ADBIS 2023

MADEISD
Modern Approaches in Data Engineering and Information System Design

Towards Automatic Conceptual Database Design based on Heterogeneous Source Artifacts

Goran Banjac, Drazen Brdjanin, Danijela Banjac

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 and open issues
- Implemented tool
- Illustrative example
- 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 long-term research project:

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 long-term research project:

Automatic database design

based on sources of different nature

(models, text, speech, …)

Main M-lab achievements:

AMADEOS

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

– The first online web-based tool for automatic CDM derivation from collections of differently represented and differently serialized 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 that provides functionality of CDM derivation from recorded 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 long-term research project:

Automatic database design

based on sources of different nature

(models, text, speech, …)

Main M-lab achievements:

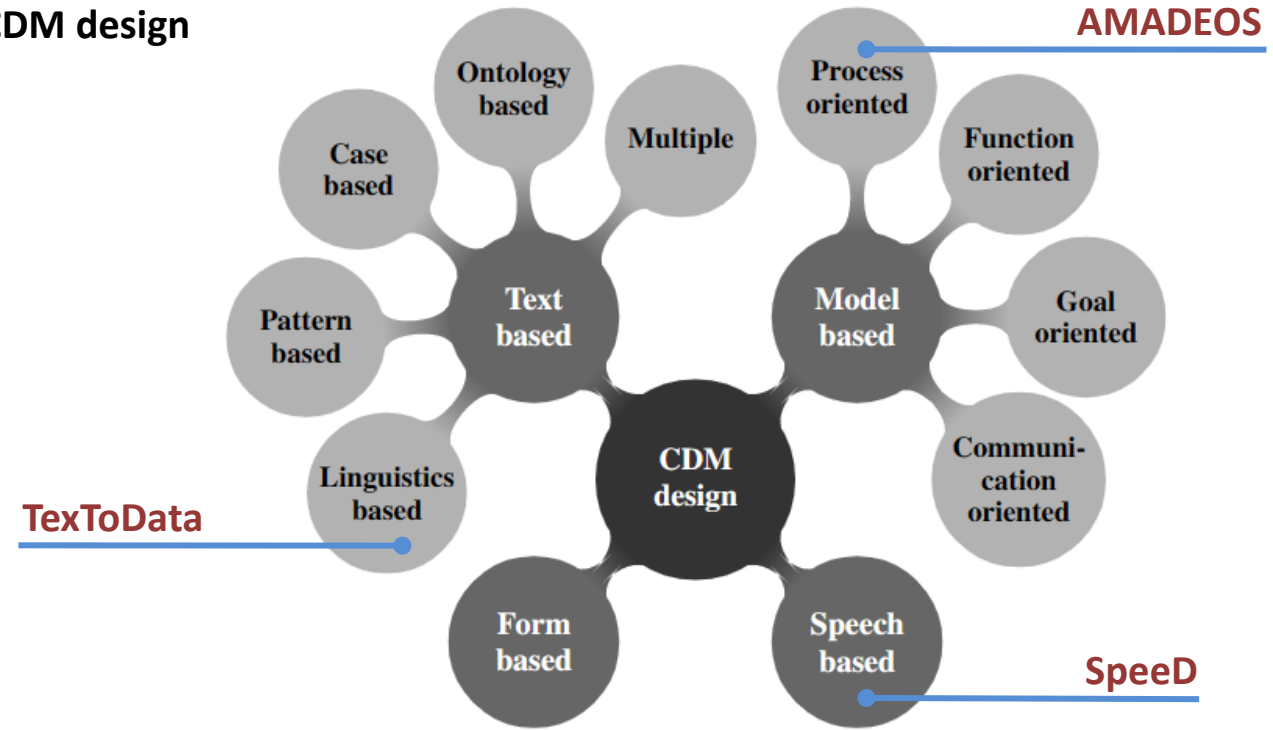
DBomnia

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

- **The first online web-based tool enabling automatic derivation of CDMs from heterogeneous source artifacts (BPMs and textual specifications)**

