The 15th International Symposium on Intelligent Distributed Computing
Sep 14-16, 2022, Bremen, Germany

Towards an Online Multilingual Tool for Automated Conceptual Database Design

Drazen Brdjanin, Mladen Grumic, Goran Banjac, Milan Miscevic, Igor Dujlovic, Aleksandar Kelec, Nikola Obradovic, Danijela Banjac, Dragana Volas, Slavko Maric

M-lab Research Group @ Faculty of Electrical Engineering
University of Banja Luka, Bosnia & Herzegovina

Presentation Outline

- Research context and motivation
- Research objectives and contributions
- Approach outline
- Implemented tool
- Illustrative examples
- Conclusion and future work

Research Context & Motivation

Model-driven Software Engineering Laboratory

Faculty of Electrical Engineering • University of Banja Luka

http://m-lab.etf.unibl.org

M-lab research focus:
Automatic database design based on sources of different nature (models, text, speech, …)

Main achievements:

http://m-lab.etf.unibl.org:8080/amadeos

AMADEOS

the first online web-based tool for automatic derivation of conceptual database models from collections of differently represented and differently serialized business process models

http://m-lab.etf.unibl.org:8080/Textodata

TexToData

the first online multilingual web-based tool for automatic derivation of conceptual database models from natural language text

Research Context & Motivation

Limitations of Model-driven approaches

- Lack of semantic capacity of BPMs for automatic generation of complete data models:
  - Ability to automatic generation of a highly complete data model structure (~80-100% of entity types and their relationships)
  - Ability to automatic generation of modest percentage of attributes in entity types
- Necessity to combine models and some other sources (possible textual specifications)

Limitations of Text-based approaches

- A lot of NLP research since Chen's eleven rules (1983) for translation of NL text into E-R, but:
  - There is still no tool able to automatically convert NL text into the corresponding CDM
  - The existing NLP-based tools typically support one single source NL (mainly English, Spanish or German) and do not provide multilingual support
  - There is no online NLP-based solution/service enabling the automated CDM design

Research objectives

Define an approach and implement an online tool/service able to automatically derive CDM from textual specification that may be specified in different NLs

(in order to be able to combine CDMs derived from different sources)

Research Objectives & Contributions

Research objectives

- **Define an approach and implement an online tool / service able to automatically derive CDM from textual specification that may be specified in different NLs**

  (in order to be able to combine CDMs derived from different sources)

Research Contributions

- **Approach**
  - **An orchestration of (publicly) available services for text translation, NLP, layouting, ...**
- **Implemented tool – TexToData**
  - online web-based tool (set of online services)
  - multilingual support (extraction of CDM from textual specifications in different source NLs)
  - automatic layouting and UML-bsaed representation of generated CDM (editing and formatting functionalities, XMI-export to support model portability, ...)

http://m-lab.etf.unibl.org:8080/Textodata

Approach outline

What do we want?

- Define an approach and implement an **online web-based tool/service** able to **automatically derive CDM from textual specification** that may be **specified in different NLs**
- Maximize **employment of existing publicly available free-of-charge services**

What do we need?

- NLP service(s) to analyze source textual specifications
- CDM generator service
- Diagram layouting & editing services

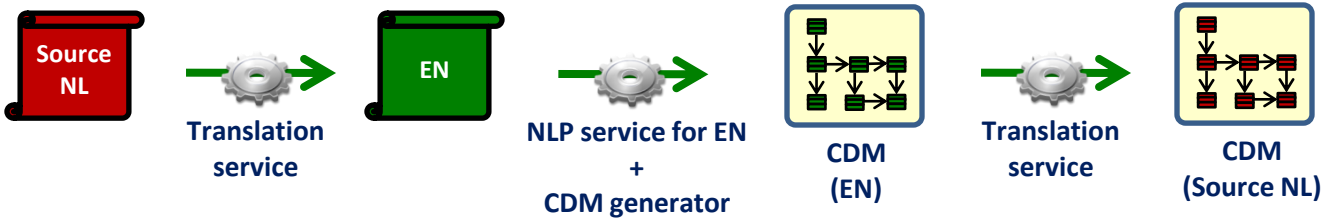
What do we have?

- online NLP services for analysis of English text (TextRazor, Bitext, …)
- online translation services (Google Translate, Yandex, …)

What is missing?

