МАТФ
Универзитет у Београду
Математички факултет

11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Развој апликације за интеграцију података из различитих извора о повезаности гена и болести

Лука Милошевић

доц. др Јована Ковачевић

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Опис проблема

- Креирање апликације под називом **GDA** - **Gene Disease Annotation** која за циљ има приказ и креирање анотацијске датотеке
- Анотацијска датотека садржи информације о повезаности гена и болести, и додатне информације о генима и болестима из различитих извора
- Значај анотацијске датотеке:
  - Помоћ приликом истраживања
  - Надомешћивање недостатака осталих база

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Структура анотацијске датотеке

- Осам колона које носе информацију о вези ген-болест

Ген	Ген-Болест	Болест
Gene Symbol	**Source**	**DOID**
Entrez ID		**Disease Name**
UniProt ID		**DOID Source**
Ensembl ID		

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Подаци

- Подаци представљају најбитнији део процеса креирања анотацијске датотеке
- Прикупљани су годинама од стране различитих организација и представљени као специјализоване базе података у различитим форматима (*.tsv*, *.csv*, *.obo*, *.dat*, *.xml* или *.txt*)
- Постоје два критеријума на основу којих се базе бирају:
  - **Ажурност базе** - база која није ажурирана више од годину дана не узима се у обзир приликом одабира
  - **Лиценца података у бази -** узимају се у обзир само базе података са отвореном лиценцом или базе које дозвољавају коришћење података у некомерцијалне сврхе
- Приликом израде анотацијске датотеке биће коришћено **тринаест** специјализованих база података које ће бити подељене на две групе, **изворне** и **помоћне** базе

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Изворне базе података

- Изворне базе користе се као основа тј. слогови који се налазе унутар изворних база формирају анотацијску датотеку
- За креирање анотацијске датотеке коришћено је **седам** изворних база: **DisGeNet**, **Cosmic**, **HumsaVar**, **Orphanet**, **ClinVar**, **HPO** и **Diseases**

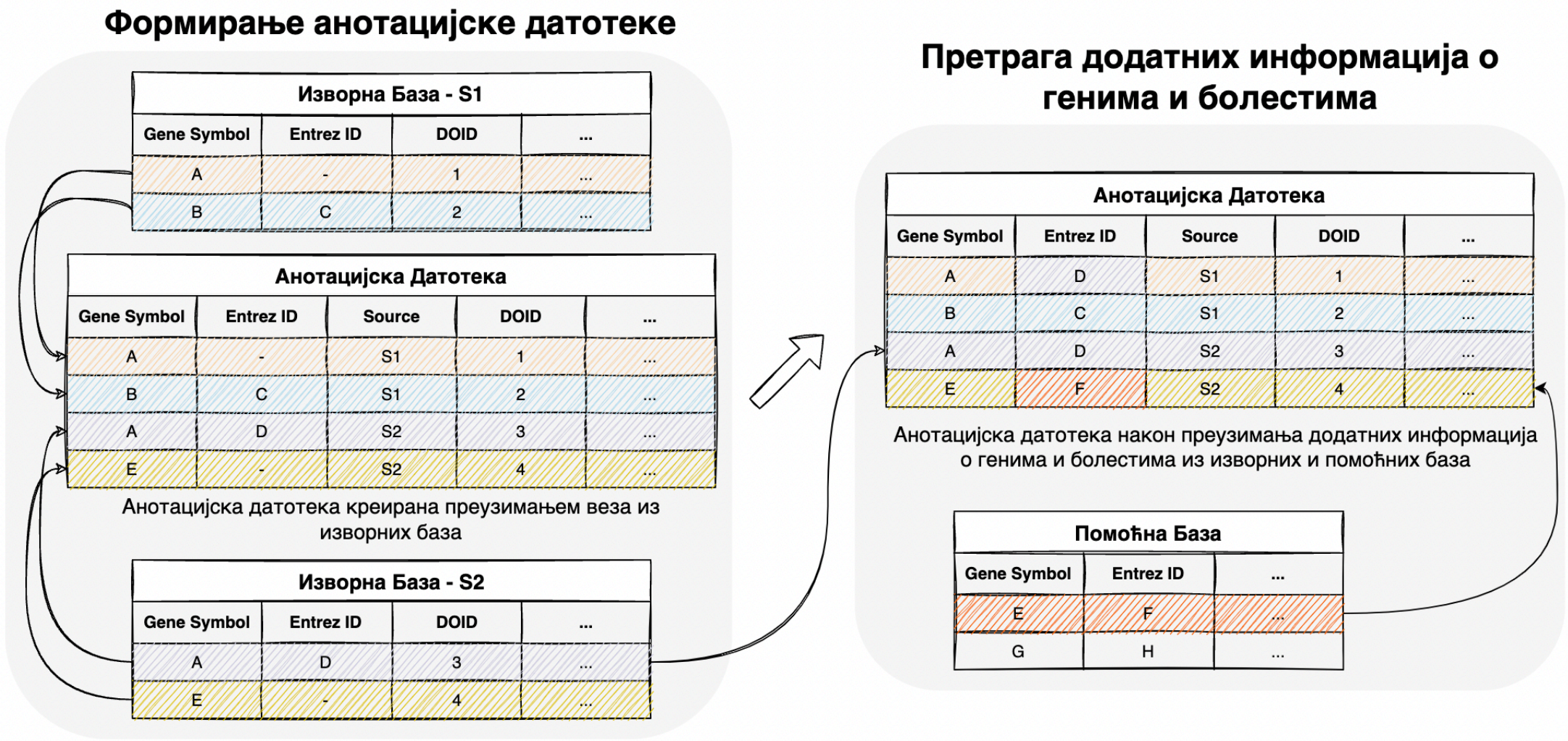
МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Помоћне базе података