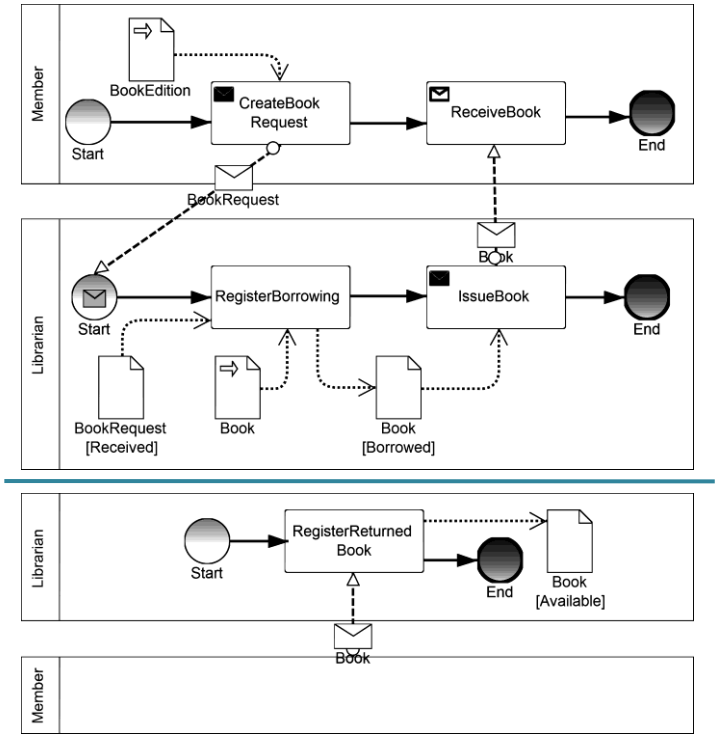
Research Context & Motivation

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



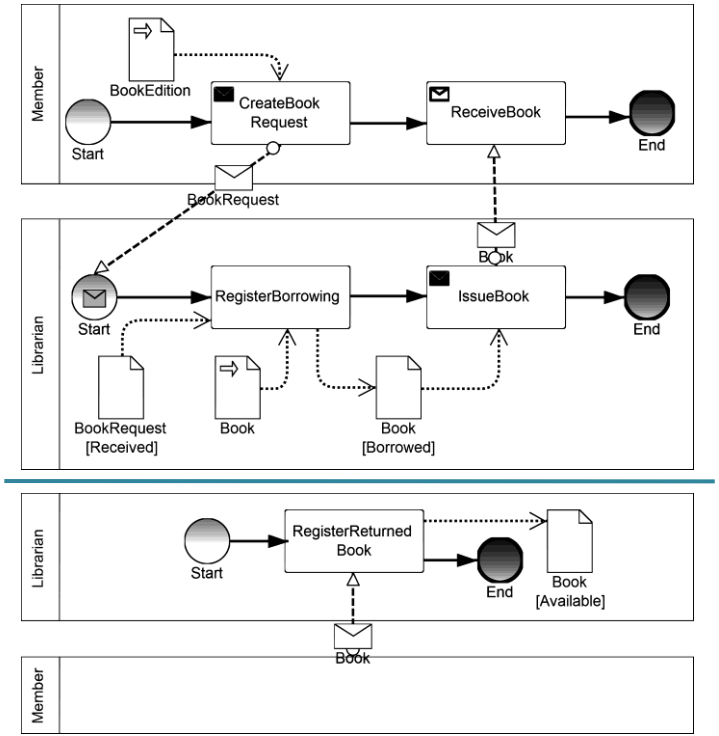
Research Context & Motivation

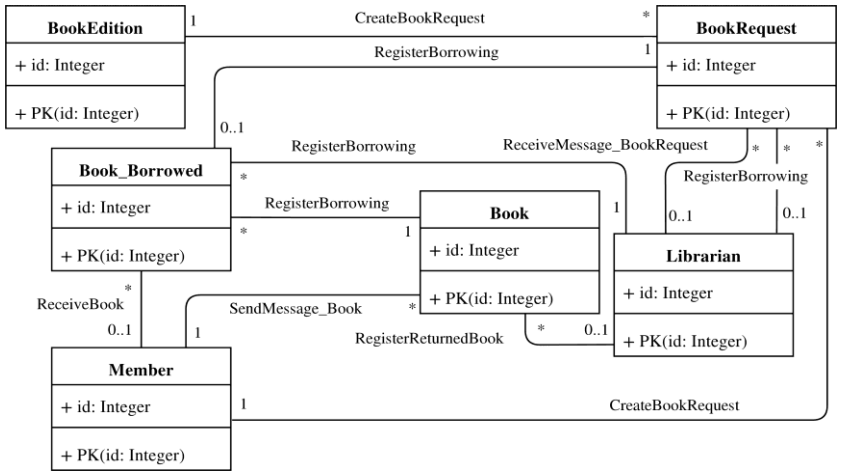
Capabilities of AMADEOS



Research Context & Motivation

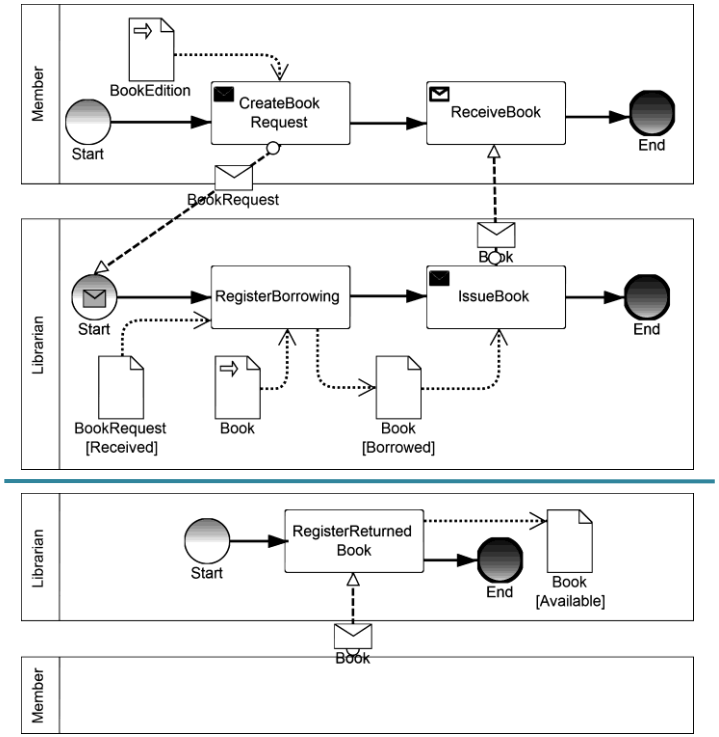
Capabilities of AMADEOS

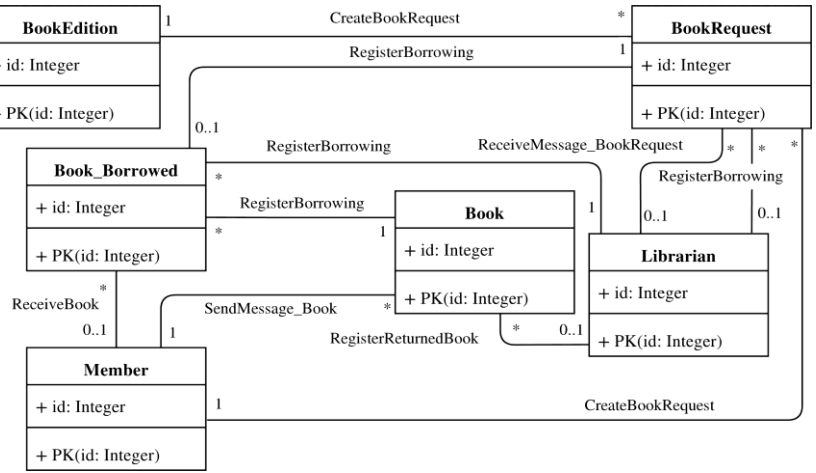




Research Context & Motivation

Capabilities of AMADEOS





- **ADV:** Ability to automatic generation of a highly complete data model structure (~80-100% of entity types and their relationships)
- **DIS:** Modest percentage of attributes in entity types (only *id* attribute in each entity type)

Research Context & Motivation

Capabilities of TexToData

Library users are librarians or members. Library user has name, email, username, and password. Librarian has residence. Member has date of birth.

Book edition has title, year, isbn, authors names, publishers names, fields, and UDC groups. Book has tag. Books belong to book edition.

Member creates borrowing requests. Borrowing request has date. Borrowing requests belongs to book edition. Librarian registers borrowings and issues books.

Member returns borrowed book. Librarian registers returned book.

Research Context & Motivation

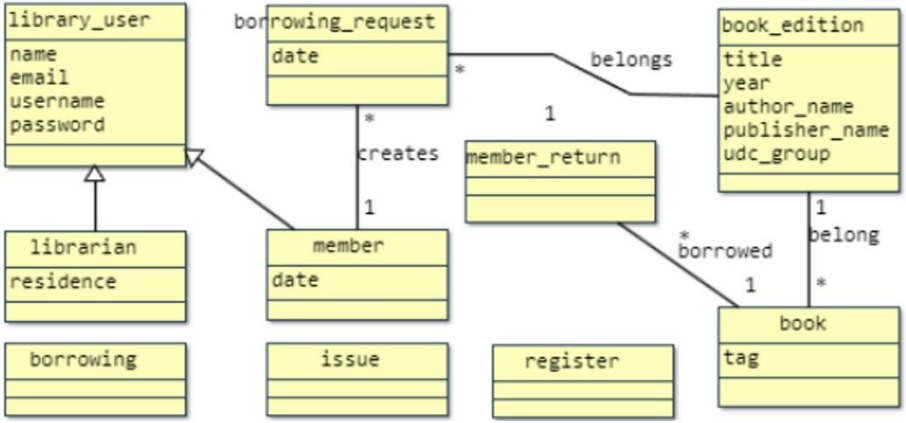
Capabilities of TexToData

Library users are librarians or members. Library user has name, email, username, and password. Librarian has residence. Member has date of birth.

Book edition has title, year, isbn, authors names, publishers names, fields, and UDC groups. Book has tag. Books belong to book edition.

Member creates borrowing requests. Borrowing request has date. Borrowing requests belongs to book edition. Librarian registers borrowings and issues books.

Member returns borrowed book. Librarian registers returned book.



Research Context & Motivation

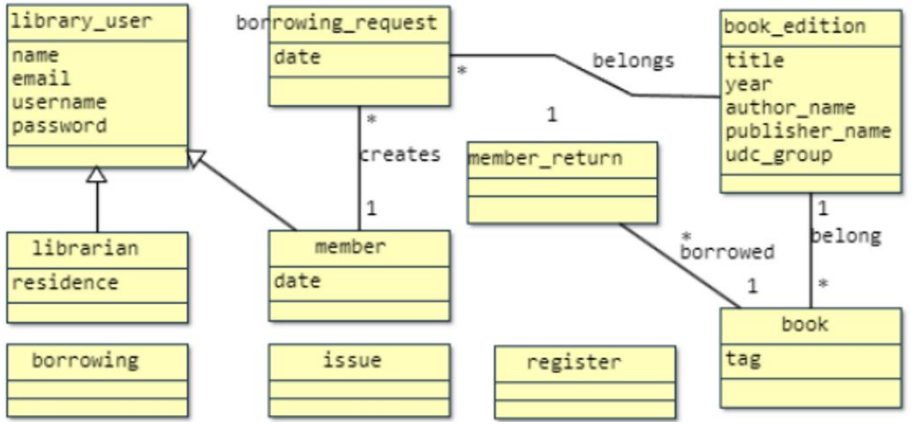
Capabilities of TexToData

Library users are librarians or members. Library user has name, email, username, and password. Librarian has residence. Member has date of birth.

Book edition has title, year, isbn, authors names, publishers names, fields, and UDC groups. Book has tag. Books belong to book edition.

Member creates borrowing requests. Borrowing request has date. Borrowing requests belongs to book edition. Librarian registers borrowings and issues books.

Member returns borrowed book. Librarian registers returned book.