- online NLP services for analysis of other NLs

The main idea

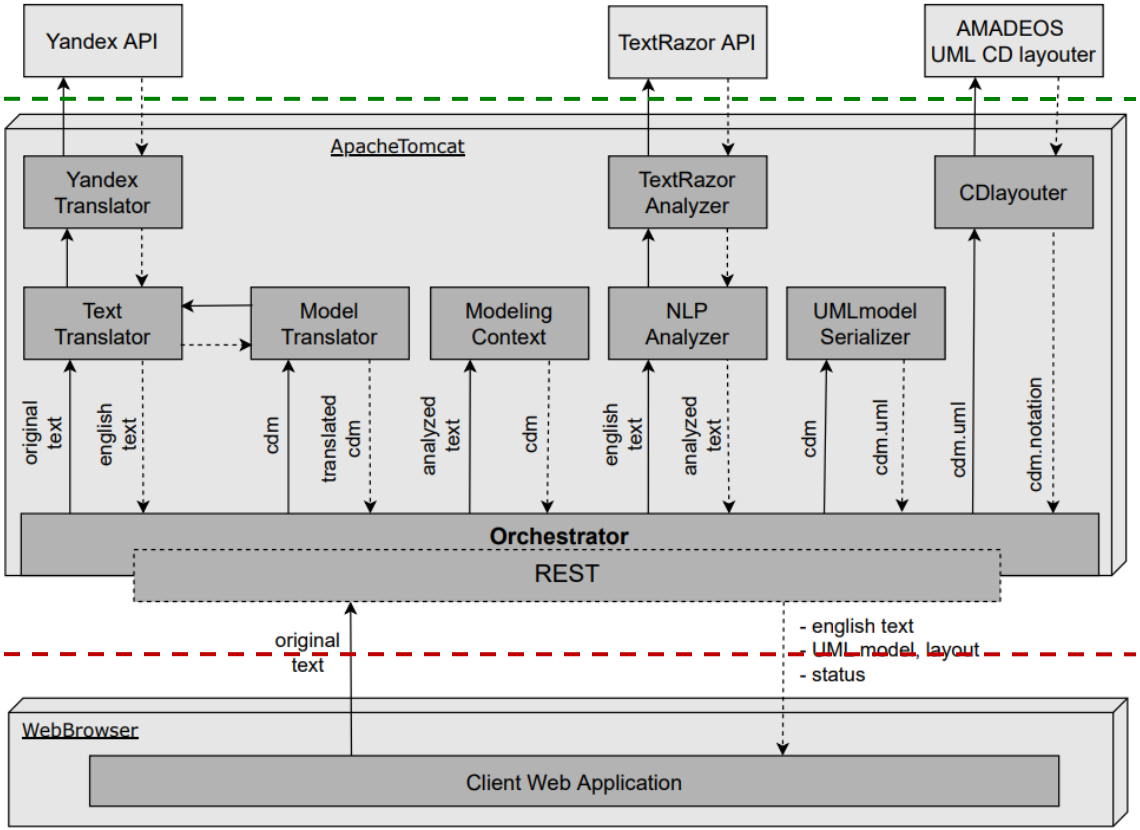


TexToData – System Architecture

External publicly available services

Server side

- Service-oriented Architecture
- The whole process of CDM generation is implemented as an orchestration of local and remote services



