The 6th Workshop in Modern Approaches in Data Engineering and Information System Design
Aug 28, 2024, Bayonne, France

ADBIS

MADEISD
Modern Approaches in Data Engineering and Information System Design

Employing Multiple Online Translation Services in a Multilingual Database Design Tool

Danijela Banjac, Milica Matic, Nedeljko Cvijanovic, Drazen Brdjanin, Goran Banjac, and Djordje Stojisavljevic

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

Presentation Outline

- Research context and motivation
- Research objectives and contributions
- TexToData tool
- Support for Multiple Translation Services
- Evaluation and 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, …)

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, …)

AMADEOS

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

The first online web-based tool for automatic CDM derivation from collections of BPMs

TexToData

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

The first online multilingual web-based tool for automatic CDM derivation from natural language text

SpeeD

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

The first tool enabling CDM derivation from recorded speech

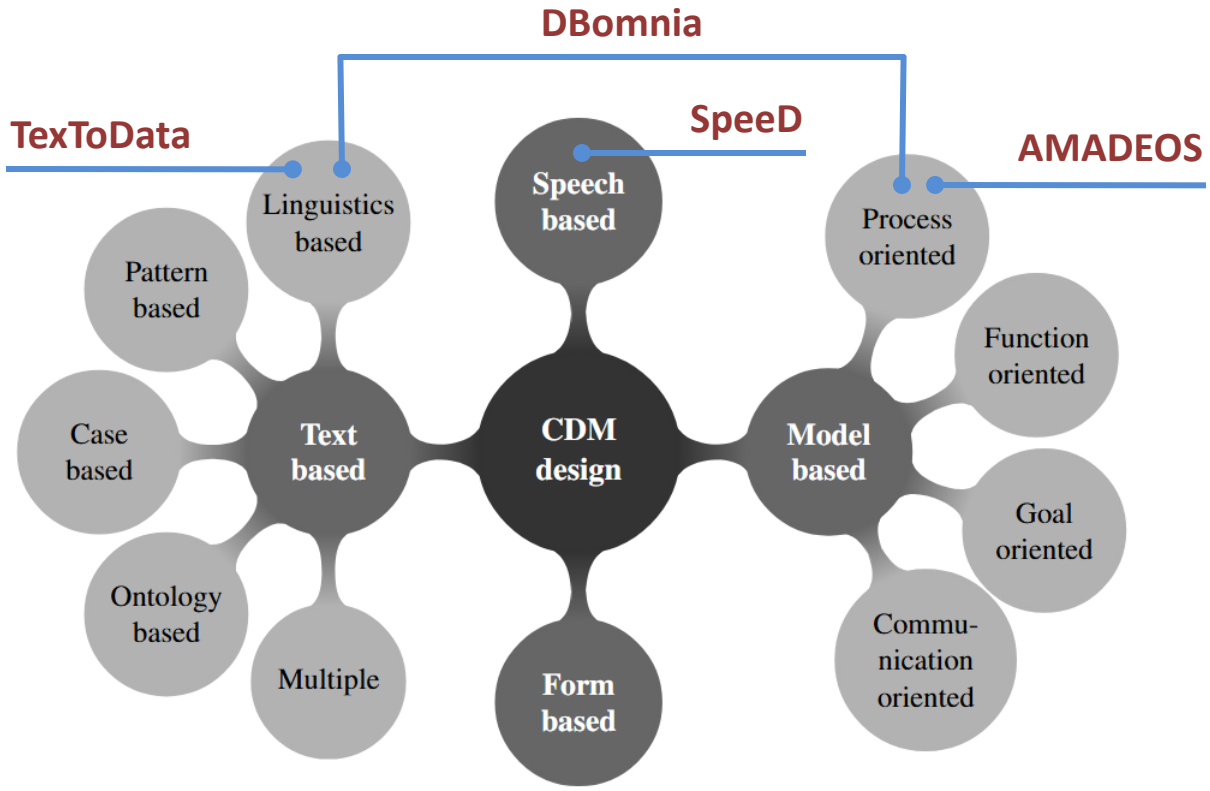
DBomnia

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

The first online web-based tool enabling automatic CDM derivation from heterogeneous source artifacts

Research Context & Motivation

Taxonomy of existing approaches to (semi-)automatic CDM design



Research Context & Motivation

Automatic CDM creation

- Many research publications focus on automatic CDM creation
- About 90% of all the requirements in industrial practices are written in natural language:
  - There is a significant potential demand for the tool that can automatically transform NL text into the CDM
- A lot of NLP research since Chen's eleven rules (1983) for translation of NL text into E-R

Research Context & Motivation

Automatic CDM creation

- Many research publications focus on automatic CDM creation
- About 90% of all the requirements in industrial practices are written in natural language:
  - There is a significant potential demand for the tool that can automatically transform NL text into the CDM
- A lot of NLP research since Chen's eleven rules (1983) for translation of NL text into E-R

Text-based approaches

- Most text-based tools typically support one single source NL (mainly English) and do not provide multilingual support
- Only the TexToData tool enables automatic CDM derivation from textual specifications in different source NLs:
  - In case the source NL is not English, it forwards the text to an external translation service
  - Uses one online external translation service, which represents a bottleneck of the entire process

Research Context & Motivation

Automatic CDM creation

- Many research publications focus on automatic CDM creation
- About 90% of all the requirements in industrial practices are written in natural language:
  - There is a significant potential demand for the tool that can automatically transform NL text into the CDM
- A lot of NLP research since Chen's eleven rules (1983) for translation of NL text into E-R

Text-based approaches

- Most text-based tools typically support one single source NL (mainly English) and do not provide multilingual support
- Only the TexToData tool enables automatic CDM derivation from textual specifications in different source NLs:
  - In case the source NL is not English, it forwards the text to an external translation service
  - Uses one online external translation service, which represents a bottleneck of the entire process

Research objectives

Define an approach and improve the TexToData tool to enable automated CDM derivation with reduced dependence on a particular external translation service