- **ADV:** Ability to automatic generation of more complete set of attributes in each entity type
- **DIS:** Less complete and less correct data model structure

Research Context & Motivation

Capabilities of AMADEOS

- **ADV:** Ability to automatic generation of a highly complete data model structure (~80-100% of entity types and their relationships)
- **DIS:** Modest percentage of attributes in entity types (only *id* attribute in each entity type)

Capabilities of TexToData

- **ADV:** Ability to automatic generation of more complete set of attributes in each entity type
- **DIS:** Less complete and less correct data model structure

Research Context & Motivation

Capabilities of AMADEOS

- **ADV:** Ability to automatic generation of a highly complete data model structure (~80-100% of entity types and their relationships)
- **DIS:** Modest percentage of attributes in entity types (only *id* attribute in each entity type)

Capabilities of TexToData

- **ADV:** Ability to automatic generation of more complete set of attributes in each entity type
- **DIS:** Less complete and less correct data model structure

Research objectives

Define an approach and implement a tool enabling automatic CDM derivation from a set of heterogeneous source artifacts

(try to maximize the correctness and completeness of the CDM by integrating CDMs derived from different sources)

Research Objectives & Contributions

Research objectives

- **Define an approach and implement a tool enabling automatic CDM derivation from a set of heterogeneous source artifacts**

  (try to maximize the correctness and completeness of the CDM by integrating CDMs derived from different sources)

Research Objectives & Contributions

Research objectives

- **Define an approach and implement a tool enabling automatic CDM derivation from a set of heterogeneous source artifacts**

  (try to maximize the correctness and completeness of the CDM by integrating CDMs derived from different sources)

Research Contributions

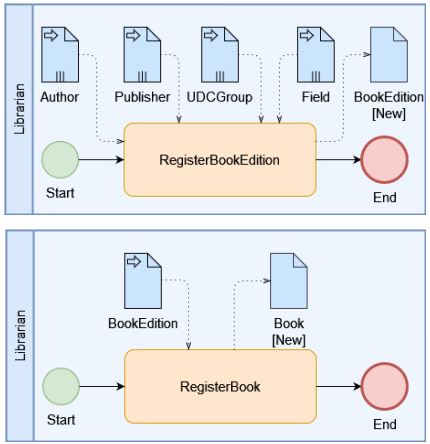
- **Approach**
  - **Employment of existing tools for generation of CDMs from specific source artifacts**
  - **Integration of those *uncertain* CDMs**
- **Implemented tool – DBomnia**
  - online web-based tool
  - support for two types of source artifacts (BPMs and textual specifications)
  - automatic layouting and UML-based representation of generated CDM (editing and formatting functionalities, XMI-export to support model portability, ...)

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

Approach Outline

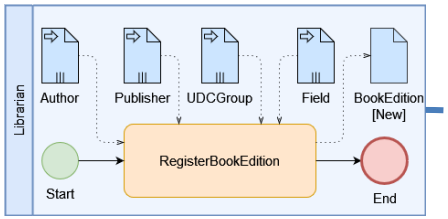
Task

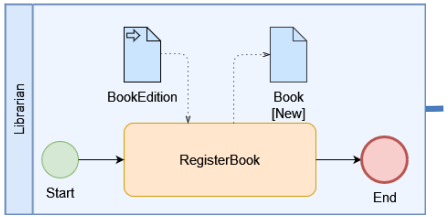
- Integration (matching & merging) of multiple (unreliable) CDMs into a single unified CDM
- Currently, we have only two types of source artifacts (BPMs and textual specifications), so we focus on the integration of two CDMs

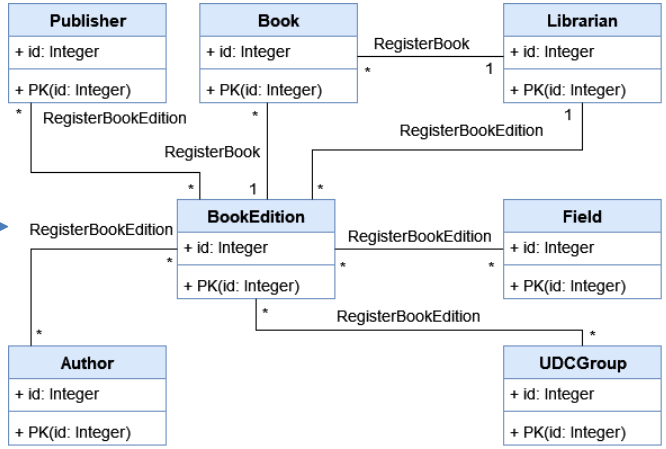


Library member has id and name. Books have name, author and ISSN. Library member can borrow books.

Approach Outline

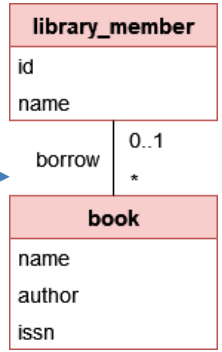






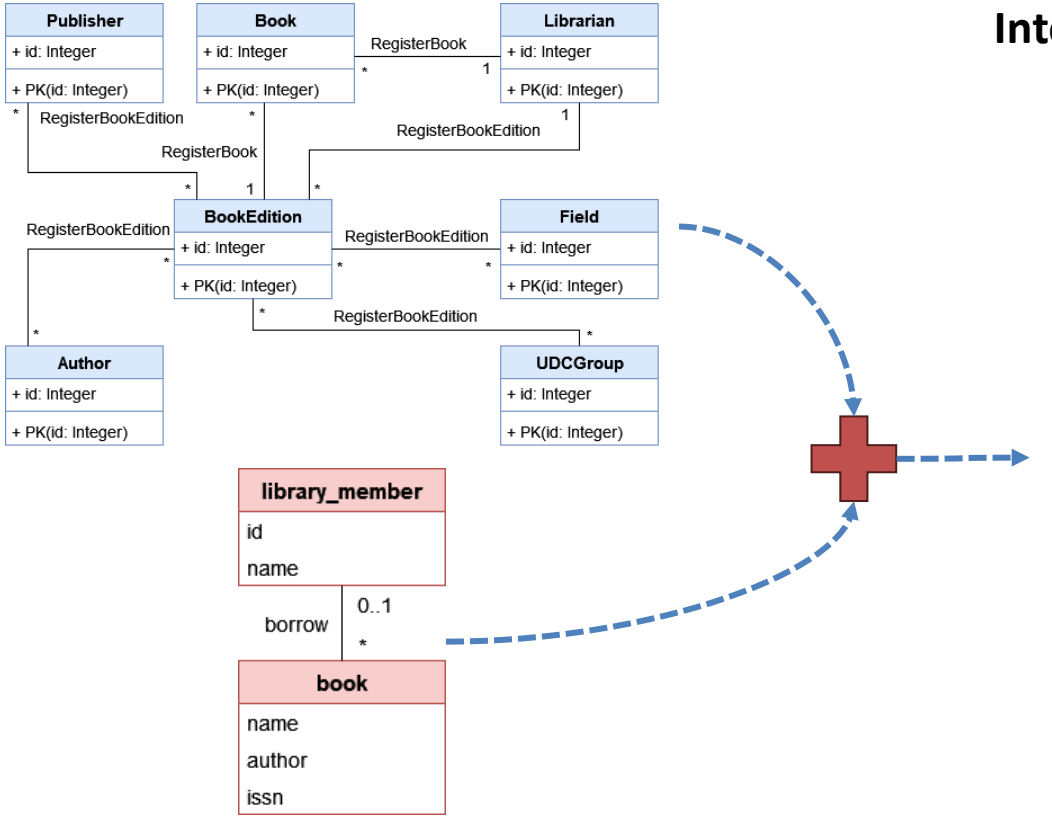
CDM generated by AMADEOS

Library member has id and name. Books have name, author and ISSN. Library member can borrow books.