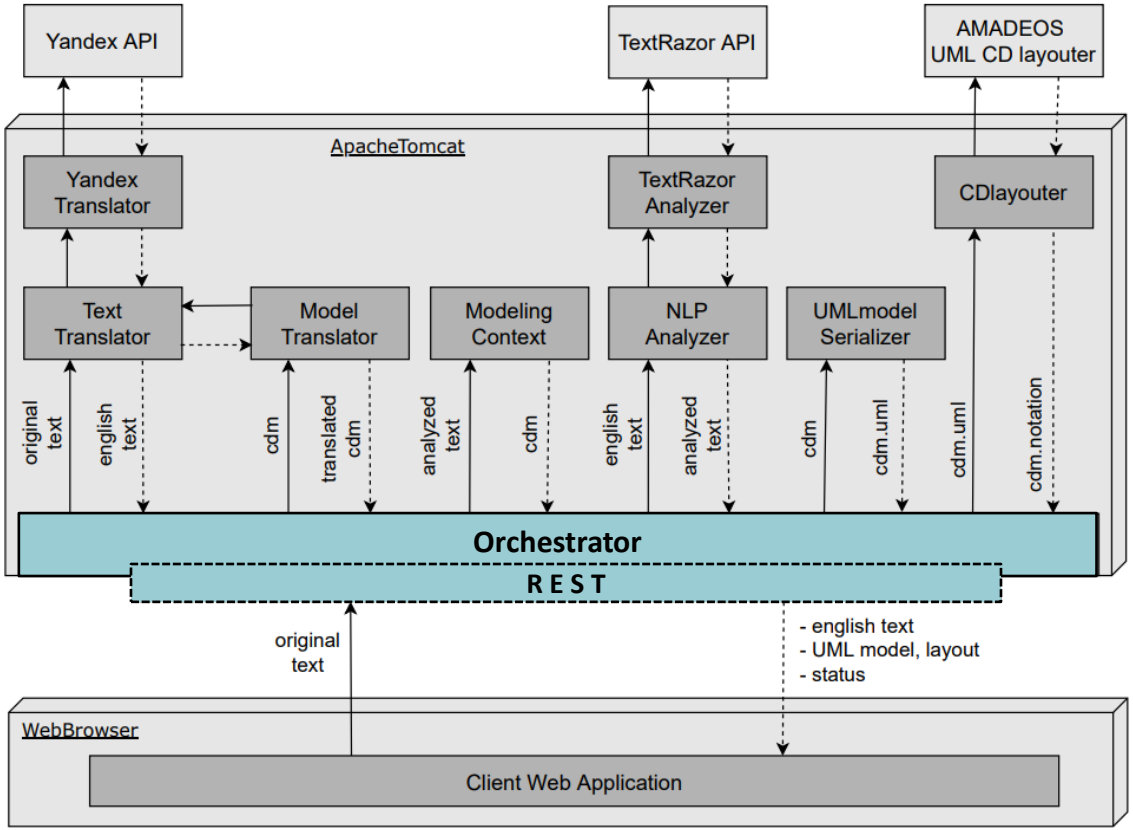
Client side

- GIU, text upload, model manipulations, …

TexToData – System Architecture

Service Orchestration

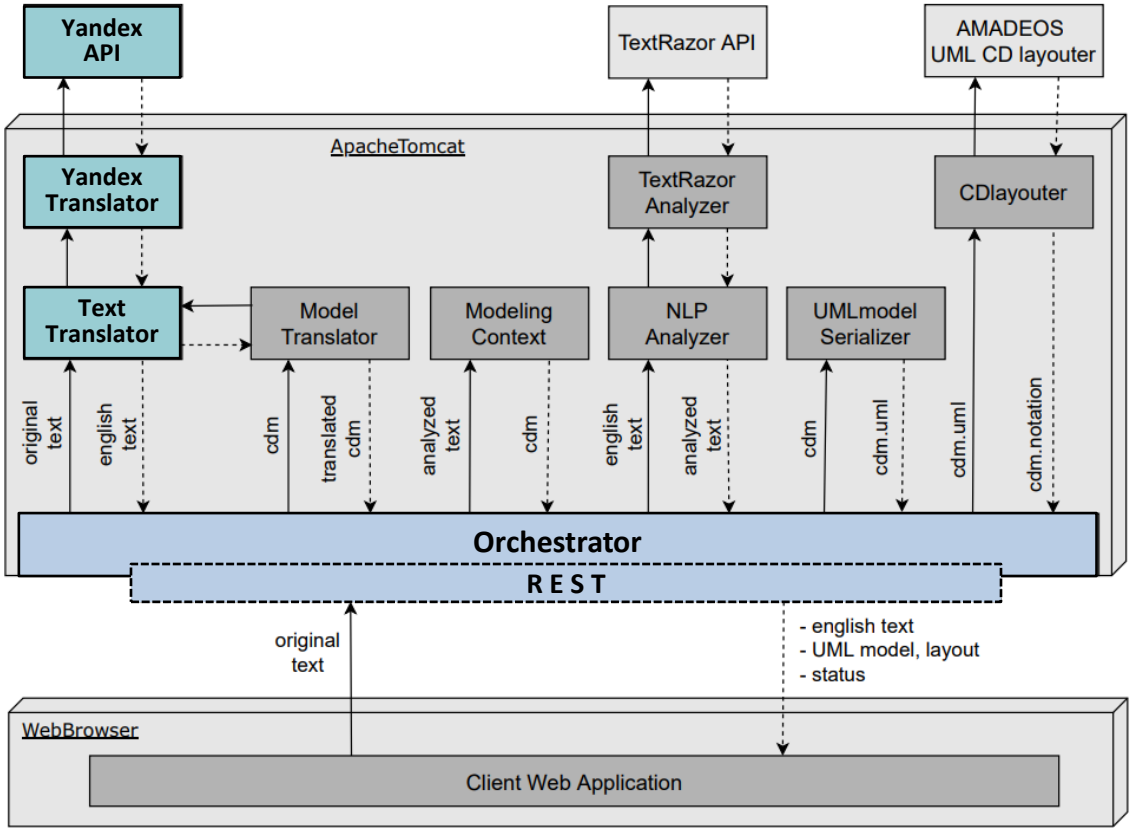
- Orchestrator service orchestrates the whole process
- In a positive usage scenario, the orchestrator receives a text (source NL), and returns the automatically generated CDM



TexToData – System Architecture

Text Translation services

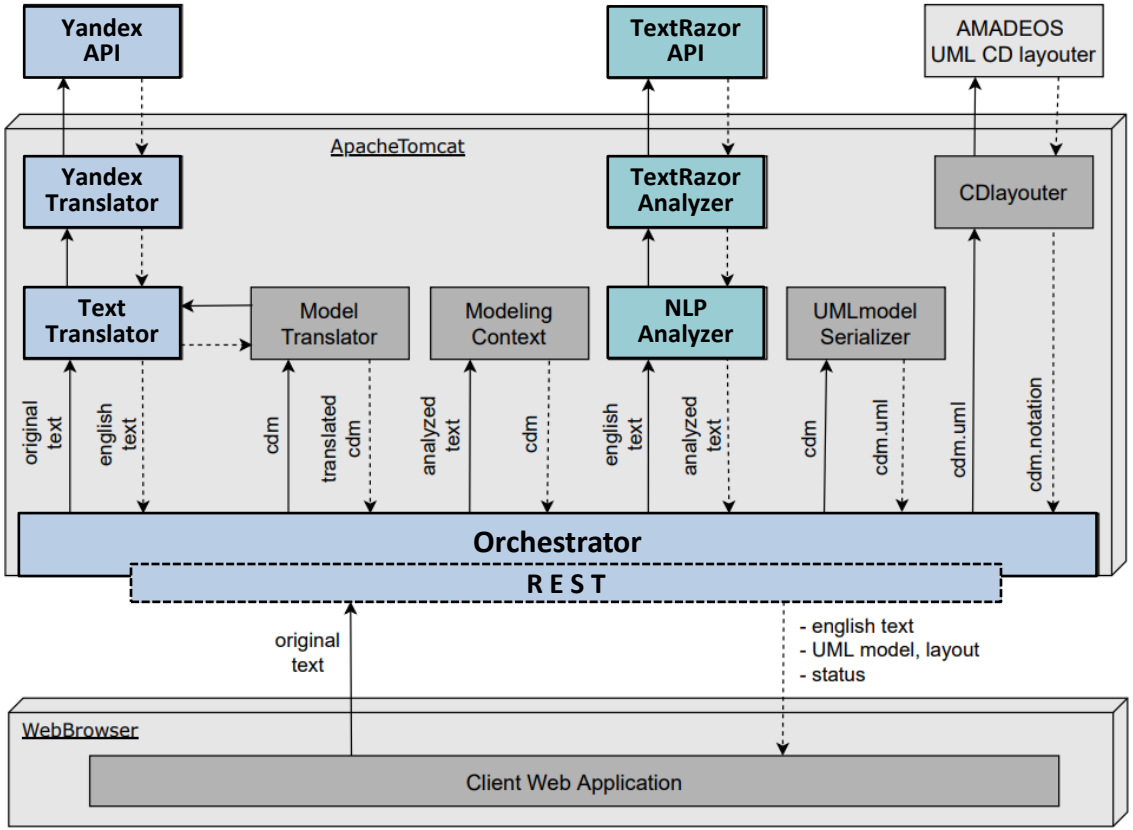
- The source text is firstly sent to the **TextTranslator** service which detects the source NL
- In case the source NL is not English, **TextTranslator** forwards the text to the external translation service through the corresponding adapter
- Currently we employ the **Yandex service** via the **YandexTranslator adapter**