- Анотацијска датотека као основу садржи везе ген-болест преузете из изворних база као и додатне информације о генима и болестима
- Помоћне базе представљају додатни ресурс изворним базама приликом претраге и преузимања додатних информација о генима и болестима
- Приликом креирања анотацијске датотеке коришћено је шест помоћних база: **Uniprot**, **HUGO**, **OBO**, **RGD**, **Orphanet Xref** и **Ensembl**

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

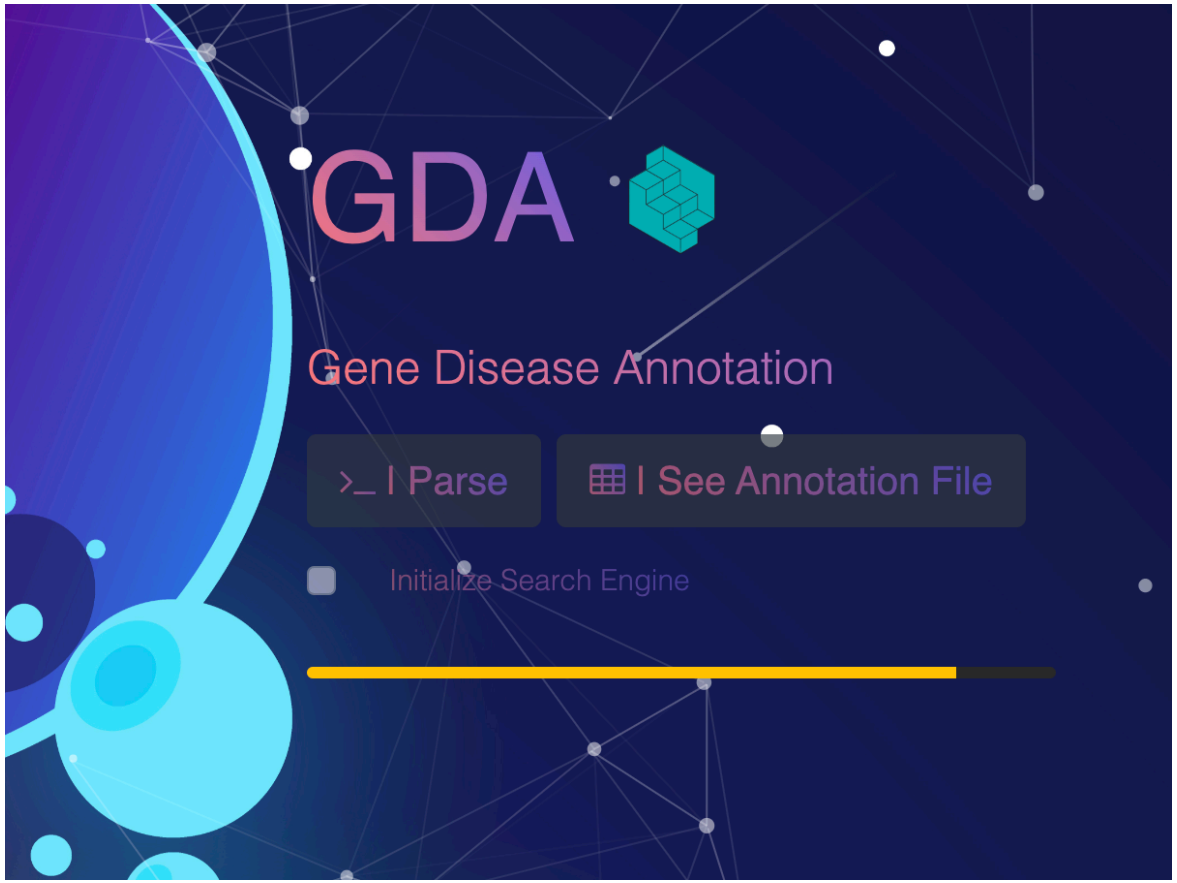
Поступак креирања анотацијске датотеке

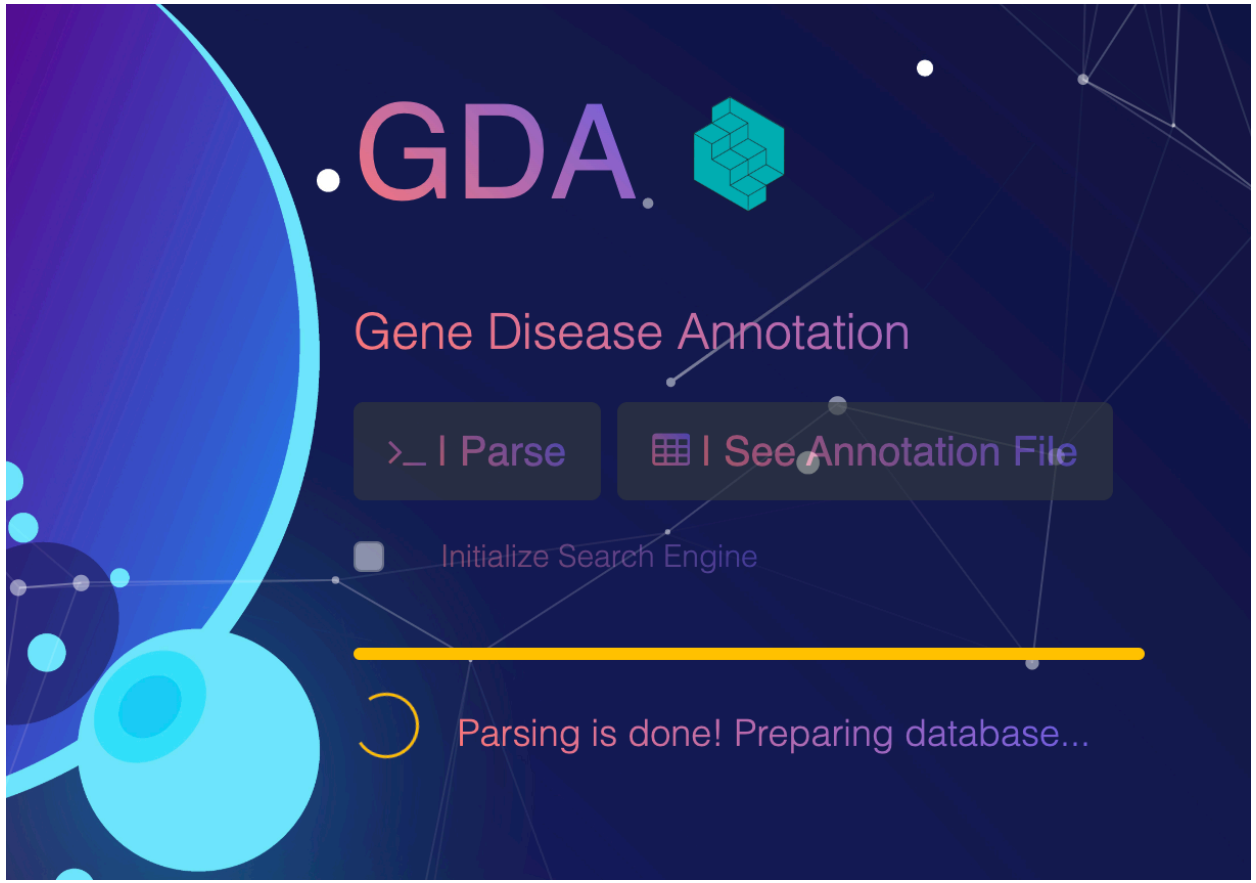


МАТФ
Универзитет у Београду
Математички факултет

11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Преглед апликације





МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Функционалности интерактивне табеле