CDM generated by TexToData

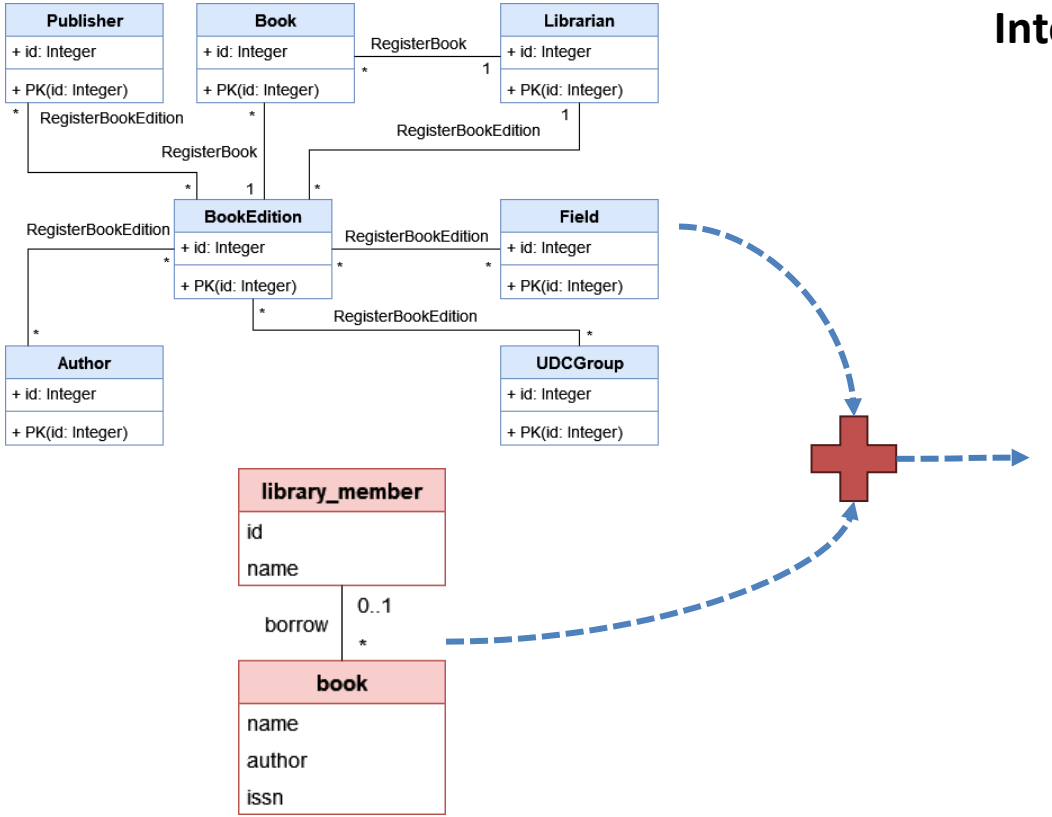
Approach Outline



Integration

- Conducted experiments suggest that CDM generated by AMADEOS is **more reliable** (more complete and more precise structure) than CDM generated by TexToData
- CDM generated by AMADEOS is used as the starting point (basis)

Approach Outline

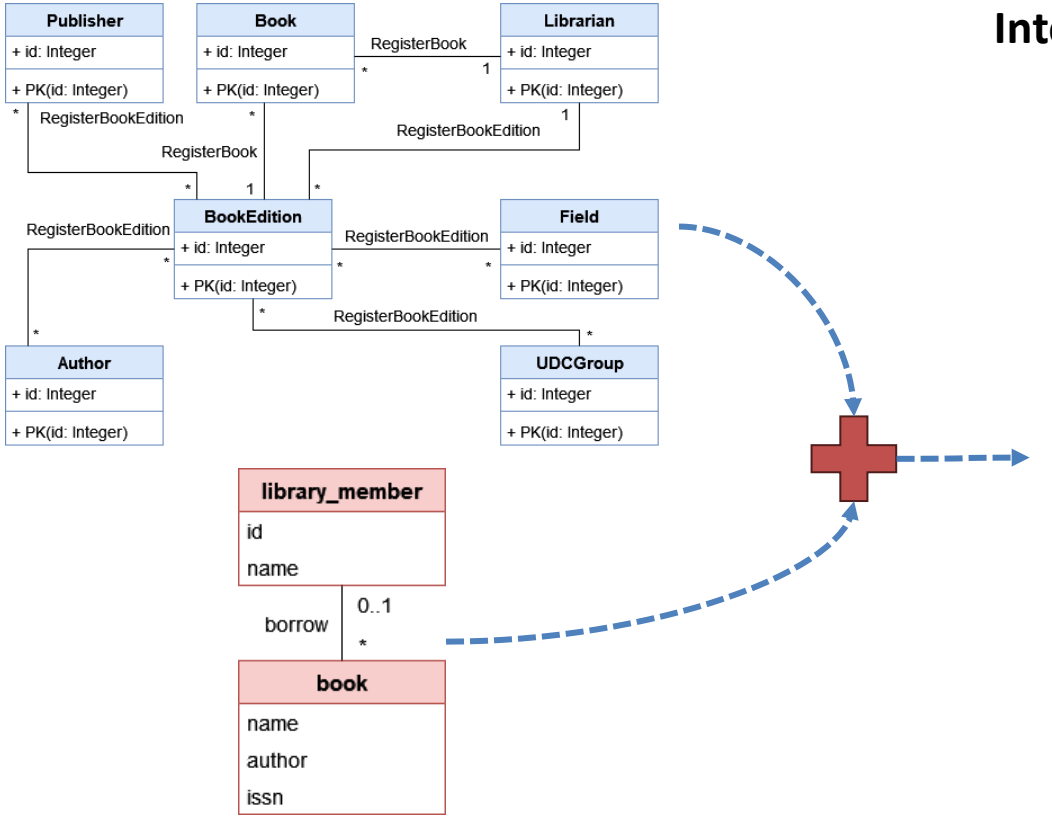


Integration

Classes

- Linguistic matching
- Recognition and resolution of semantic anomalies and typographical errors remains to be solved
- Structure-based matching will be part of the future improvement
- *Relevant* unmatched classes from both CDMs are kept in the final CDM
- Merging step – attributes from classes in the CDM generated by TexToData are added to the corresponding classes in the CDM generated by AMADEOS

Approach Outline

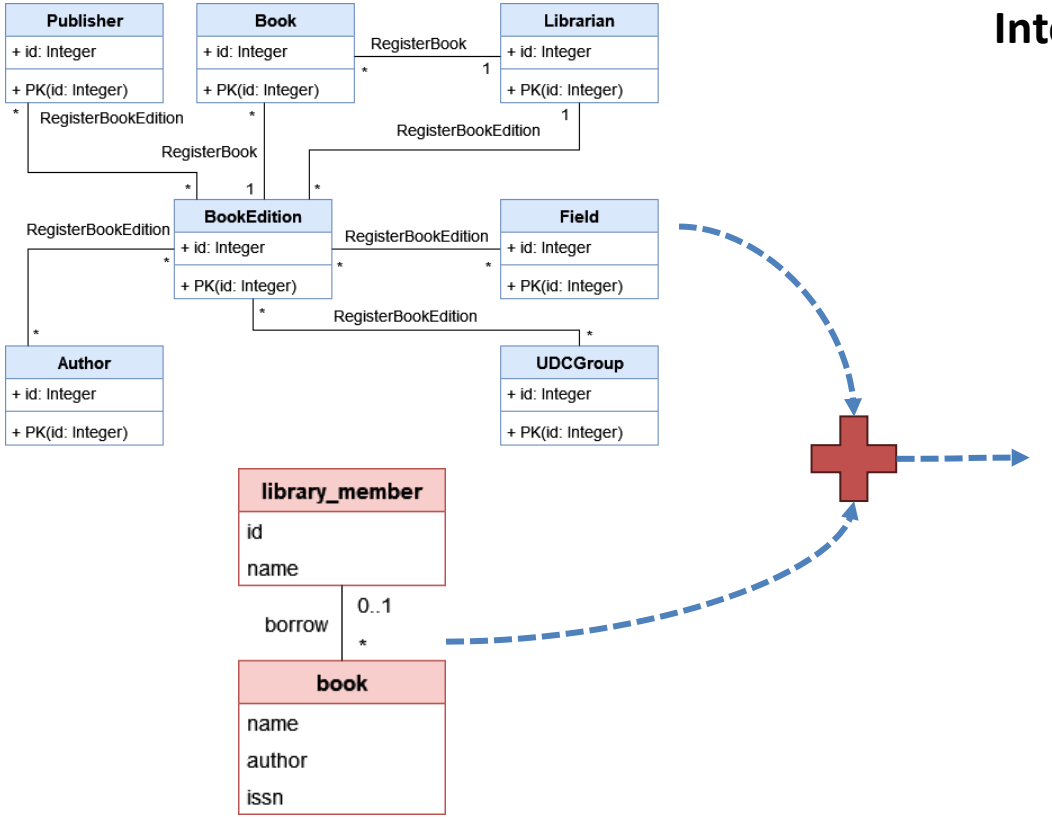


Integration

Associations

- CDM generated by AMADEOS is much more precise and much more complete than CDM generated by TexToData regarding associations
- In order to try to increase the completeness of the associations in the final CDM, currently we try to add associations from CDM generated by TexToData that are missing in the CDM generated by AMADEOS
- Structure-based matching
- Future work will include improvements of both tools and integration approach