(in order to create more robust system)

Research Objectives & Contributions

Research objectives

- **Define an approach and improve the TexToData tool to enable automated CDM derivation with reduced dependence on a particular external translation service**

  (in order to create more robust system)

Research Objectives & Contributions

Research objectives

- **Define an approach and improve the TexToData tool to enable automated CDM derivation with reduced dependence on a particular external translation service**

  (in order to create more robust system)

Research Contributions

- **Approach**
  - **Analize publicly available online translation services, and integrate into TexToData without changing the current translation process**
- **Improved tool – TexToData**
  - Added support for five free online translation services
  - Evaluated their impact on the effectiveness of CDM derivation

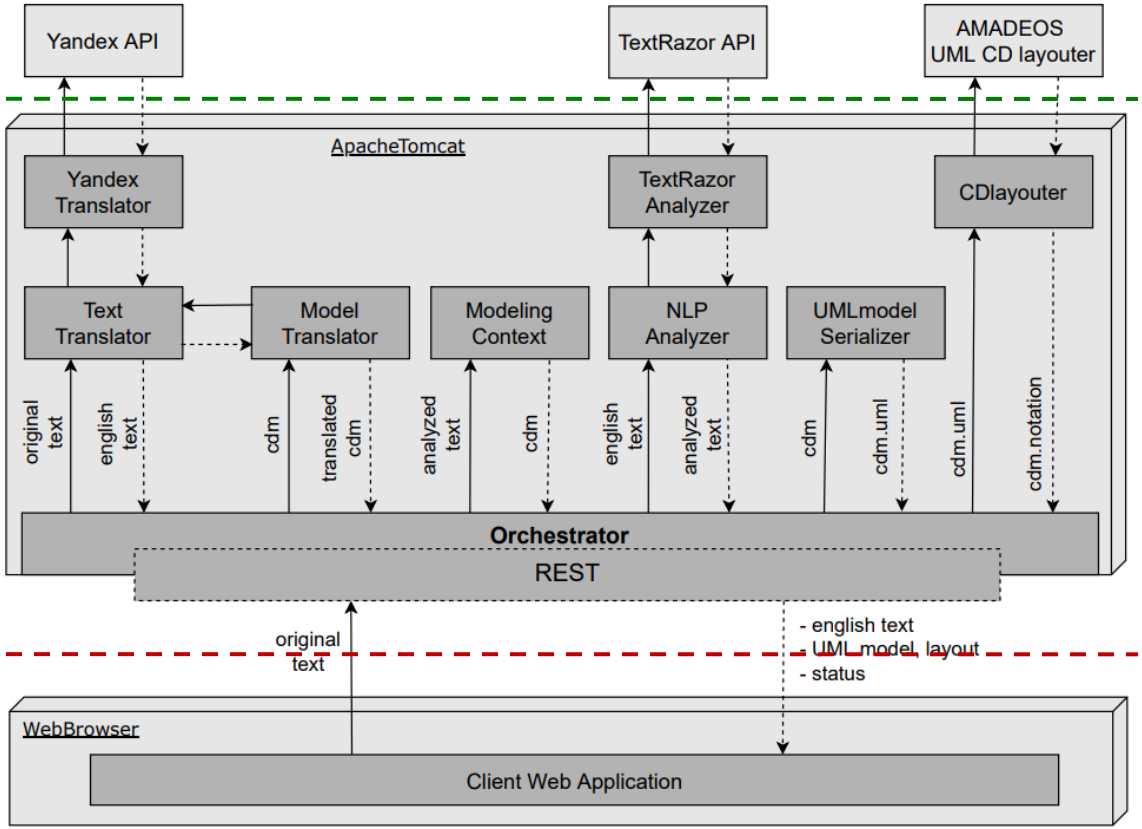
http://m-lab.etf.unibl.org:8080/Textodata