- Приликом коришћења апликације кориснику је омогућено да врши различите манипулације са подацима на нивоу интерактивне табеле
- Неке од функционалности које апликација подржава су: сортирање, претрага целокупне анотацијске датотеке, претрага на нивоу атрибута, извоз у одређени формат (*PDF*, *Excel* и *CSV*) и штампање

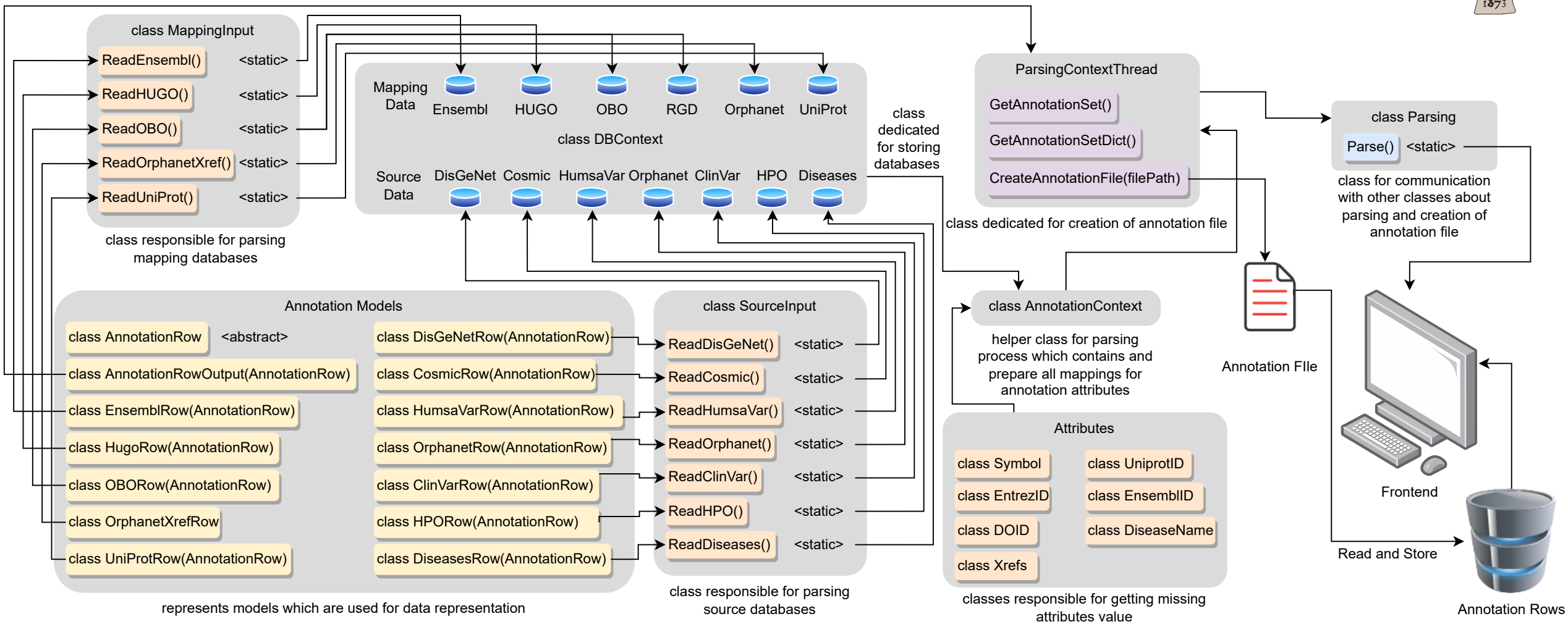
МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Серверски део апликације - креирање анотацијске датотеке