TexToData – System Architecture

NLP services

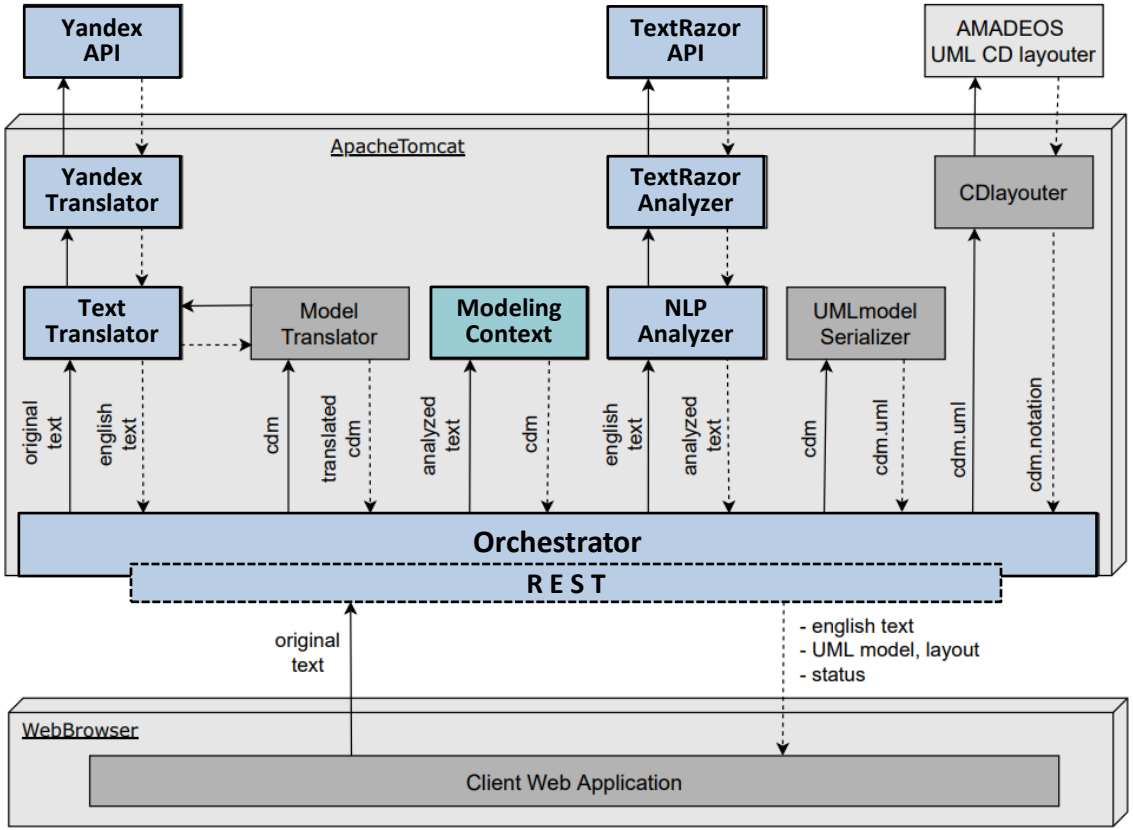
- When we have English text, the Orchestrator sends the English text to the **NLPAnalyzer service** that is responsible for NLP
- The **NLPAnalyzer service** employs the external NLP service via the corresponding adapter
- Currently we employ the **TextRazor service** via the **TextRazorAnalyzer adapter**



TexToData – System Architecture

CDM generation service

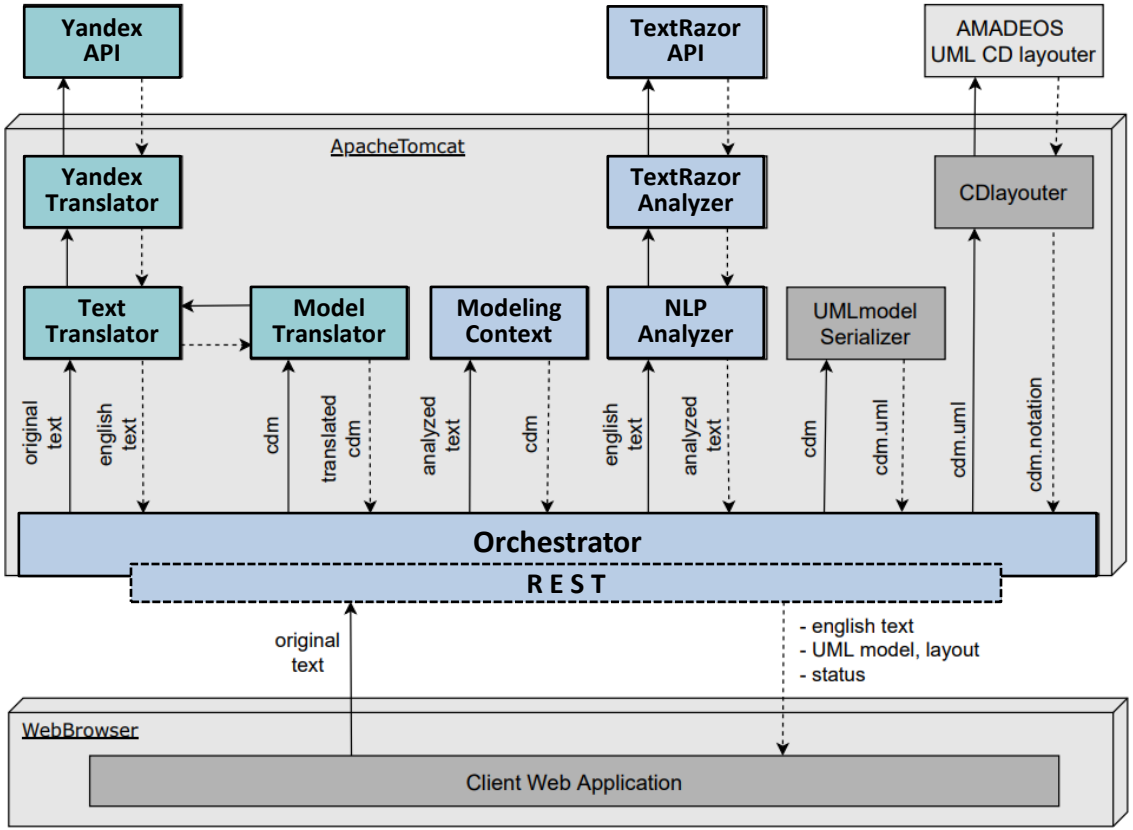
– After NLP is finished, the analyzed text is sent to the **ModelingContext service** which generates an internal representation of the CDM