Approach Outline

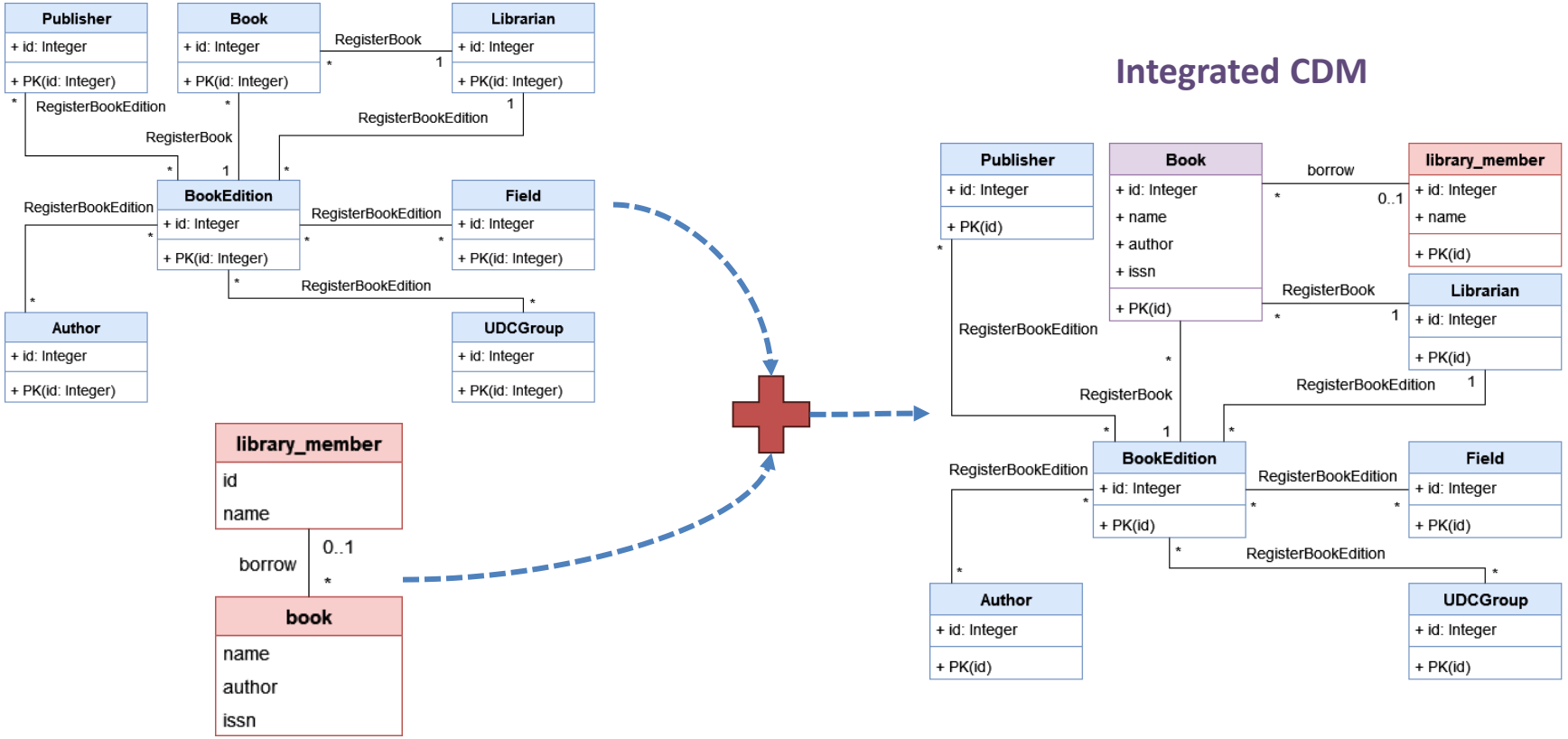


Integration

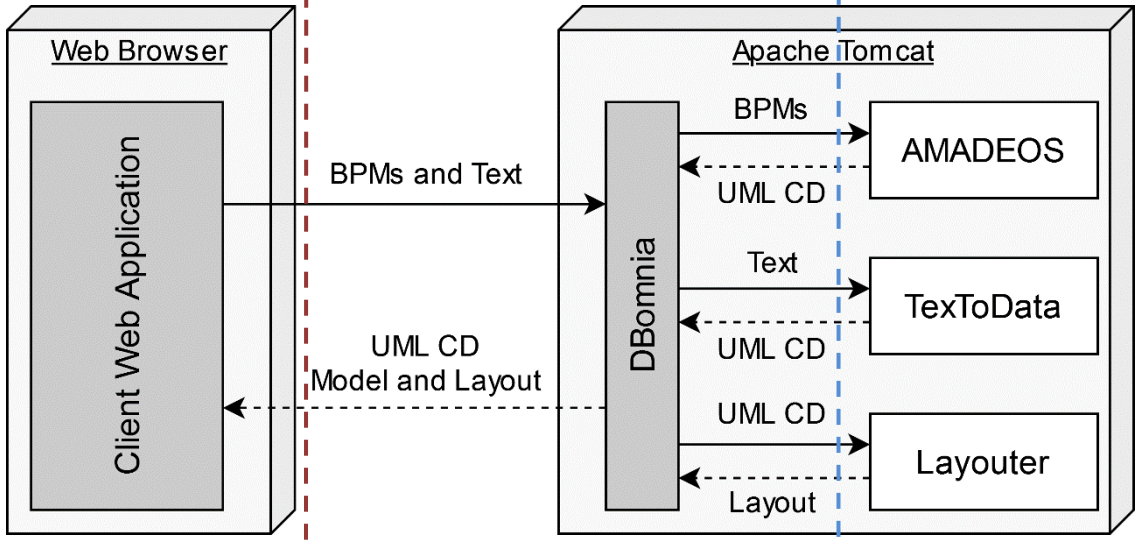
Generalizations

- AMADEOS does not generate generalizations
- Generalizations from CDM generated by TexToData are added to CDM generated by AMADEOS

Approach Outline



DBomnia – System Architecture



Client side

- GUI, files upload, text input, model manipulations, …

Server side

- Service-oriented Architecture
- The process of the CDM synthesis consists of four steps

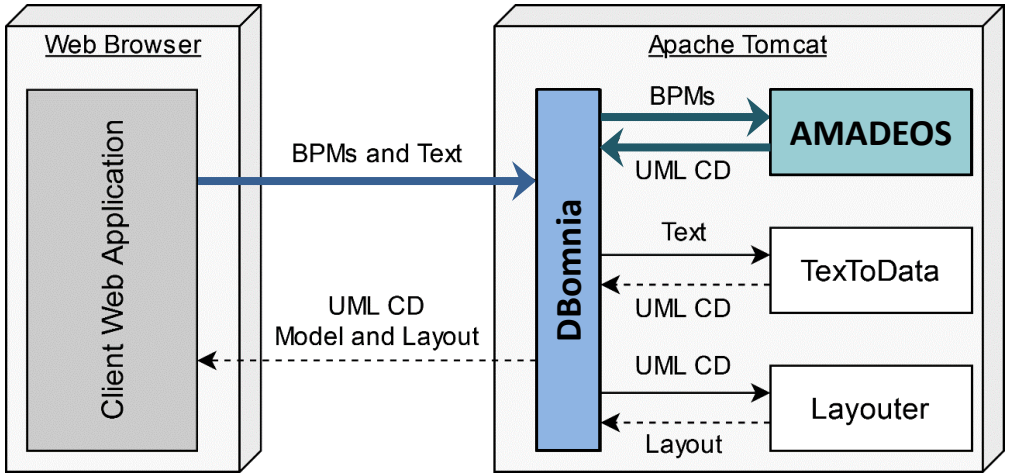
Publicly available services

- AMADEOS
- TexToData
- Layouter

DBomnia – CDM Synthesis Process

The process of the CDM synthesis consists of four steps:

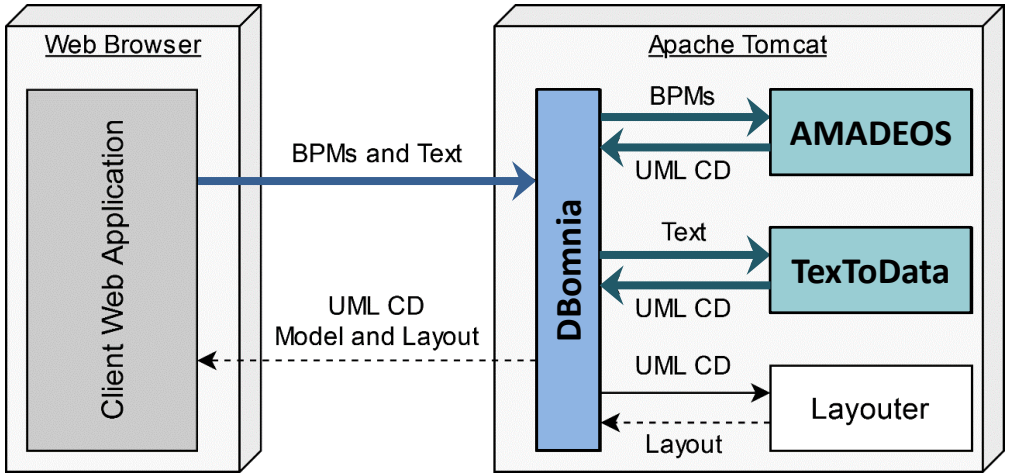
1. **Generation of the CDM from the source collection of BPMs**
   - The source **collection of BPMs** is sent to **AMADEOS**
   - **AMADEOS** generates the corresponding CDM and responds with the JSON object which contains **generated CDM**, execution status, etc.



DBomnia – CDM Synthesis Process

The process of the CDM synthesis consists of four steps:

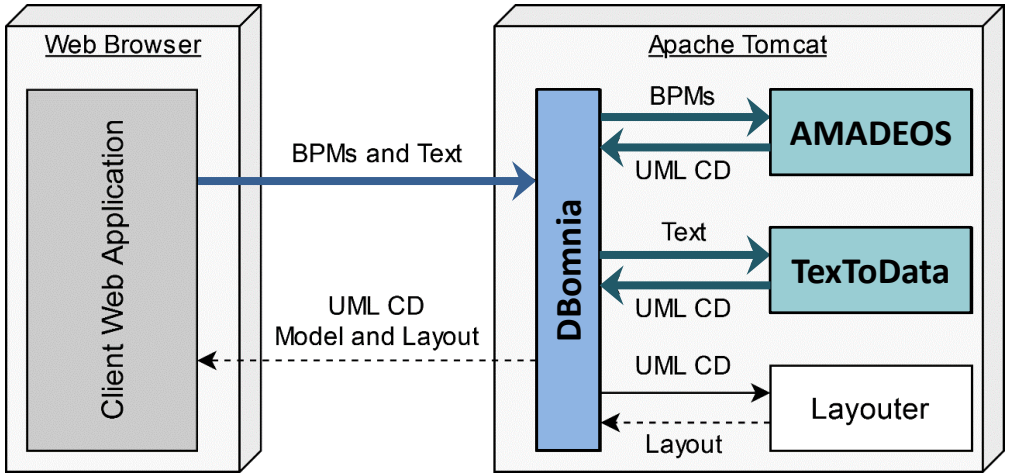
1. **Generation of the CDM from the source collection of BPMs**
   - The source **collection of BPMs** is sent to **AMADEOS**
   - **AMADEOS** generates the corresponding CDM and responds with the JSON object which contains **generated CDM**, execution status, etc.
2. **Generation of the CDM from the input textual specification**
   - The input **textual specification** is sent to **TexToData**
   - **TexToData** also generates the corresponding CDM and responds with the JSON object which contains **generated CDM**, error messages (if any), etc.



DBomnia – CDM Synthesis Process

The process of the CDM synthesis consists of four steps:

3. **Integration of the CDMs generated in the first two steps**



DBomnia – CDM Synthesis Process

The process of the CDM synthesis consists of four steps:

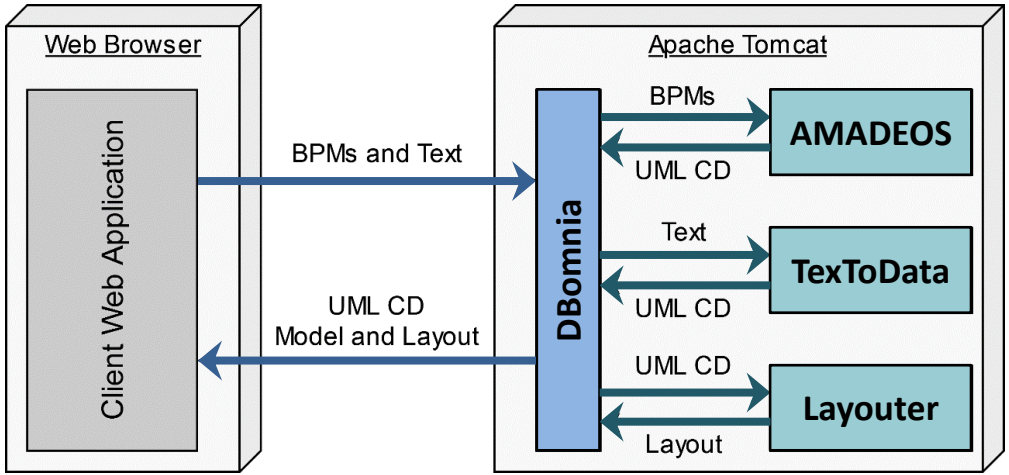
3. **Integration of the CDMs generated in the first two steps**

4. **Generation of the diagram layout for the integrated CDM**

   – The integrated **CDM** is sent to the **Layouter service**

   – **Layouter** is the pre-existing service (also used in **TexToData**) that provides the functionality of generating a diagram layout for the input UML class diagram

   – **Layouter** generates the corresponding diagram layout and responds with the **file** containing the generated layout