- Односи се на целокупан процес добијања анотацијске датотеке од почетног читања база до финалног парсирања и мапирања
- Имплементиран је у програмском језику **Python**
- Процес добијања анотацијске датотеке може се поделити на три фазе:
  1. Читање и парсирање изворних и помоћних база
  2. Претпроцесирање ускладиштених података
  3. Претрага недостајућих атрибута и креирање анотацијске датотеке

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Архитектура апликације



МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Читање и парсирање изворних и помоћних база података

- Базе су парсиране на **четири** различита начина у зависности од формата саме базе:
  1. Парсирање библиотеком **Pandas**
     - **.csv** и **.tsv** формати
     - **DisGeNet**, **COSMIC**, **ClinVar**, **Diseases**, **Ensembl** и **HUGO**
  2. Парсирање библиотеком **ElementTree**
     - **XML** формат
     - **Orphanet** и **Orphanet Xref**
  3. Парсирање библиотеком **pronto**
     - **.obo** формат
     - **OBO** и **RGD**
  4. **Неконвенцијално парсирање** - захтева парсирање специфичних формата база
     - **.txt** и **.dat** формати
     - **HumsaVar**, **HPO** и **UniProt**

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Претпроцесирање ускладиштених података

- Класа задужена за ову фазу креирања анотацијске датотеке назива се **AnnotationContext**
- У овој фази врши се инцијализација елемената неопходних за наредну фазу као што су:
  - Речници задужени за претрагу
  - Атрибутске класе које служе као интерфејс за претрагу сваког од атрибута анотацијске датотеке
  - Претраживач (**Typesense**) који се користи као једна од метода претраге атрибута **DOID**

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Претрага недостајућих атрибута и креирање анотацијске датотеке

- Класа **ParsingContextThread** задужена је за дохватање и претрагу недостајућих атрибута везе ген-болест
- Претрага је имплементирана на начин да користи паралелизацију и за сваку базу врши се у засебној нити
- Претрага се заснива на коришћењу атрибутских класа иницијализованих у претходној фази
- На крају претраге засебни резултати база се обједињују и тако добијени слогови чувају се у анотацијској датотеци на диску под називом **annotation_file.txt**

