Систем за обраду података о заједничким радовима истраживача из Србије и иностранства

Лука Поткоњак

Вредновање научног рада

- Наукоментрија и библиометрија
- Цитираност
- *Impact factor*
- Мреже цитираности и мреже коауторства

Web of Science индексна база

- Преко 90 милиона записа о радовима
- Покрива 256 области науке
- Индексира радове објављене од 1900. године до данас
- Грешке…

WEB OF SCIENCE™

Формат података у *WOS*

Колоне искоришћене у раду и њихови примери

- C1 - Адреса аутора
  - [Jaksic, Smiljana] Univ Belgrade, Fac Forestry, Kneza Viseslava 1, Belgrade 11000, Serbia; [Maksimovic, Snjezana] Univ Banjo Luka, Fac Elect Engn, Patre 5, Banja Luka 78000, Bosnia & Herceg; [Pilipovic, Stevan] Univ Novi Sad, Fac Sci, Trg D Obradov 4, Novi Sad 21000, Serbia
- OI - ORCID идентификатор (Open Researcher and Contributor ID)
  - Gasic, Uros/0000-0001-5384-8396; Durovic, Sasa/0000-0003-2022-2447
- SC - Области истраживања
  - Mathematics; Zoology; Physics; Astronomy & Astrophysics; Geology…

Формат података у *WOS*

Додатне колоне погодне за прецизнију идентификацију

- PM – *PubMed* идентификатор *MEDLINE* базе биомедицинских чланака
- RI – *ResearcherID* је идентификатор истраживача специфичан за *WOS*
- SO – назив публикације, односно часописа
- ID, DE – кључне речи
- EM – адреса електронске поште аутора
- CR – референце

Честе грешке у подацима о радовима

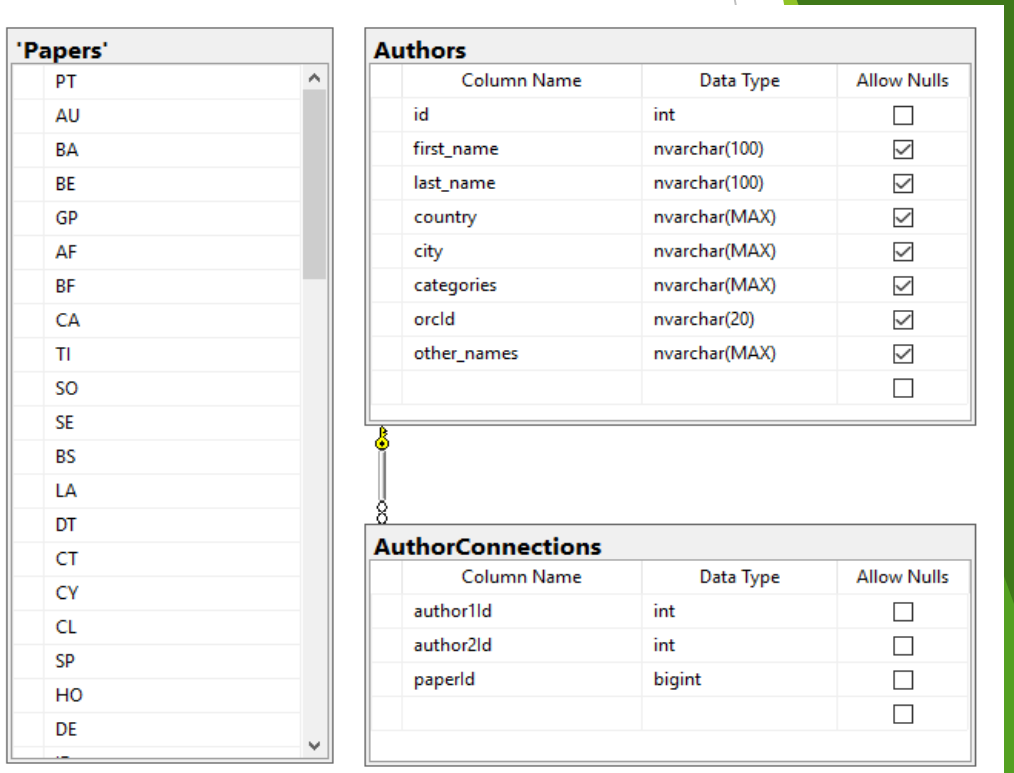
- Грешке при преписивању
- Нестандардизован запис имена
- Адресе веома варирају у формату
- Специфичности одређених језика
  - Редослед имена и презимена
  - Дифрафи
  - Ђ