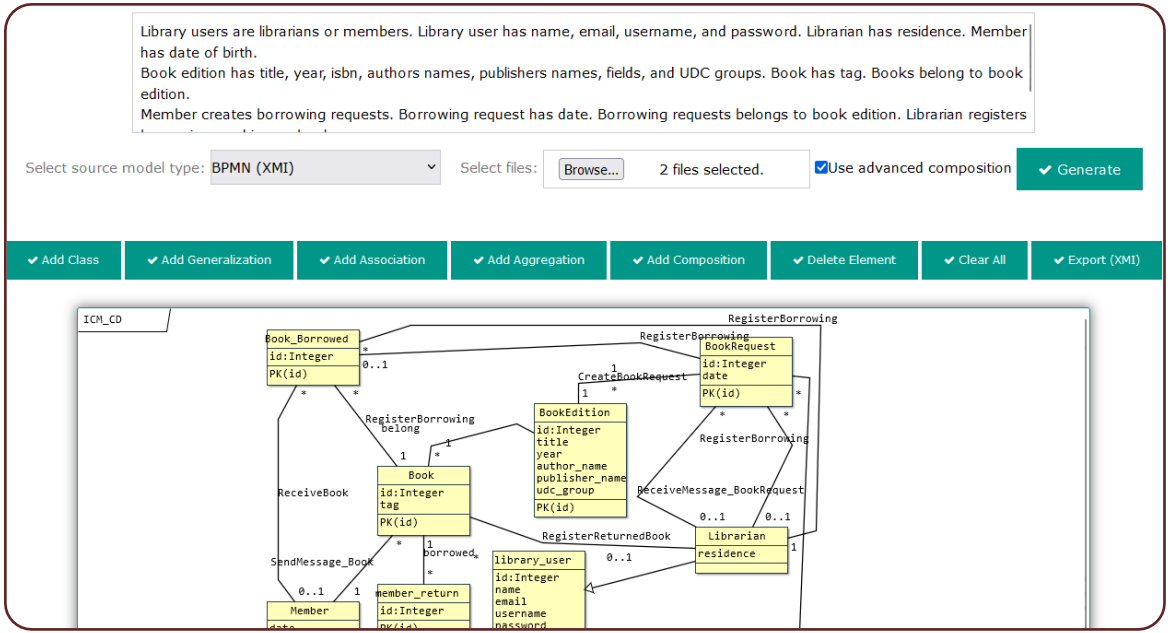


DBomnia – Client Side

Client web application

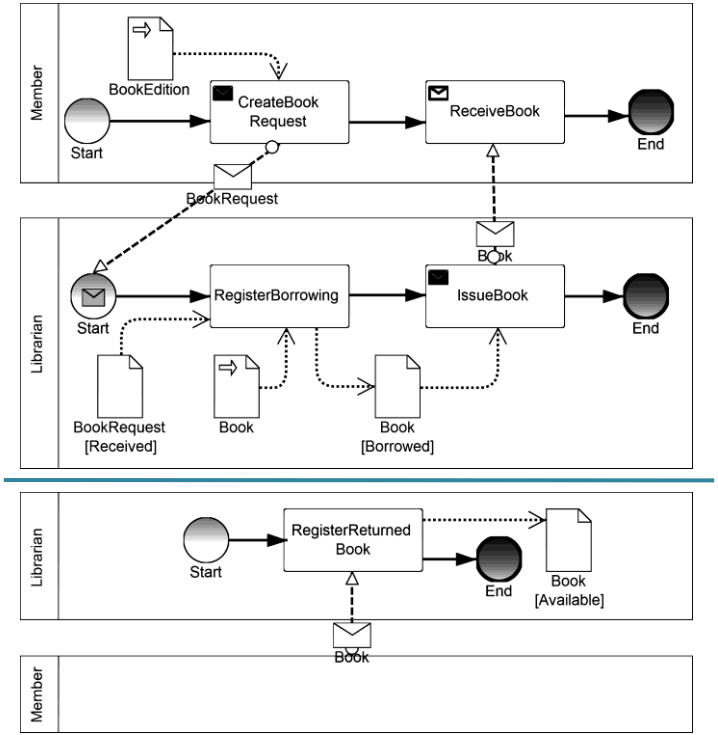
http://m-lab.etf.unibl.org:8080/dbomnia

- Allows users to input a textual specification and upload a collection of source BPMs
- Upon the user's request (click on the button *Generate*), all source artifacts are sent to **DBomnia**
- When the entire synthesis process is finished, the *client web application* receives the JSON response and visualizes the class diagram in the browser
- The visualized diagram is editable so users can additionally improve it



Illustrative Example

BPMs



Textual specification

Library users are librarians or members. Library user has name, email, username, and password. Librarian has residence. Member has date of birth.

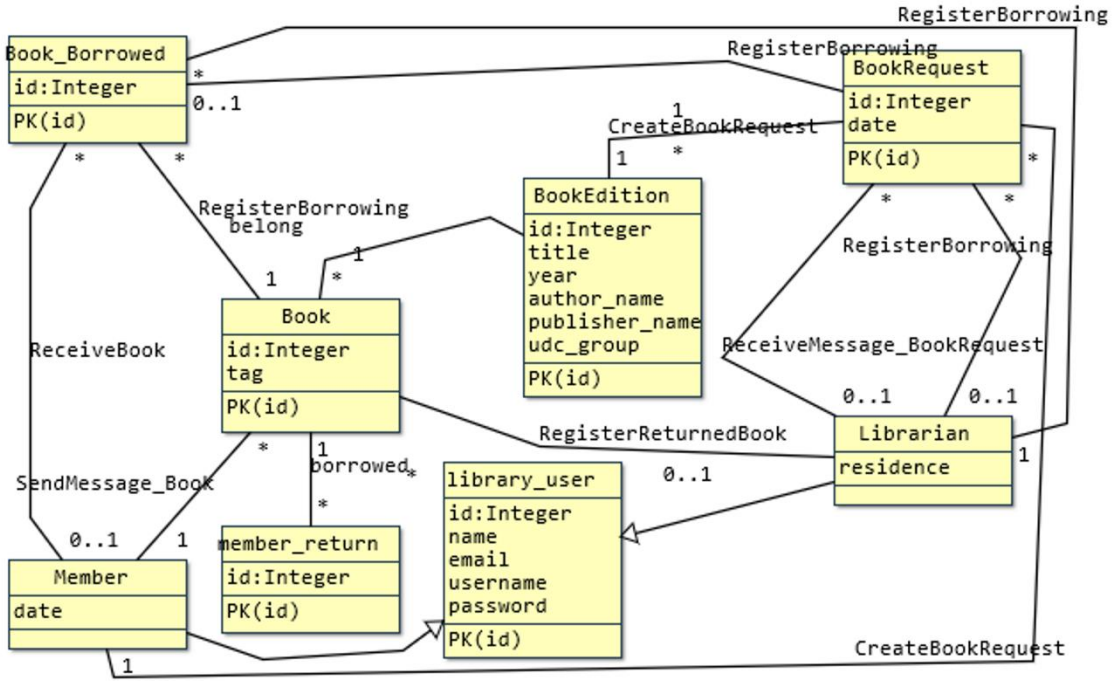
Book edition has title, year, isbn, authors names, publishers names, fields, and UDC groups. Book has tag. Books belong to book edition.

Member creates borrowing requests. Borrowing request has date. Borrowing requests belongs to book edition. Librarian registers borrowings and issues books.

Member returns borrowed book. Librarian registers returned book.

Illustrative Example

CDM generated by DBomnia

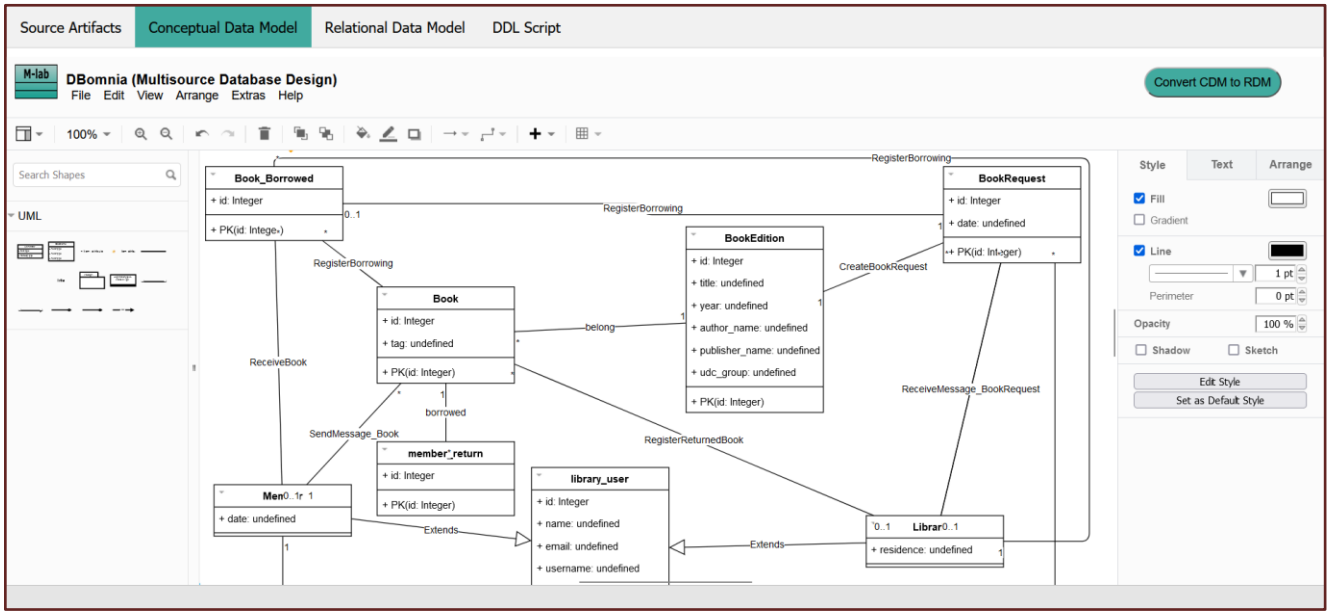


- **DBomnia** still does not generate 100% complete nor 100% correct target model
- However, the generated CDM also shows that the approach has **great potential** since the generated CDM is **more complete than each particular CDM** derived from the source artifacts of one single type