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



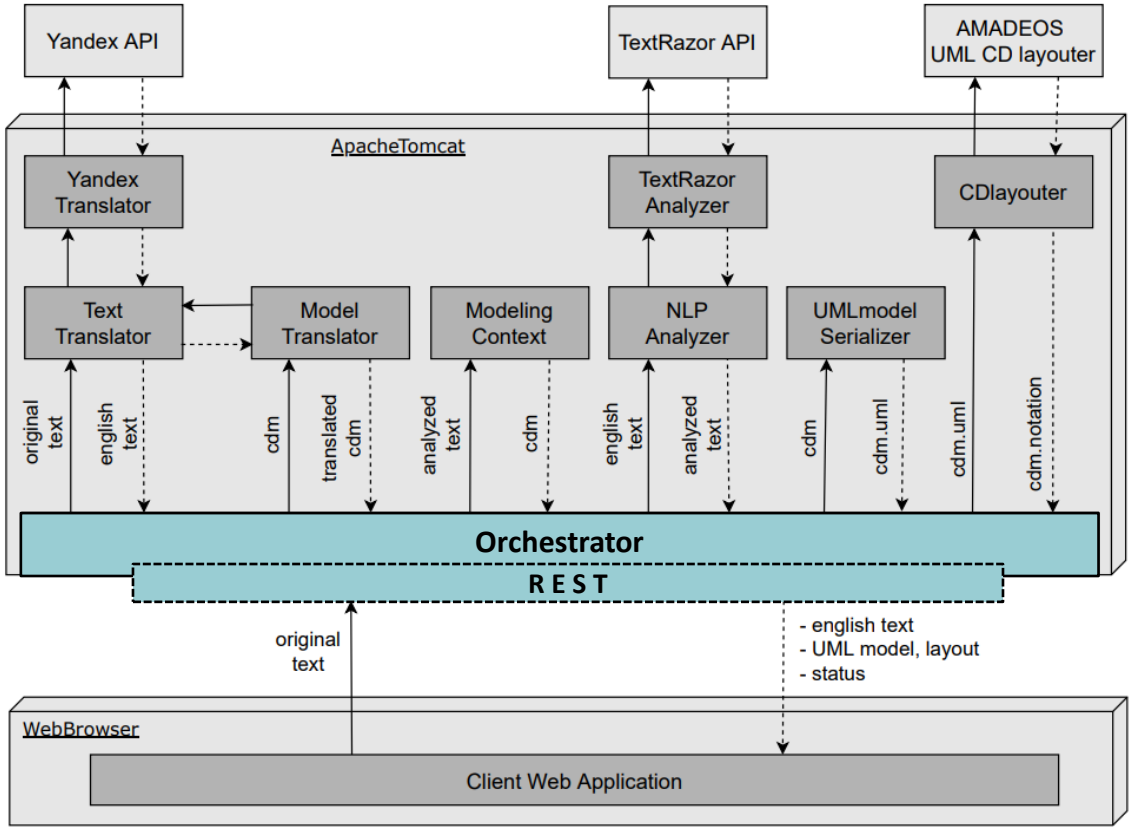
Client side

- GIU, text upload, model manipulations, …

TexToData – System Architecture

Service Orchestration

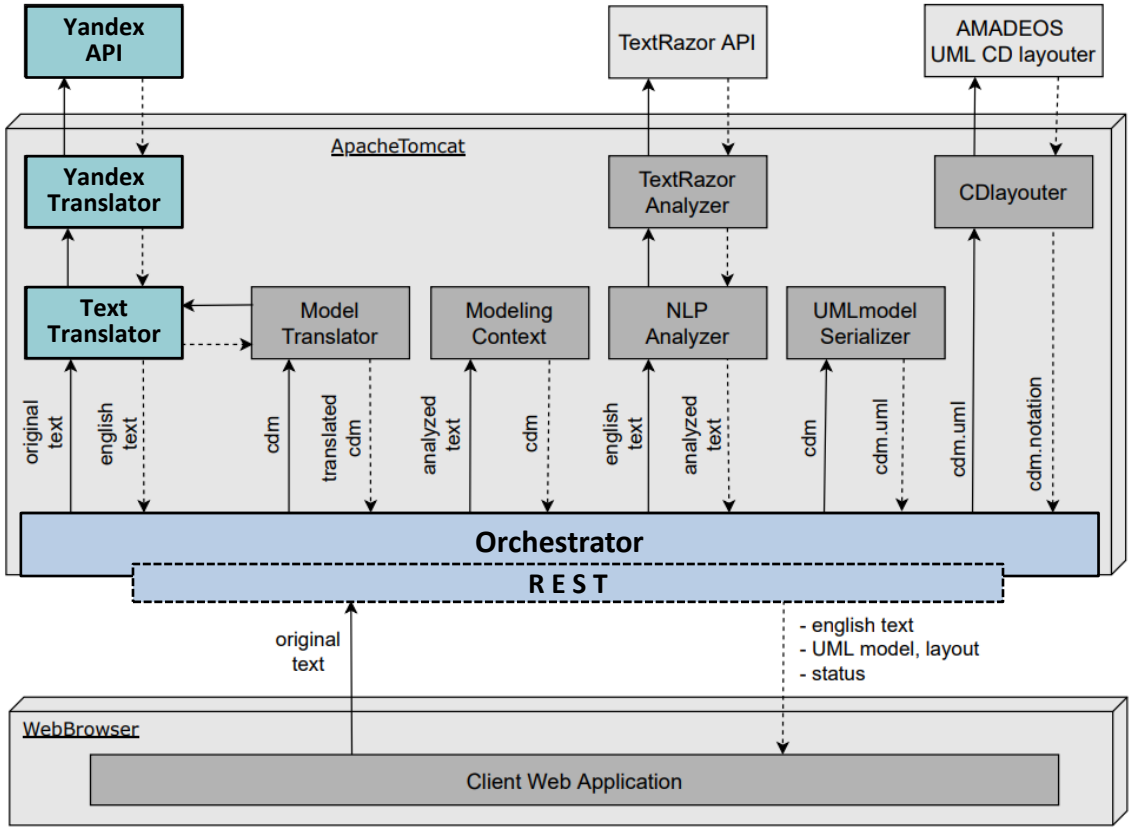
- Orchestrator service orchestrates the whole process
- In a positive usage scenario, the orchestrator receives a text (source NL), and returns the automatically generated CDM