Опис података

- Изабрани су *SCIE* (*The Science Citation Index Expanded*) радови који
  - У адреси имају Србију
  - Објављени су од 2014. до 2018. године
  - Категорија им је *Article* (чланак) или *Review* (преглед/ревизија)
- Избачени радови који имају само Србију у пољу са адресама
  - Резултат су сви радови који имају адресу бар једног аутора у Србији
- Избачени су сви *multiauthor* радови из резултата јер они теже да промене изглед и резултате мреже коауторства
- Уочени су додатни *multiauthor* радови који нису адекватно обележени, и они су ручно избачени из резултата

Апликација за идентификацију аутора

- C# и SQL Server
- Класе програма се мапирају на табеле базе
- Улазни подаци у *Excel* формату
- Regex издвајање адресе
  - (\[(.*?)\].*?,.*?,.*?(;|$))
- *Scoring* функција сличности аутора
  - Левенштајн даљина
  - Имена и иницијали
  - Категорије
  - Адреса



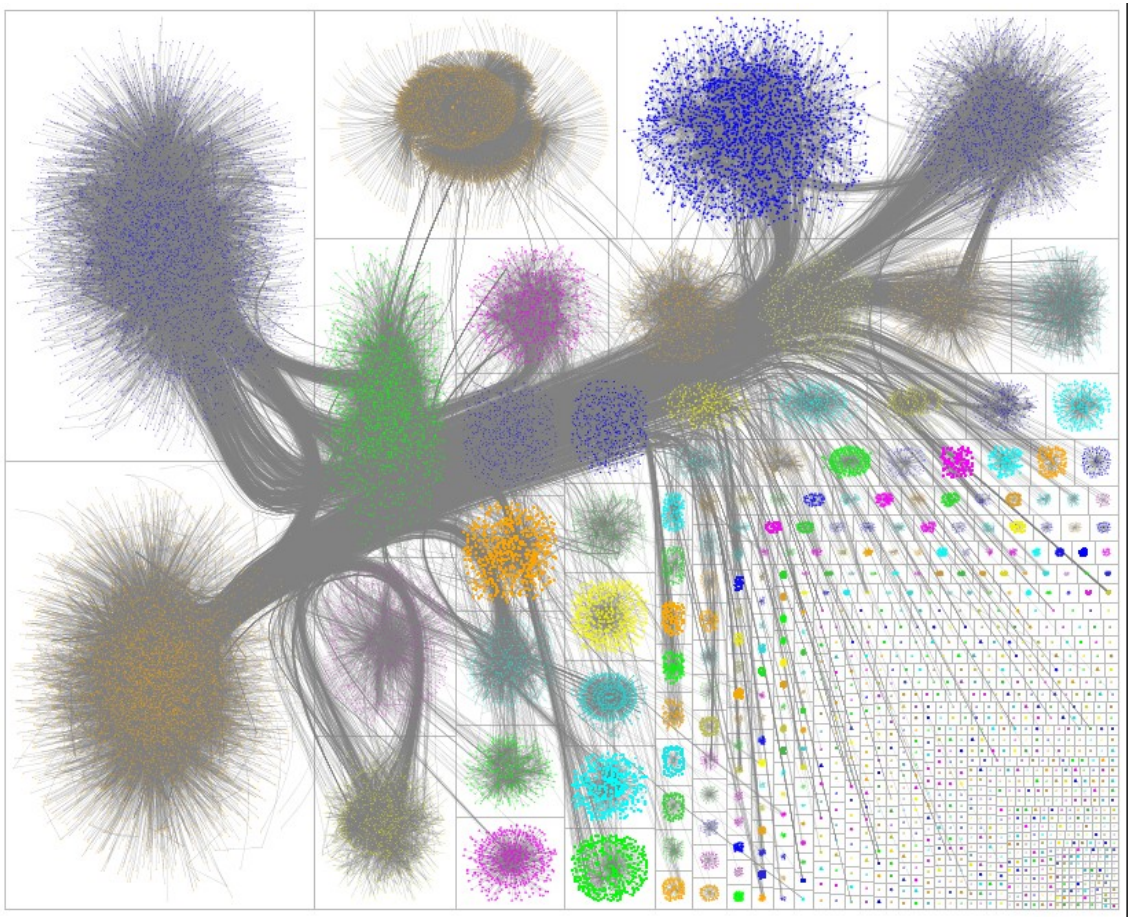
Апликација за идентификацију аутора

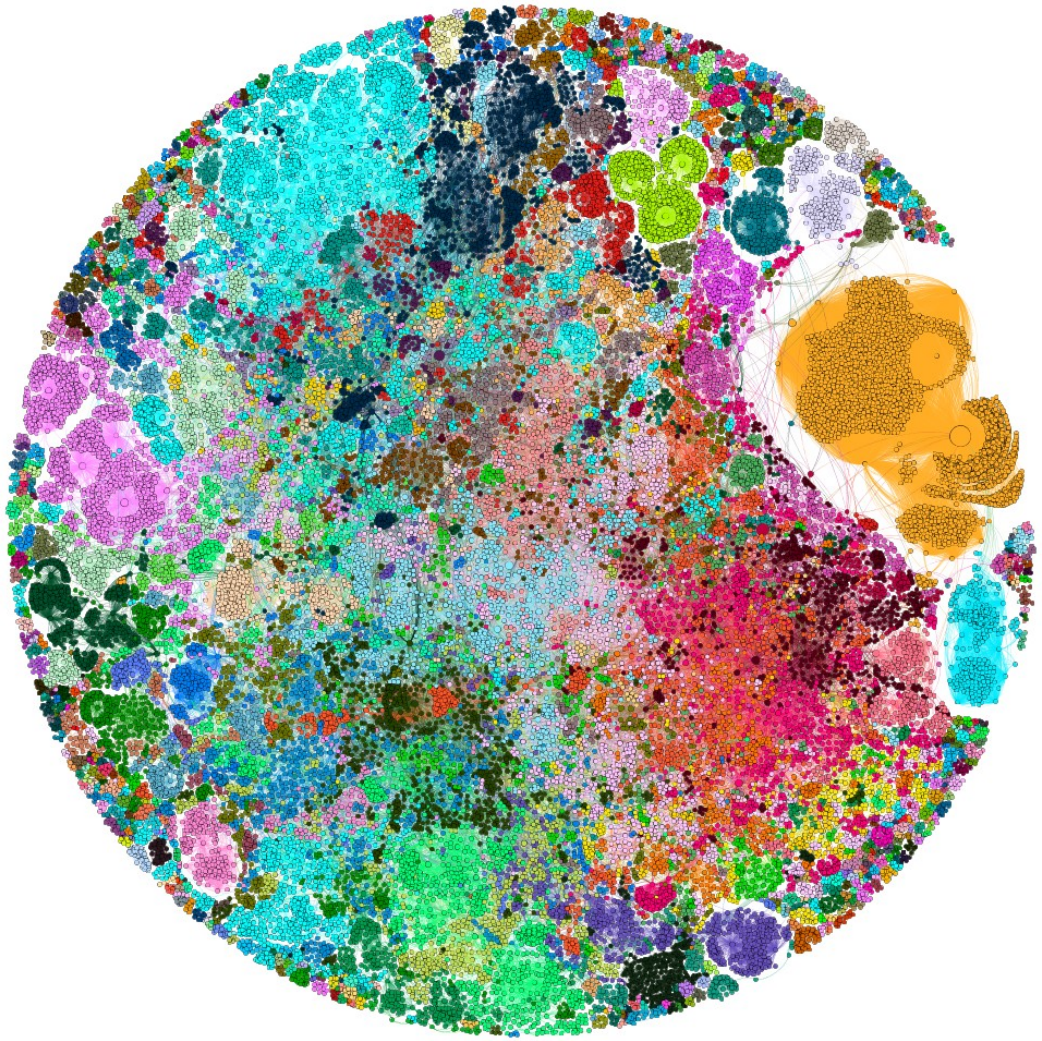
13.9.2018. 20:41:31 Importing Excel to DB
13.9.2018. 20:41:31 Excel imported to DB
13.9.2018. 20:41:31 Generating data from DB
13.9.2018. 20:41:34 Bad input count: 41
13.9.2018. 20:41:34 One name count: 3
13.9.2018. 20:41:34 No WOS count: 0
13.9.2018. 20:41:34 **Multi institution count: 12497**
13.9.2018. 20:41:34 Data extracted from DB
13.9.2018. 20:41:34 Different name authors count: 44426
13.9.2018. 20:41:34 Total authors count: 77631
13.9.2018. 20:41:34 Merging same name Authors
13.9.2018. 20:45:53 Merged authors count: 44426/44426
13.9.2018. 20:45:53 Same name authors merged
13.9.2018. 20:45:53 Different name authors count: 44426
13.9.2018. 20:45:53 Total authors count before same name merge: 77631
13.9.2018. 20:45:53 Total authors count after same name merge: 51025
13.9.2018. 20:45:53 **Same name authors merged count: 26606**
13.9.2018. 20:45:53 Duplicate name authors left count: 6599
