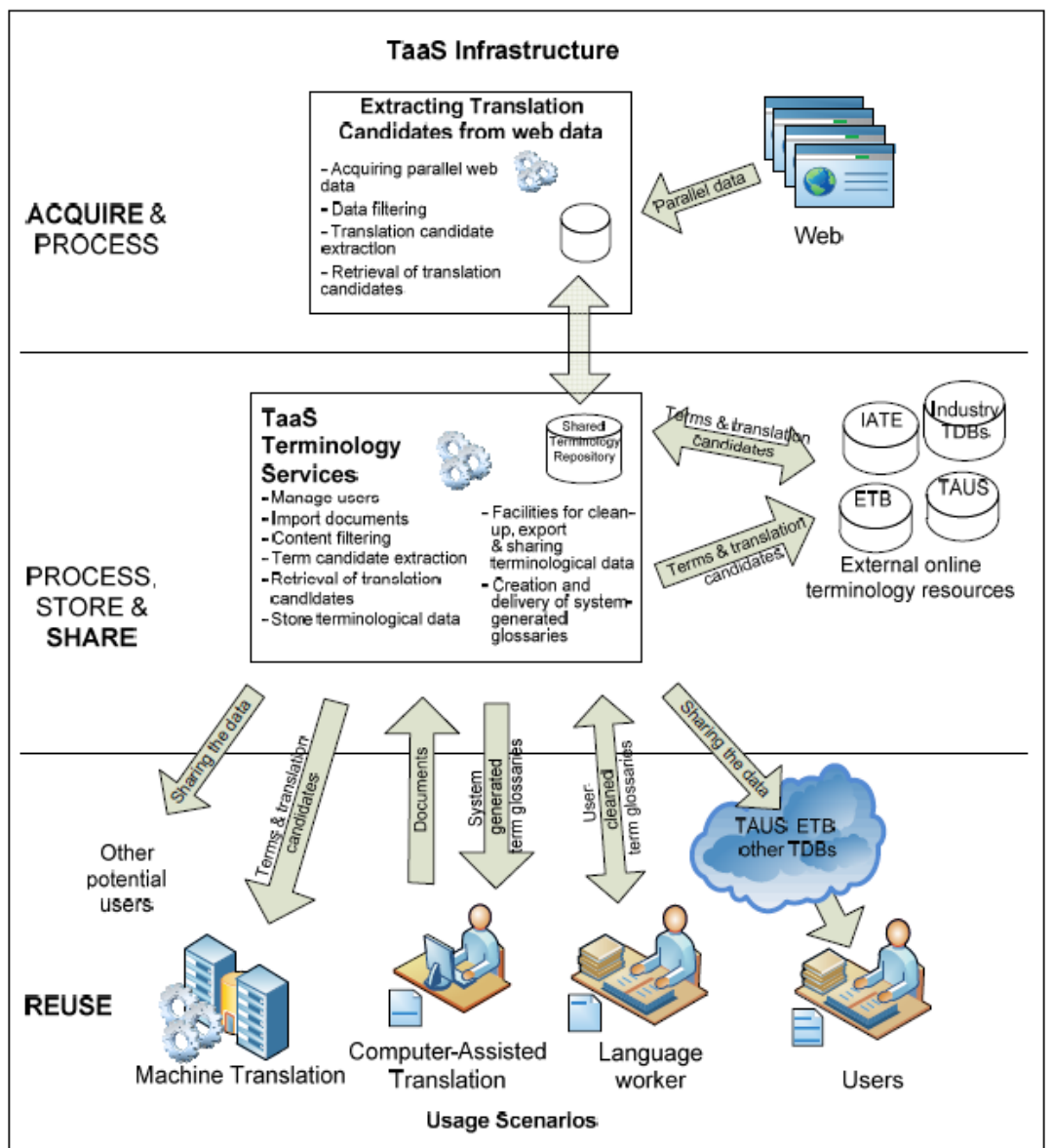
Terminology as a service

a cloud-based platform

for acquiring, cleaning up, sharing, and reusing

multilingual terminological data

# TaaS basic services



# LetsMT!

Let's MT! beta

Systems Corpora Translate Tools About Sign up Login

## Build your own machine translation system!

With LetsMT you can easily build and run your own custom machine translation systems. Simply upload your own corpora and/or choose to use any of the publicly available corpora. Train your systems and use them for all your translation needs.

### EASY

- ✔ Frees you from hardware and software infrastructure
- ✔ Store corpora and engines in one place
- ✔ Access to many free public texts and translation engines
- ✔ Instantly increase productivity with CAT plug-in for localization process and web browser widget

### IS LETS MT FOR YOU?

- ✔ Localization & translation service providers
- ✔ Holders of linguistic resources
- ✔ Companies with a need to translate large amounts of information
- ✔ Those who don't want to trust their resources to public systems

### THE BASICS


Translate

Translate now using available systems

Build

Build your own machine translation system

Store and Share

Upload, organize, store, and share your corpora

# SMT adaptation use case

SMT system adaptation to narrow domain

- automotive manufacturing

We had:

- limited amount of in-domain parallel texts from a client
- no in-domain texts in the target language
- extracted terms from parallel texts
- additional comparable texts collected from the web
- bilingual in-domain terms tagged and mapped automatically in the collected texts

We got:

- **32%** increase in BLEU against a broad domain system

Terminology is on the cusp between
semantic and language technologies

Terminology bridges the three communities
LOD, MW & NLP

Terminology has the potential to vastly enhance
the degree of automation for LOD

Terminology facilitates the creation
of multilingual ontologies, taxonomies, etc.

Terminology helps to automate the creation
of multilingual & cross-lingual metadata

# Thank you for your attention and time!

www.tilde.com

tatiana.gornostay@tilde.lv

The research within the projects LetsMT!, ACCURAT, TTC, TaaS leading to these results has received funding from the European Commission ICT Policy Support Programme and FP7 Programme