TexToData – System Architecture

Model Translation services

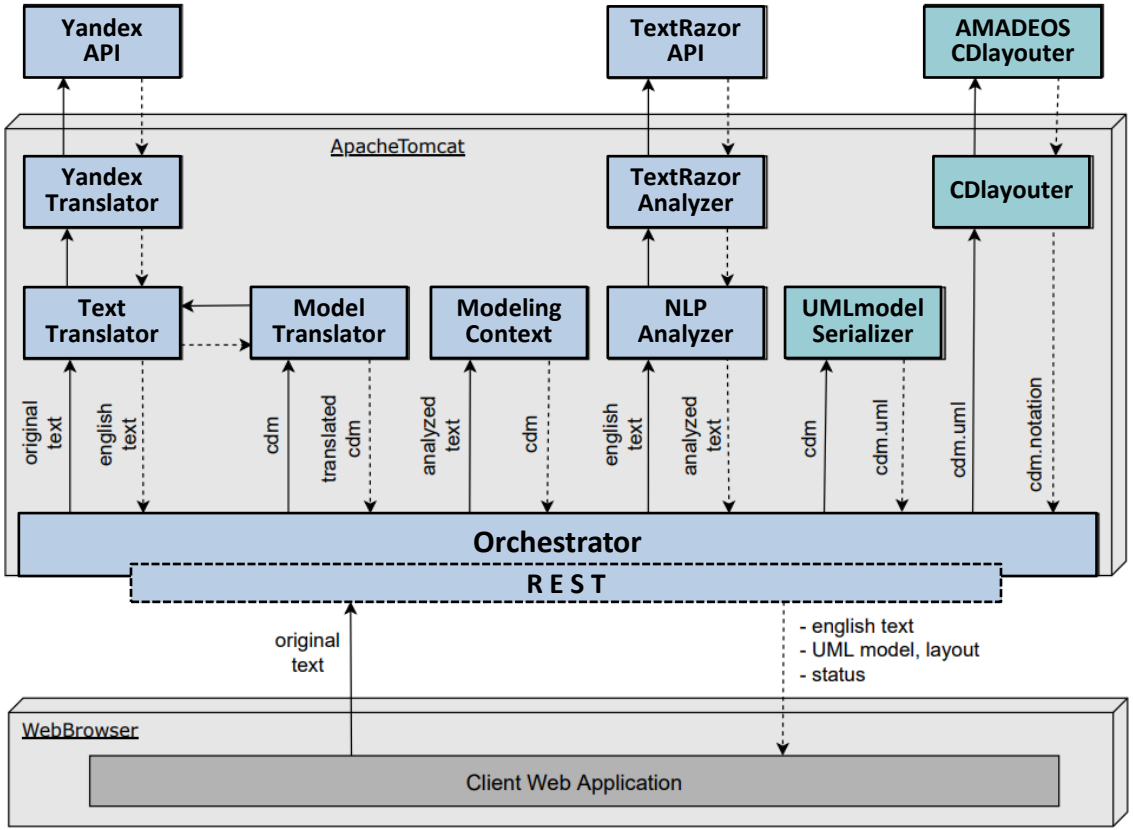
- If the source NL is not English, then the CDM is sent to the **ModelTranslator service**, in order to translate CDM back to the source language
- The **ModelTranslator service** employs the **TextTranslator service** to translate each model element back to the source language