13.9.2018. 20:45:53 Merging same OrcId authors
13.9.2018. 20:46:03 Same OrcId authors merged
13.9.2018. 20:46:03 Total authors count after same OrcId merge: 49933
13.9.2018. 20:46:03 **Same OrcId merged: 1092**
13.9.2018. 20:46:03 Merging similar name Authors
13.9.2018. 20:55:28 Scores calculated count: 49932/49932
13.9.2018. 20:55:38 Iteration: 0 Current max score: 9; Duplicate authors merged count: 3
13.9.2018. 20:55:44 Iteration: 1 Current max score: 8; Duplicate authors merged count: 5
13.9.2018. 20:55:51 Iteration: 2 Current max score: 7; Duplicate authors merged count: 42
13.9.2018. 20:56:02 Iteration: 3 Current max score: 6; Duplicate authors merged count: 275
13.9.2018. 20:56:07 Iteration: 4 Current max score: 7; Duplicate authors merged count: 3
13.9.2018. 20:56:13 Iteration: 5 Current max score: 6; Duplicate authors merged count: 7
13.9.2018. 20:56:35 Iteration: 6 Current max score: 5; Duplicate authors merged count: 944
13.9.2018. 20:56:41 Iteration: 7 Current max score: 6; Duplicate authors merged count: 1
13.9.2018. 20:56:47 Iteration: 8 Current max score: 5; Duplicate authors merged count: 9
13.9.2018. 20:58:40 Iteration: 9 Current max score: 4; Duplicate authors merged count: 5962
13.9.2018. 20:58:47 Iteration: 10 Current max score: 6; Duplicate authors merged count: 1
13.9.2018. 20:58:53 Iteration: 11 Current max score: 5; Duplicate authors merged count: 9
13.9.2018. 20:59:00 Iteration: 12 Current max score: 4; Duplicate authors merged count: 37
13.9.2018. 20:59:01 Similar name authors merged
13.9.2018. 20:59:01 Total authors count after similar name merge: 42635
13.9.2018. 20:59:01 **Similar names merged: 7298**
13.9.2018. 20:59:01 Commiting to DB
13.9.2018. 20:59:15 Commited 42635 out of 42635 authors
13.9.2018. 20:59:15 Authors commited
13.9.2018. 20:59:42 Commited 4275346 out of 4275346 connections
13.9.2018. 20:59:42 Connections commited
13.9.2018. 20:59:42 Committed to DB