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

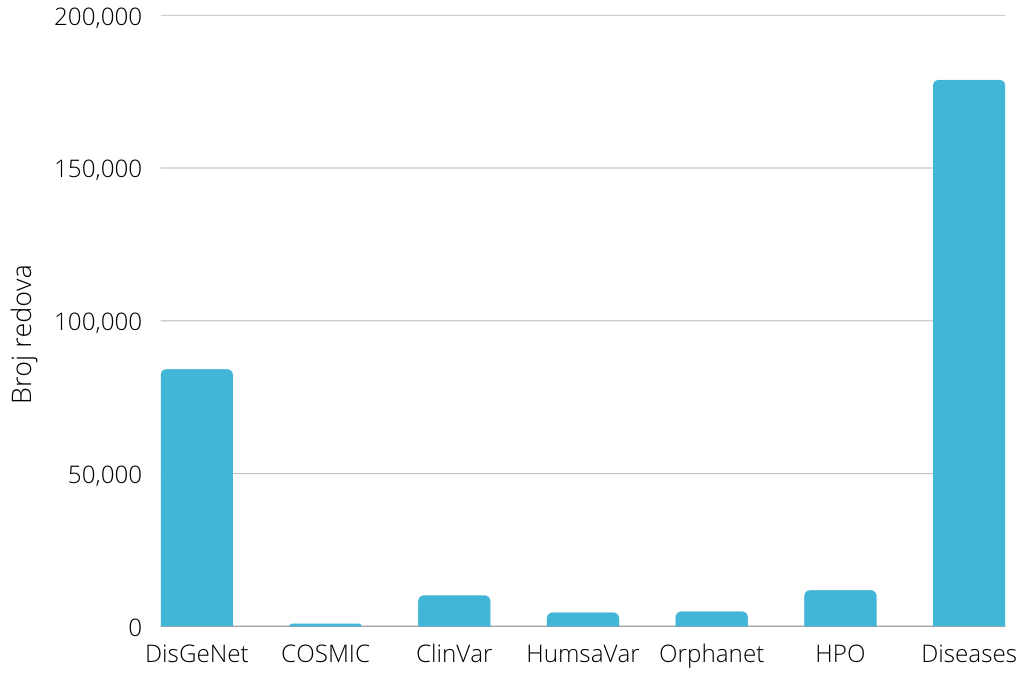
Интерфејс апликације

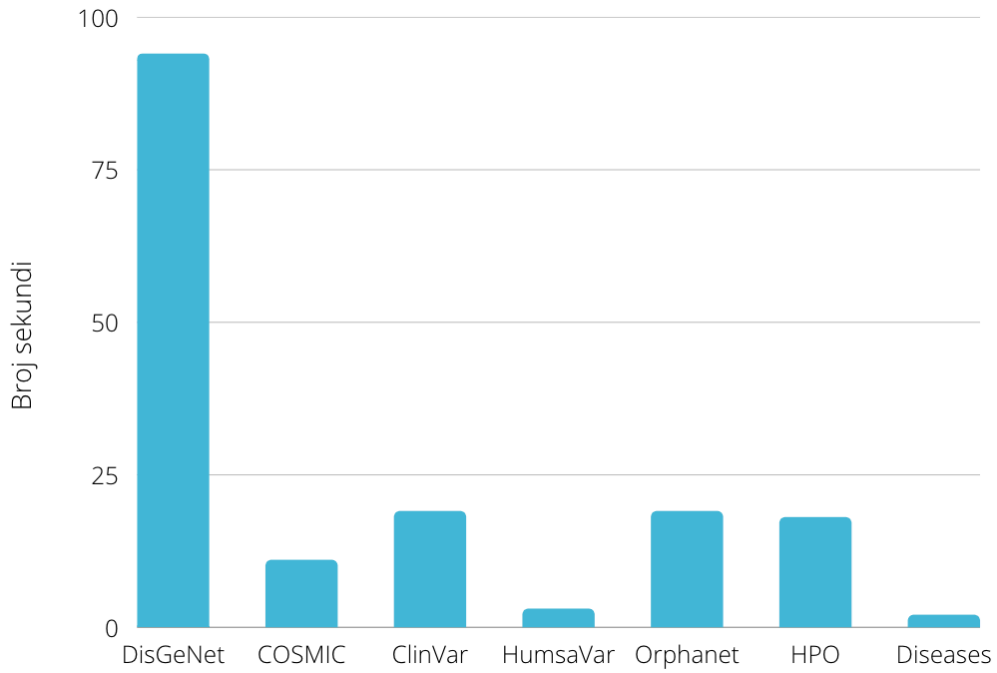
- Формат анотацијске датотеке у свом изворном облику као текстуална датотека може бити непогодна за коришћење, с тим у вези интерфејс апликације користи се за угоднији приказ анотацијске датотеке потенцијалним корисницима
- Представља веб интерфејс који се састоји од почетне странице и странице која садржи анотацијску датотеку приказану у форми интерактивне табеле
- Имплементиран је коришћењем **Django** развојног оквира, док је за интерактивну табелу коришћен додатак **DataTables** библиотеке **JQuery** програмског језика **JavaScript**

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Резултати извршавања

- Број ентитета који се користи приликом креирања анотацијске датотеке износи 862263
- Финално парсирање започиње са 294483 ентитета, а завршава са 294971





МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Резултати који се односе на прецизност атрибута **DOID**

- На крају процеса креирања анотацијске датотеке мери се прецизност атрибута **DOID**
- Проценат тачно пронађених атрибута **DOID** износи **97%**
- Проценат пронађених атрибута **DOID** на основу начина претраге:
  - Атрибут **DOID** је садржан унутар базе која се парсира **61%**
  - Коришћењем **Xref** вредности **20%**
  - Коришћењем **frozenset**-а **17%**
  - Коришћењем претраживача **Typesense 2%**

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Покретање и извршавање апликације

- Апликација је доступна за оперативне системе **Windows**, **Linux** и **MacOS**
- Портабилност апликације се постиже поступком докеризовања апликације
- За покретање апликације користе се две скрипте **GDA.sh** (*Linux* i *MacOS*) и **GDA.bat** (*Windows*)
- Наведене скрипте имају неколико опција за покретање и заустављање апликације, такође опције имају два облика (краћи и дужи)
- Пример команде за прво покретање апликације:
  - Windows: **GDA.bat -u**
  - Linux i MacOS: **./GDA.sh -up**

МАТФ
Универзитет у Београду
Математички факултет

11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Хвала на пажњи!

МАТФ
Универзитет у Београду
Математички факултет

11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Питања?