TexToData – System Architecture

Text Translation services

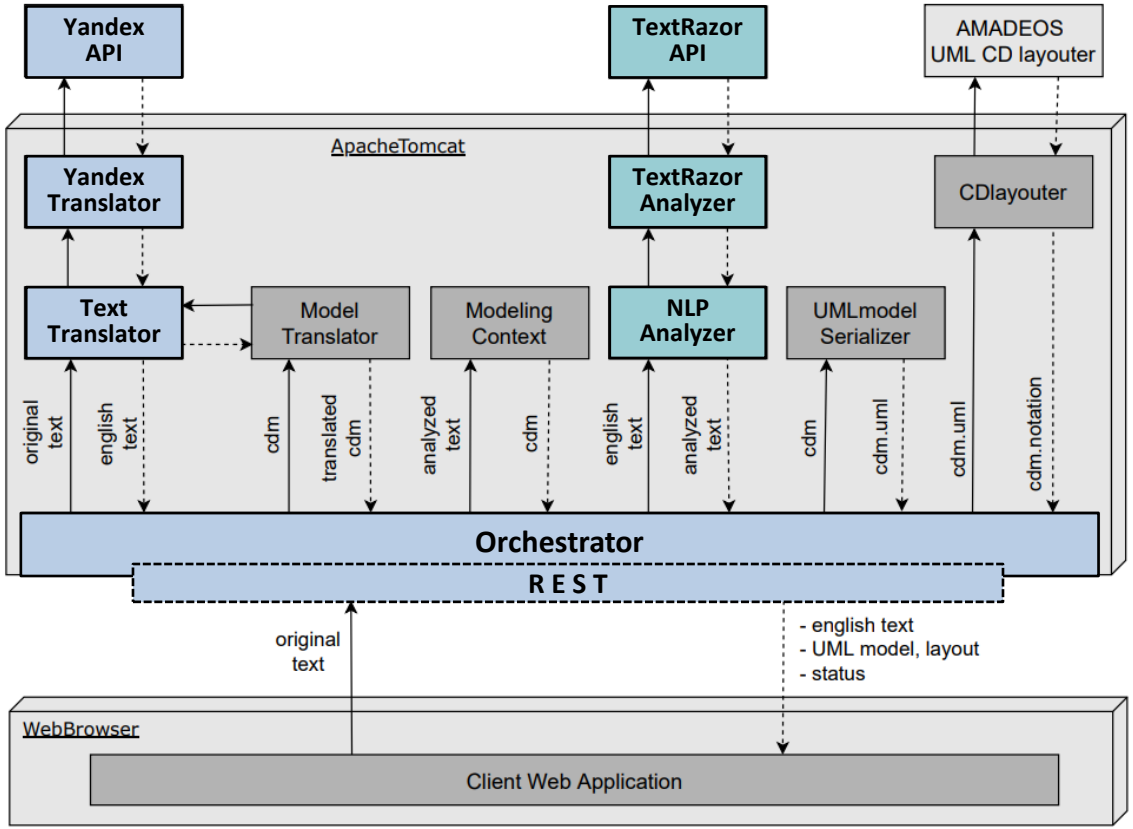
- The source text is firstly sent to the **TextTranslator** service which detects the source NL
- In case the source NL is not English, **TextTranslator** forwards the text to the external translation service through the corresponding adapter
- Currently we employ the **Yandex service** via the **YandexTranslator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



TexToData – System Architecture

NLP services

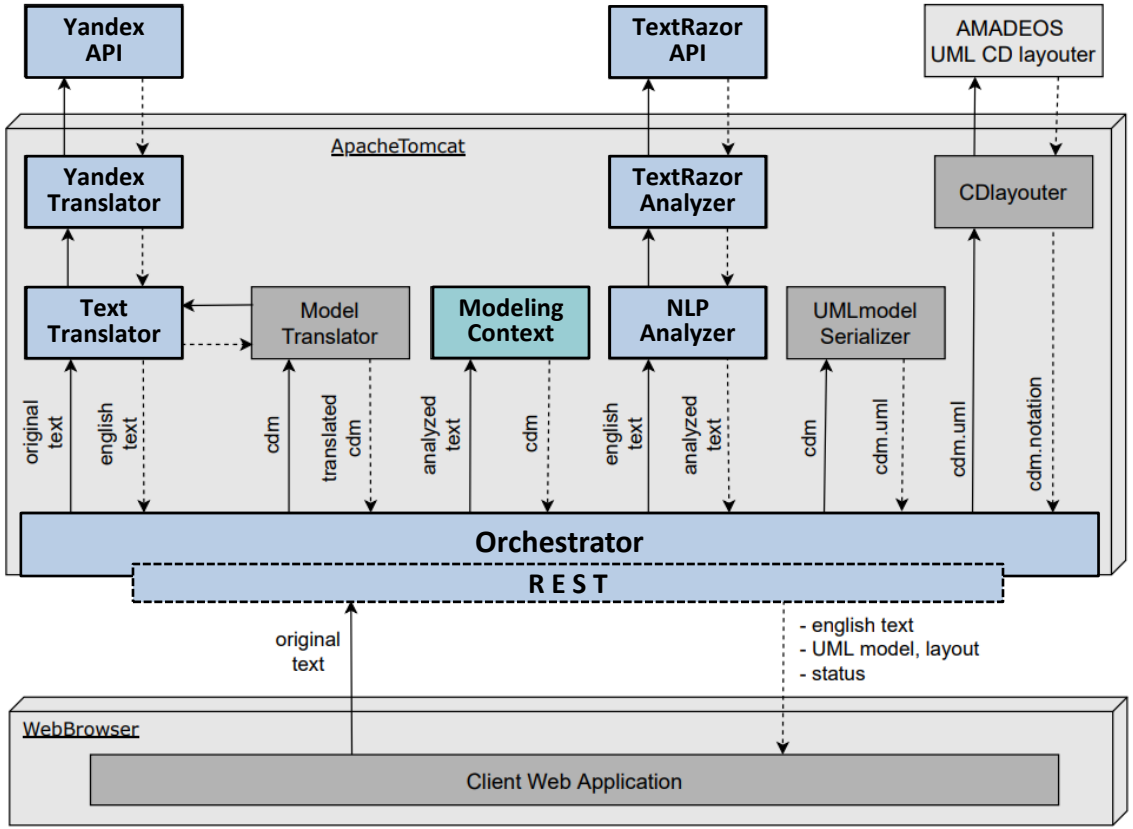
- When we have English text, the Orchestrator sends the English text to the **NLPAnalyzer service** that is responsible for NLP
- The **NLPAnalyzer service** employs the external NLP service via the corresponding adapter
- Currently we employ the **TextRazor service** via the **TextRazorAnalyzer adapter**