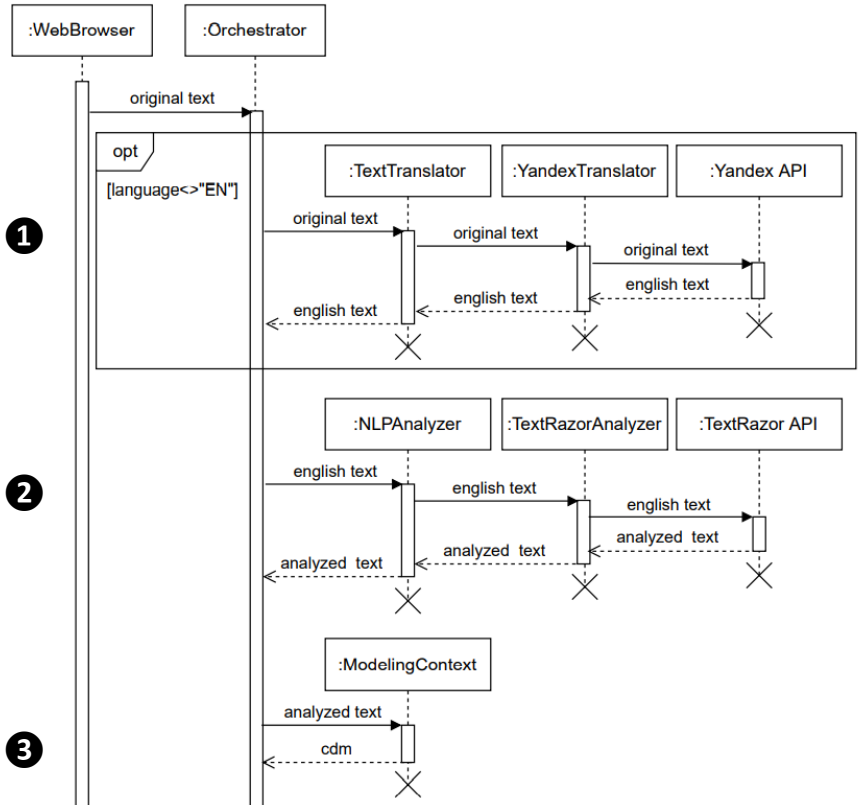
TexToData – System Architecture

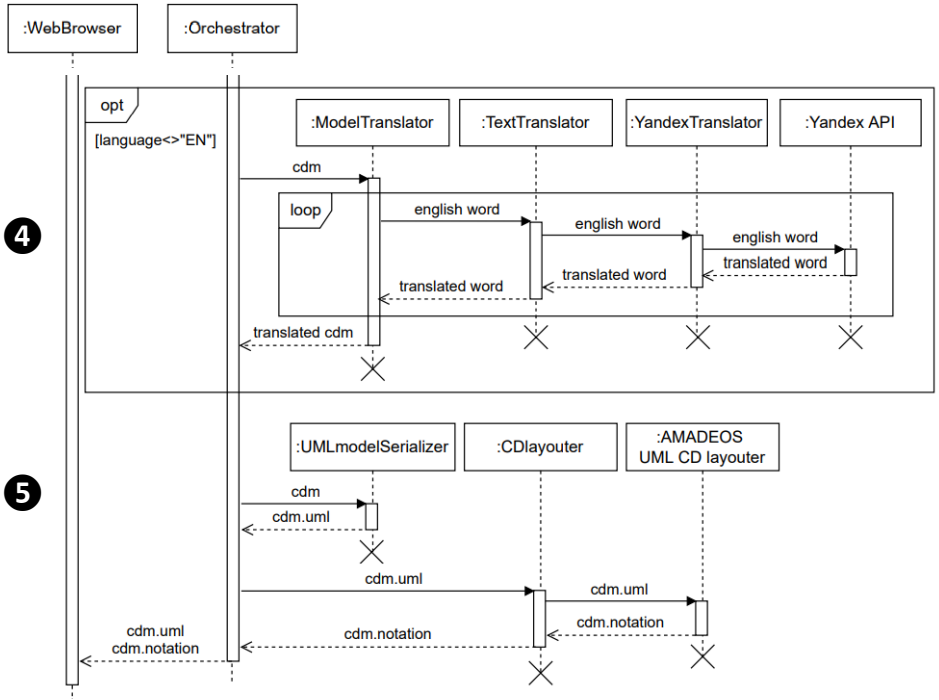
Model & Diagram serialization services

- When the CDM is generated, and translated back to the source language, the Orchestrator service further sends CDM to the **UMLmodelSerializer service** which serializes the generated class diagram in the XMI format
- After the serialization, the model is sent to the **CDlayouter service**, which employs the corresponding **AMADEOS layouting service** and returns a layout of the class diagram
- Finally, the model and the diagram are merged into a single JSON object, and returned to the client.



TexToData – Service Orchestration

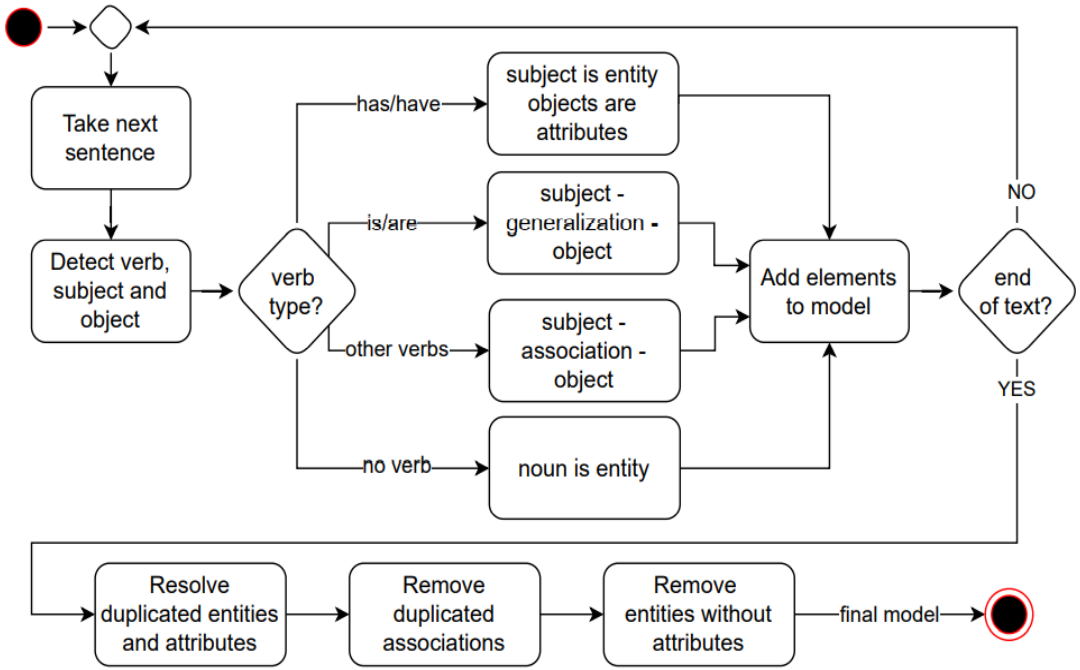




TexToData – From English text to CDM

ModelingContext service

- The core service, responsible for the CDM generation, takes English text split into sentences as input, where each input word is tagged with the corresponding PoS and dependency to other words.
- Currently, we apply a simple strategy, with rules split into two logical phases:
  1. sentences are analyzed separately, adding detected classes, attributes and relationships to the initial model;
  2. cleaning up the initial model, where duplicated classes, relationships, as well as classes without attributes are removed, producing the final model as output.