Визуализација мреже аутора

- Анализа социјалних мрежа
  - Степен
  - Кластери
  - Централност
- *Gephi* и *NodeXL*
- Мрежа коауторства и оптимизације

Визуализација мреже аутора





Визуализација мреже аутора

Број коаутора по раду	Број радова	Број аутора	Број веза коауторства	Број веза коауторства приказаних у графу
Сви	8945	42721	4275346	158246
До 100	8907	36984	462913	134405
До 50	8839	34058	320362	126412
До 20	8572	28316	187025	111946
До 7	6326	16426	63660	49091

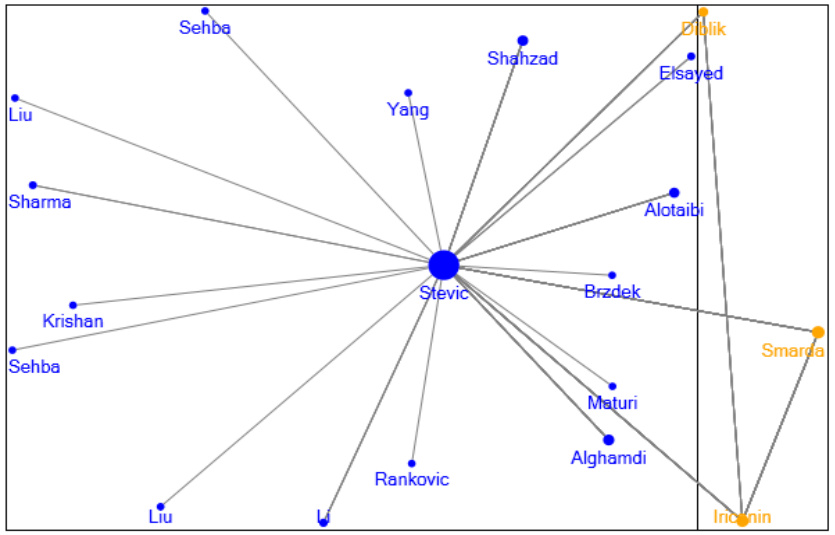
Мрежа коауторства за радове са 7 или мање аутора:

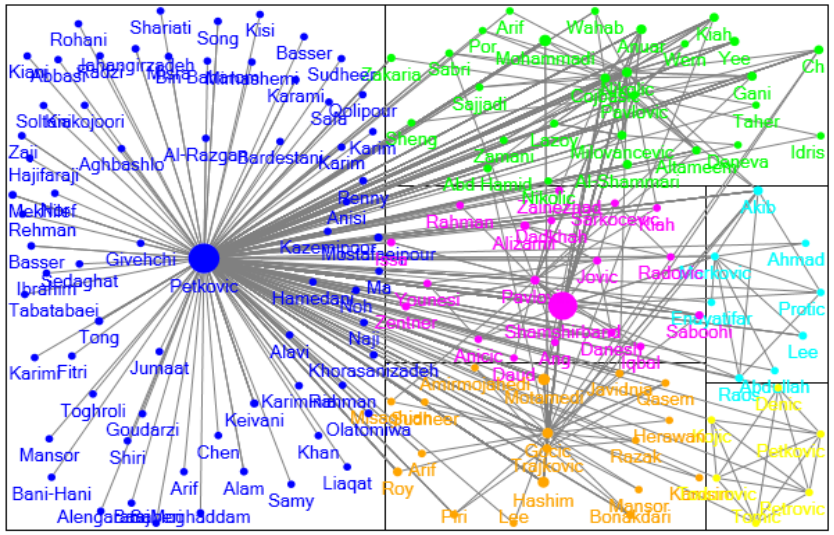
Анализа резултата

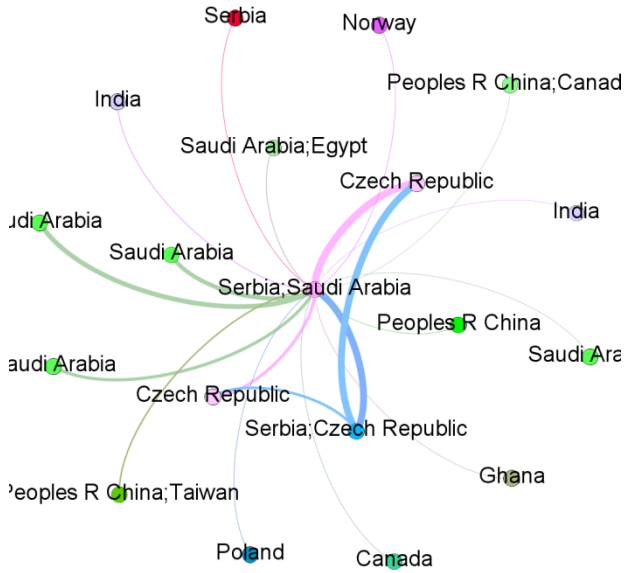
- Најпродуктивнији аутори из Србије са не више од 7 аутора по раду у периоду 2014.-2018.

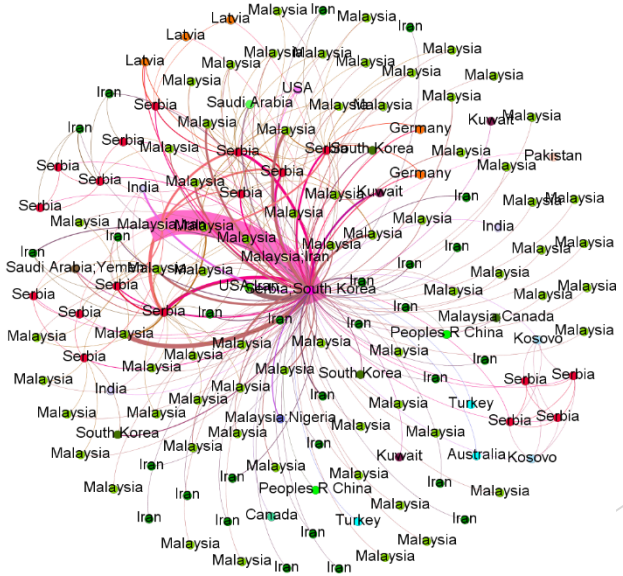
Име	Презиме	Број радова	Просечан број коаутора	Betweenness centrality
Dalibor	Petkovic	96	4,65	1380391,753
Ivan	Gutman	88	0,82	1226144,408
Stevo	Stevic	61	1,18	177210,500
Stojan	Radenovic	54	2,46	2159771,345
Milivoj	Belic	52	3,21	1181591,686

Анализа резултата – Его мреже









Закључак

- Основи библиографије
- Преглед *Web of Science* базе
- Апликација за идентификацију и дедупликацију имена аутора
  - Број дупликата је значајно смањен
  - Постоји велики простор за унапређење апликације, како хеуристиком, тако и машинским учењем
- Мрежа коауторства
  - Визуализације
  - Анализа