TexToData – System Architecture

CDM generation service

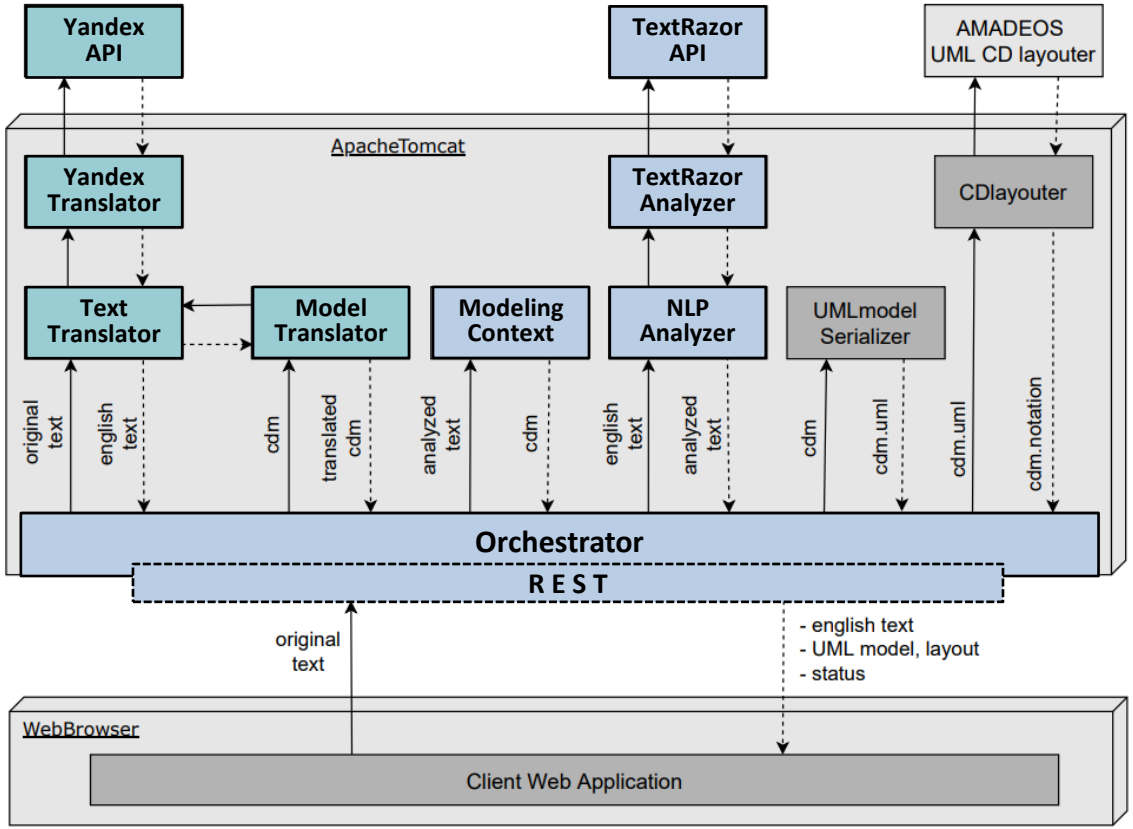
– After NLP is finished, the analyzed text is sent to the **ModelingContext service** which generates an internal representation of the CDM



TexToData – System Architecture

Model Translation services

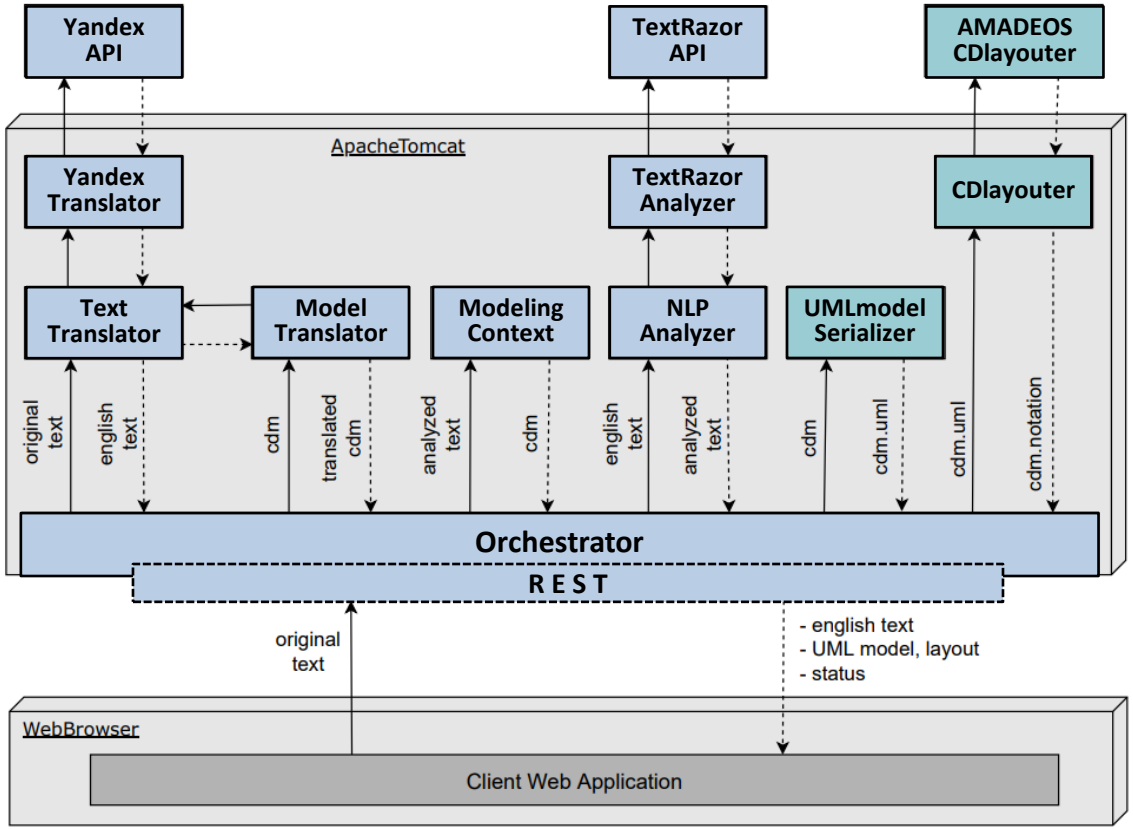
- If the source NL is not English, then the CDM is sent to the **ModelTranslator service**, in order to translate CDM back to the source language
- The **ModelTranslator service** employs the **TextTranslator service** to translate each model element back to the source language



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



Support for Multiple Translation Services

- After an analysis of publicly available online translation services, we identified a set of five additional services:

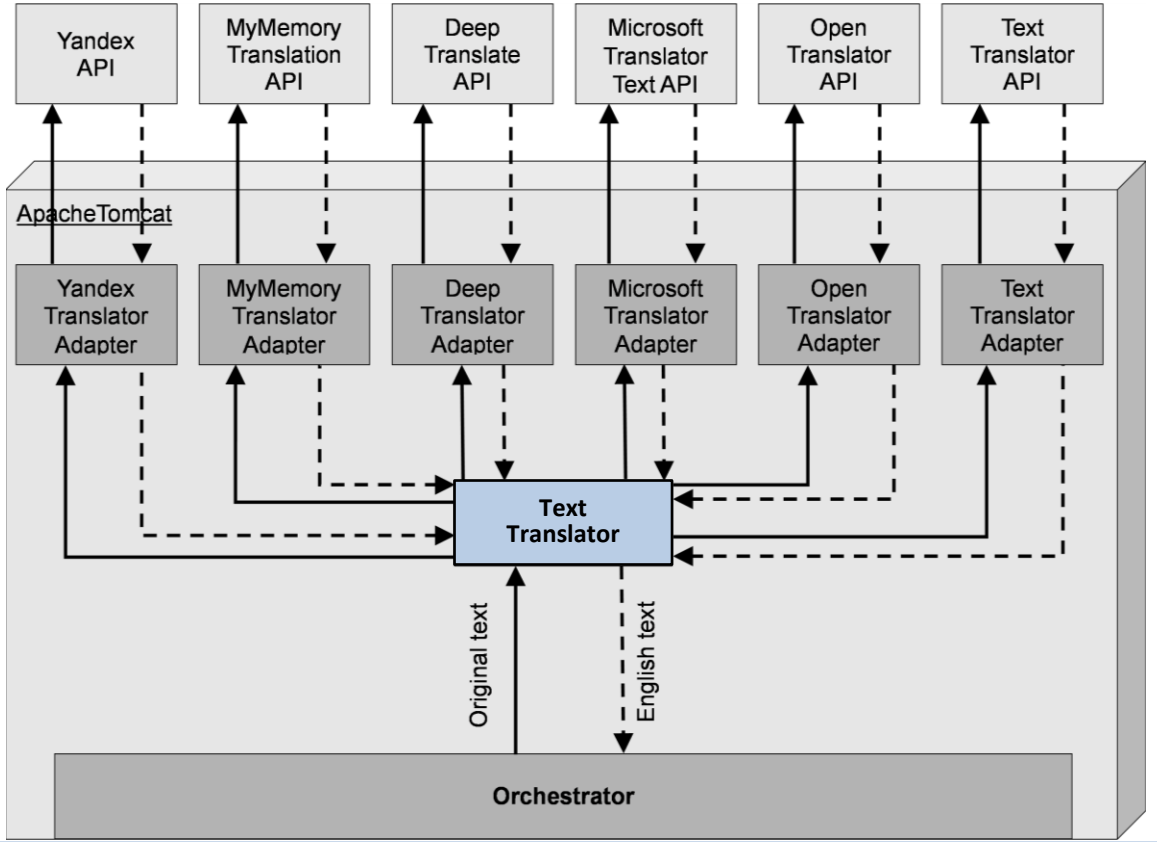
Service name	Automatic NL detection	Number of supported NLs
MyMemory	No	100+
DeepTranslate	Yes	~120
Microsoft Translator Text	Yes	~130
OpenTranslator	Yes	~105
Text Translator	Yes	100+

- These online services expose REST API and return the result as a JSON object
- We analyze only free (basic) subscription plans for all services

TexToData – Extended Architecture

Support of multiple translation services

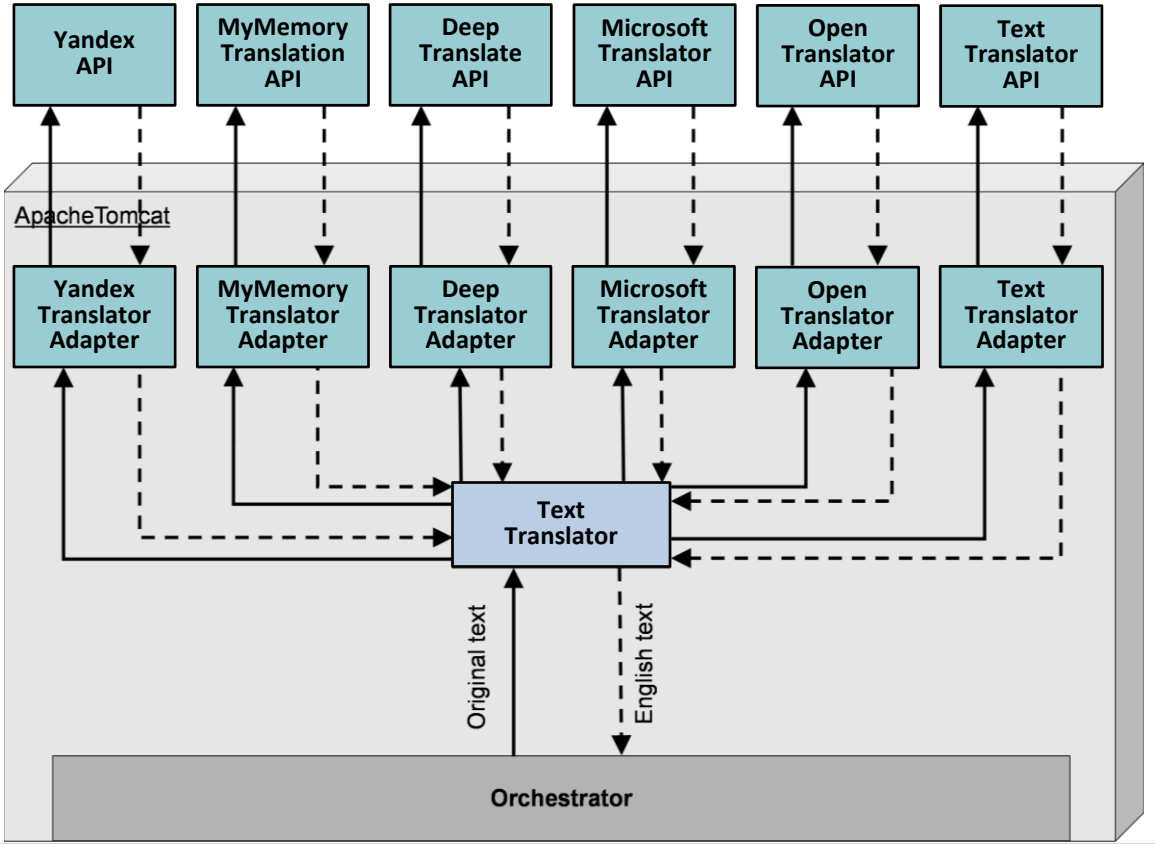
– The improved tool contains additional implementations of the **TextTranslator** interface