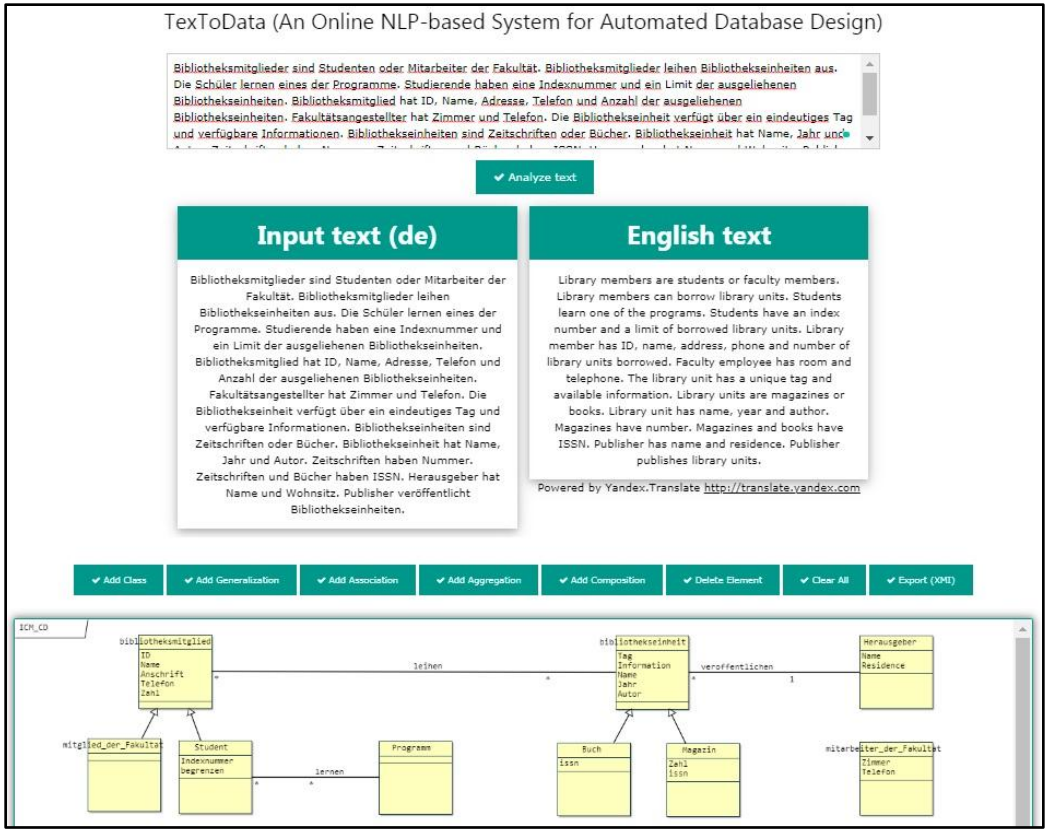


TexToData – Client Side

Client web application

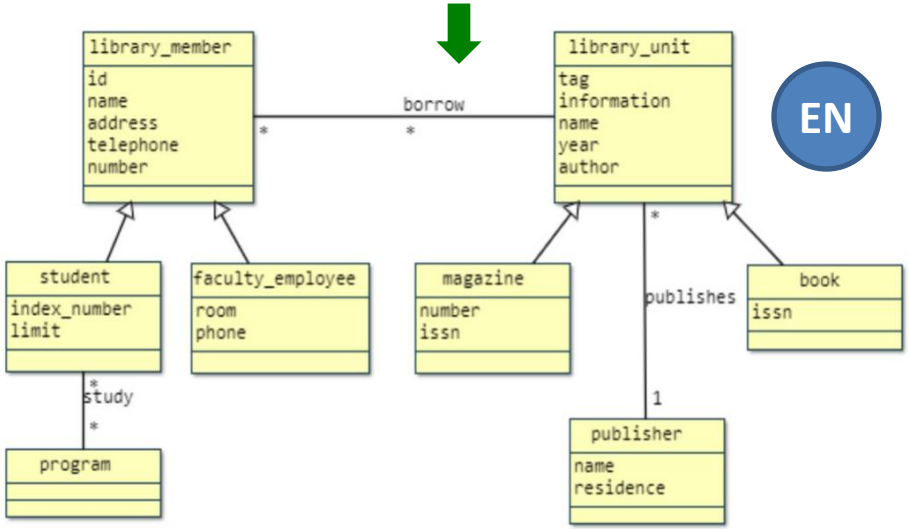
- The client web application allows users to upload a source NL text.
- When the entire synthesis process is finished, the client application receives the JSON response and visualizes the class diagram in the browser.
- The visualized diagram is editable so users can additionally improve it.
- It is also possible to export the model in the XMI format, and further use it in some other platform.

http://m-lab.etf.unibl.org:8080/Textodata

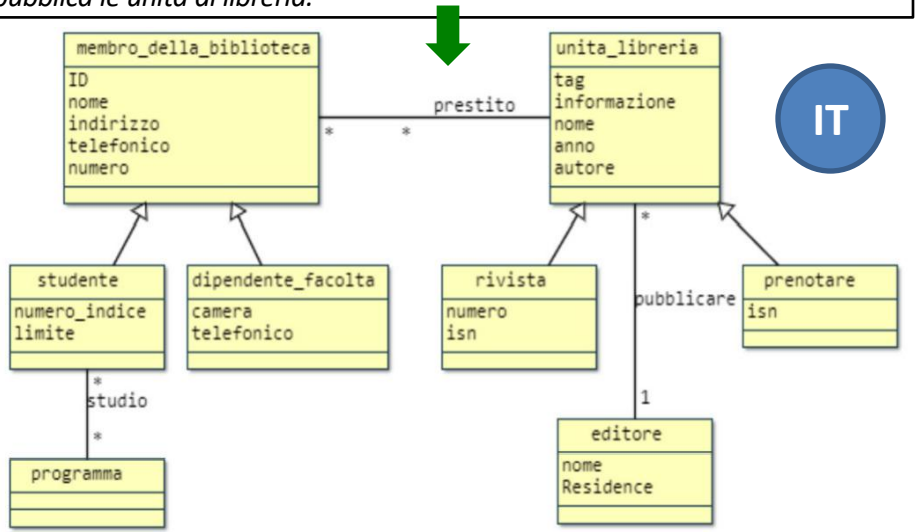


Illustrative examples of automatic CDM generation

Library members are students or faculty employees. Library members borrow library units. Students study one of the programs. Students have index number and limit of library units borrowed. Library member has id, name, address, telephone and number of library units borrowed. Faculty employee has room and phone. Library unit has unique tag and available information. Library units are magazines or books. Library unit has name, year and author. Magazines have number. Magazines and books have ISSN. Publisher has name and residence. Publisher publishes library units.