The Most Recent Improvements

Support for CDM design

- **Automatically generate** an initial CDM (from a collection of BPMs and textual specifications) which can be further improved
- **Import** an existing CDM from a file
- **Create** a new CDM from scratch



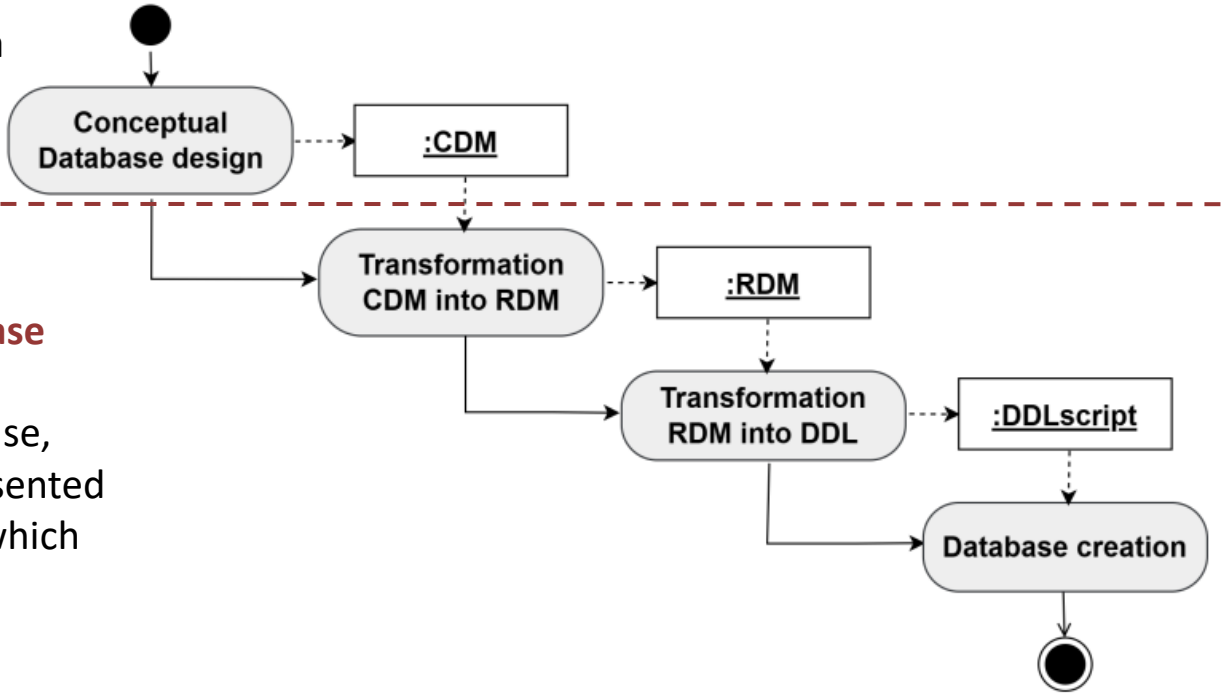
DBomnia has been significantly improved in UI and UX

(implementation of the client side is based on JavaScript and mxGraph library)

The Most Recent Improvements

Typical design process for relational DBs

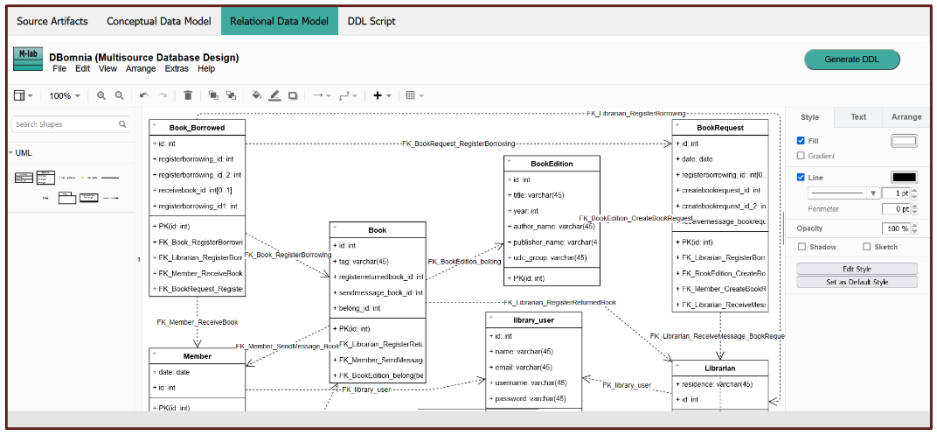
In the previous version, DBomnia supported CDM design only



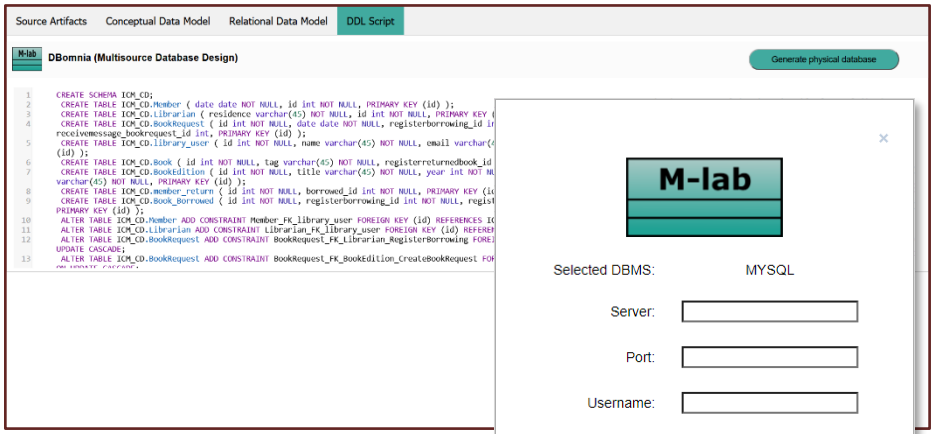
Now, **the coverage of the database design process is completed**, from CDM to the physical database, whereby RDM (as is CDM) represented by the standard UML notation (which eliminates the portability issues)

The Most Recent Improvements

CDM → RDM



RDM → DDL script



Generation of DB schema

Conslusion and Future Work

- In this paper, we presented **DBomnia** – the first online web-based tool enabling **automatic CDM derivation from a heterogeneous set of source artifacts**
- Currently supported source artifacts are **BPMs** and **textual specifications**
- **DBomnia** employs other tools to generate CDMs from specific source artifacts (**AMADEOS** derives CDM from BPMs, while **TexToData** derives CDM from textual specifications)
- Then, **DBomnia** **integrates** the generated CDMs into a single unified CDM
- The employed tools generate CDMs that are not 100% complete nor 100% correct, which makes them models with **reduced reliability**, and this constitutes the main challenge
- The presented tool represents an **early prototype** of the system, so a plethora of open issues should be resolved in the **future**:
  - Further improvements of the approach and tool
  - Further improvement of the specific CDM generators
  - The inclusion of other types of source artifacts
  - Further improvement of the UI and UX
  - Thorough validation and verification
  - …

ADBIS 2023

MADEISD
Modern Approaches in Data Engineering and Information System Design

Thank you!

Goran Banjac, Drazen Brdjanin, Danijela Banjac

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