TexToData – Extended Architecture

Support of multiple translation services

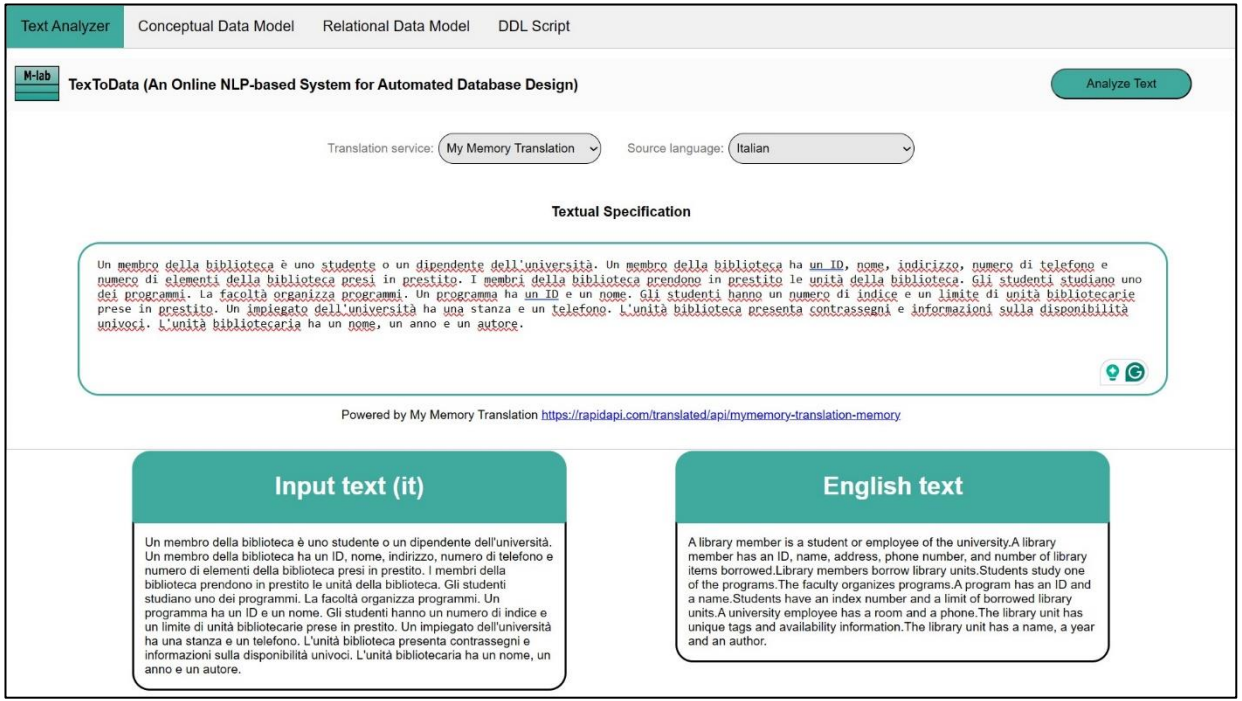
- The improved tool contains additional implementations of the **TextTranslator** interface
- Each implementation communicates with the corresponding external service to translate text from the source NL to English, and after the CDM is created, to translate it from English back to the source NL
- Each implementation of the TextTranslator interface provides methods for the text translating and detection of the input NL



TexToData – Client Side

Client web application

- The client web application allows users to upload a source NL text
- **The user can select one of the supported translation services**
- When the entire synthesis process is finished, the client application receives the JSON response
- The visualized diagram is editable so users can additionally improve it
- It is possible to export the model in the XMI format, and further use it in some other platform



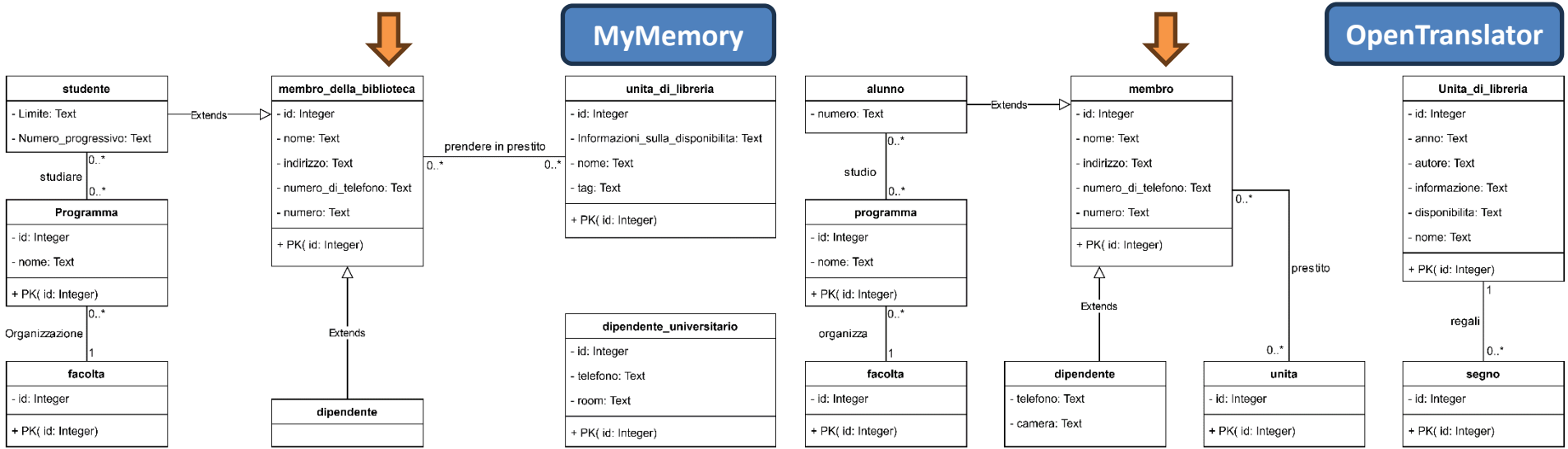
http://m-lab.etf.unibl.org:8080/Textodata