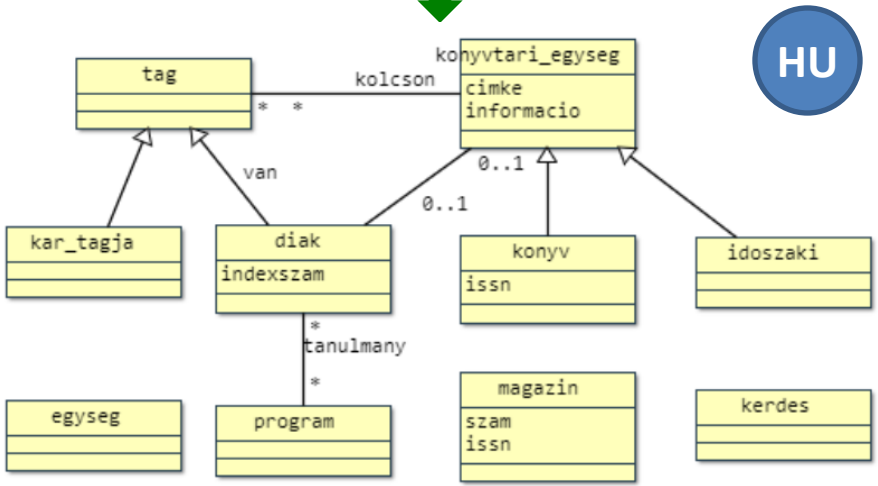
I membri della biblioteca sono studenti o dipendenti di facoltà. I membri della biblioteca prendono in prestito unità della biblioteca. Gli studenti studiano uno dei programmi. Gli studenti hanno il numero di indice e il limite delle unità di biblioteca prese in prestito. Il membro della biblioteca ha id, nome, indirizzo, telefono e numero di unità della biblioteca prese in prestito. L'impiegato della facoltà ha stanza e telefono. L'unità della libreria ha un tag univoco e le informazioni disponibili. Le unità della biblioteca sono riviste o libri. L'unità della biblioteca ha nome, anno e autore. Le riviste hanno il numero. Riviste e libri hanno ISSN. L'editore ha nome e residenza. L'editore pubblica le unità di libreria.

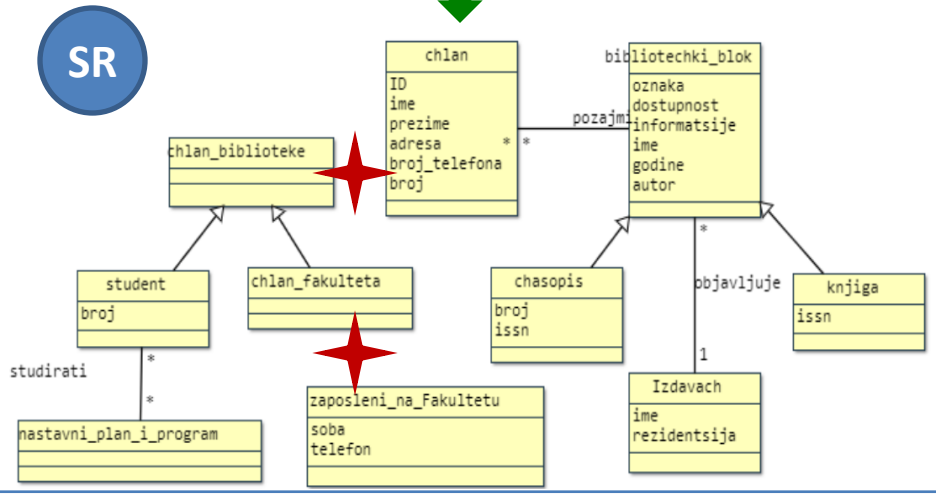


Some less effective examples of CDM generation

A könyvtár tagjai hallgatók vagy oktatók. A könyvtár tagjai könyvtári egységeket kölcsönöznek. A hallgatók az egyik programot tanulják. A hallgatók indexszámmal és a kölcsönzött könyvtári egységekkel rendelkeznek. A könyvtári tag rendelkezik azonosítóval, névvel, címmel, telefonszámmal és a kölcsönzött könyvtári egységek számával. A kar dolgozójának szobája és telefonja van. A könyvtári egység egyedi címkével és elérhető információkkal rendelkezik. A könyvtári egységek folyóiratok vagy könyvek. A könyvtári egység neve, évszáma és szerzője. A folyóiratoknak van száma. A folyóiratoknak és könyveknek van ISSN-je. A kiadónak van neve és lakhelye. A kiadó könyvtári egységeket ad ki.

Чланови библиотеке су студенти или запослени на факултету. Чланови библиотеке позајмљују библиотечке јединице. Студенти студирају један од студијских програма. Студенти имају број индекса и лимит позајмљених библиотечких јединица. Члан библиотеке има ИД, име, презиме, адресу, телефон и број позајмљених библиотечких јединица. Запослени на факултету има собу и телефон. Библиотечка јединица има јединствену ознаку и информације о доступности. Библиотечке јединице су часописи или књиге. Библиотечка јединица има назив, годину и аутора. Часописи имају број. Часописи и књиге имају ИССН. Издавач има назив и пребивалиште. Издавач објављује библиотечке јединице.





Conslusion and future work

- In this paper we presented an approach and the corresponding online web-based tool named TexToData, aimed at automatic derivation of CDMs from textual specifications specified in different source NLs.
- TexToData generates the CDM through an orchestration of web services, whereby some functionalities are performed by external services that are publicly available and free of charge.
- Although the presented examples are quite simple, they are very illustrative and show that the proposed approach and implemented tool enable automatic CDM derivation from textual specifications represented in different NLs, even from NLs with very complex morphology (such as Slavic and some other languages).
- This further implies **that is not necessary to develop and employ different NLP services for different NLs, but only one single NLP service for the English language is enough**?
- Our future work will include:
  - A more extensive evaluation of the approach and the implemented tool with respect to different translation and NLP services, as well as source NLs;
  - Further improvement of the implemented tool.

Thank you!

Drazen Brdjanin, Mladen Grumic, Goran Banjac, Milan Miscevic, Igor Dujlovic, Aleksandar Kelec, Nikola Obradovic, Danijela Banjac, Dragana Volas, Slavko Maric

M-lab Research Group @ Faculty of Electrical Engineering
University of Banja Luka, Bosnia & Herzegovina