Evaluation

- A case study-based evaluation of the TexToData was performed
- The subject of the study was a Faculty Library
- The initial specification was written in the Serbian language (Latin alphabet), and then translated into Serbian language (Cyrillic alphabet), as well as Italian, German, Greek, and French (by employing the Google Translate service)
- We used TexToData to automatically derive the corresponding CDM for each language by using all of the supported translation services (including Yandex)
- Each generated CDM was manually evaluated against the reference CDM by authors

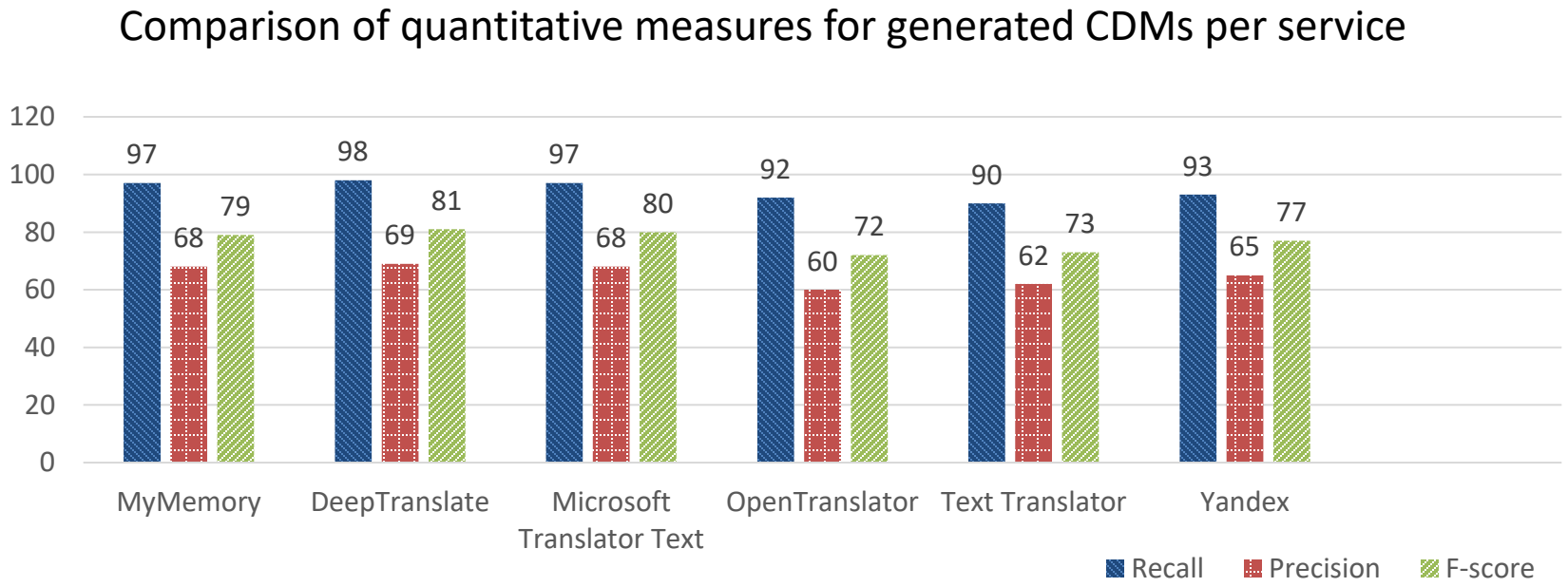
Illustrative examples of automatic CDM generation

Un membro della biblioteca è uno studente o un dipendente dell'università. Un membro della biblioteca ha un ID, nome, indirizzo, numero di telefono e numero di elementi della biblioteca presi in prestito. I membri della biblioteca prendono in prestito le unità della biblioteca. Gli studenti studiano uno dei programmi. La facoltà organizza programmi. Un programma ha un ID e un nome. Gli studenti hanno un numero di indice e un limite di unità bibliotecarie prese in prestito. Un impiegato dell'università ha una stanza e un telefono. L'unità biblioteca presenta contrassegni e informazioni sulla disponibilità univoci. L'unità bibliotecaria ha un nome, un anno e un autore.



Evaluation results

- The results also show that employing different translation services affects generated CDMs



Conslusion and future work

- In this paper we presented an improvement of the TexToData tool for automated database design with multilingual support
- TexToData enable automatic CDM derivation from textual specifications represented in different NLs
- If the source language is not English, an external translation service is employed to translate the textual specification into English and to translate the generated model back into the source natural language
- By adding support for five additional translation services, we achieved automatic CDM derivation that does not depend only on one translation service
- The case study-based evaluation proved that employing multiple online translation services enables effective automatic CDM derivation
- Our future work will include:
  - A more extensive evaluation of the approach and the implemented tool
  - Further improvement in the entire approach to the automatic CDM derivation from text

Thank you!

Danijela Banjac, Milica Matic, Nedeljko Cvijanovic, Drazen Brdjanin, Goran Banjac, and Djordje Stojisavljevic

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