Oznaka poročila: ARRS-CRP-ZP-2013-02/7

ZAKLJUČNO POROČILO CILJNEGA RAZISKOVALNEGA PROJEKTA

A. PODATKI O RAZISKOVALNEM PROJEKTU

1.Osnovni podatki o raziskovalnem projektu

Šifra projekta	V5-1015
Naslov projekta	Atlas slovenske znanosti
Vodja projekta	12570 Dunja Mladenić
Naziv težišča v okviru CRP	2.01.01 Uporaba informacijskih sistemov za povečanje učinkovitosti razvojno-raziskovalnega dela s poudarkom na znanstveno-raziskovalni dejavnosti
Obseg raziskovalnih ur	1960
Cenovni razred	C
Trajanje projekta	10.2010 - 03.2013
Nosilna raziskovalna organizacija	106 Institut "Jožef Stefan"
Raziskovalne organizacije - soizvajalke	
Raziskovalno področje po šifrantu ARRS	2 TEHNIKA 2.07 Računalništvo in informatika
Družbeno-ekonomski cilj	13.01 Naravoslovne vede - RiR financiran iz drugih virov (ne iz SUF)

2.Raziskovalno področje po šifrantu FOS[1]

Šifra	1.02
- Veda	1 Naravoslovne vede
- Področje	1.02 Računalništvo in informatika

3.Sofinancerji

	Sofinancerji		
1.	Naziv		
	Naslov		

B. REZULTATI IN DOSEŽKI RAZISKOVALNEGA PROJEKTA

4.Povzetek raziskovalnega projekta[2]

SLO

Glavni cilj projekta je vzpostavitev enotnega sistema za enostaven dostop do razvojno-raziskovalnih podatkov s ciljem spodbujanja novih idej, sodelovanja med domačimi institucijami in industrijo, promocije znanstveno-raziskovalnih dosežkov doma in v tujini ter vzpostavljanje okolja inovativne in kreativne kulture. Razvit sistem je dostopen kot spletna informacijska točka (spletni portal) na osnovi podatkov in baz, ki beležijo dosežke slovenskih raziskovalcev. S tem omogočimo odprt dostop do razvojno raziskovalnih javnih podatkov. Ob tem spletna informacijska točka ponuja vrsto inovativnih storitev za celovito empirično analizo slovenske znanosti skozi vrsto analitskih orodij za analizo dogajanja v slovenskem razvojno raziskovalnem okolju, vključno s kompetenčnimi grafi in sodelovalnimi grafi. S tem ponudimo orodja za pomoč industriji in raziskovalnim organizacijam pri iskanju potencialnih partnerjev, potrebnih kompetenc in rezultatov. Podakti o dosežkih slovenskih raziskovalcev so pridobljeni z integriracijo podaktov iz SICRIS baze podatkov o raziskovalcih, COBISS baze podatkov o objavah raziksovalcev, Videolectures.net baze video predavanj in dosežkov domačih in tujih znanstvenikov.

Raziskovalna komponenta projekta vsebuje vrsto problemov, ki so trenutno aktualni v raznih raziskovalnih področjih, ki se ukvarjajo z analizo in obdelavo podatkov. Glavni raziskovalni izzivi so: (1) povezovanje raznovrstnih baz podatkov (ugotavljanje identitete podatkovnih objektov), (2) prikaz (vizualizacija) in interaktivna analiza (angl. OLAP) kompleksnih podatkov kot so omrežja, besedila, časovni razvoj, (3) učinkovitost (skalabilnost) metod za analizo velikih količin podatkov.

Za potrebe diseminacije in motivacije smo pripravili krajše promocijske filme, ki promovirajo znanstveno-raziskovalne dosežke.

ANG

Goal of the project is establishment of a system for easy access to data on Slovenian researchers to stimulate new ideas, collaboration between research and industry, promotion of scientific results. The developed system is accessible as Web portal integrating data from several databases on Slovenian researchers. This enables easy access to publicly available data on activity of Slovenian researchers. At the same time the Web portal offers a number of innovative services for empirical analysis of Slovenian research space via analytics tools including competence graphs and collaboration graphs. The project offers tools to support industry and research organizations in searching for partners, the needed competences and results. The data on Slovenian researchers space is obtained via integration of SICRIS database on Slovenian researchers, COBISS database of publications of Slovenian researchers, Videolectures.net databse of video recordings of lectures and achievement so of national and international researchers.

The research component of the project addresses a number of problems from different research areas dealing with data analysis. The main research challenges are the following: (1) connecting relevant databases, (2) data visualization and interactive data analysis of complex data including text, social networks, time tagged data, (3) scalability of methods for data analysis

Short promotional videos of research and development results will serve dissemination purpose of the project and its results.

5.Poročilo o realizaciji predloženega programa dela na raziskovalnem projektu[3]

Delo na projektu je potekalo po delovnih sklopih (DS), kot je bilo planirano v programu dela projekta.

DS1 Podrobne zahteve in specifikacija sistema

1. Naredili smo analizo zahtev in pričakovanj uporabnika, obstoječega stanja v Sloveniji in pa obstoječih treh sistemov, ki jih bomo povezali v skupno točko.

2. Pripravljena analiza nam je predstavljala osnovo za pripravo in definiranje podrobnih zahtev za izgradnjo sistema. Na osnovi teh zahtev smo izdelali specifikacijo sistema enotne točke. Specifikacija predstavlja temeljni dokument za izgradnjo sistema in zajema vse potrebne tehnične aspekte.

Opisano v R1.

DS2 Uvoz in konsolidacija podatkov

1. Opredelili smo možne vire podatkov in implementirali CERIF podatkvoni model. Opisano v R2.1

2. Razvili smo potrebne module za enosmerni uvoz podatkov iz definiranih treh virov: SICRIS, IST-World, Videolectures. Rezultat te naloge je programski modul za uvoz podatkov. Opisano v R2.1

3. Izdelali smo modula za za filtriranje in čiščenje podatkov s katerim bomo lahko filtrirali in čistili podatke, kjer je potrebno z avtomatskimi metodami detektirati in izločati anomalije v podatkih. Rezultat naloge je programski modul za filtriranje in čiščenje podatkov. Opisano v R2.1

4. Na osnovi prečiščenih podatkov smo razvili metode za povezovanje podatkovnih objektov iz treh zgoraj navedenih virov. Rezultat je programski modul za povezovanje podatkovnih objektov. Opisano v R2.2

5. Izdelali smo modula za izvoz podatkov iz centralnega repositorija ATLASa. Rezultat je programski modul, ki generira XML zapis.

DS3 Analiza in agregacija podatkov

1. Modelirali smo podatke o raziskovalcih in njihovih sodelovanjih na projektih z uporabo metod za analizo omrežja. Poudarek je na izračunu lastnosti točk v omrežju, kar bo omogočalo določati pomembnosti in rangirati točke na razne načine. Ob tem smo opazovali sodelovanje med znanstveniki iz različnih področji znanosti agregirano po področjih (kot so podana v bazi SICRIS) in sodelovanje posameznikov neglede na področje. Ugotovili smo, da povprečen raziskovalec na projektih sodeluje s 8 drugimi raziskvoalci, in da so naravoslovne znanosti med najbolj povezanimi v smislu sodelovanja raziskovalcev iz različnih področij. Poleg tega smo opazovali lastnosti omrežja s stališča izmenjave informacij, po posameznikih in znanstvenih področjih.

2. Modelirali smo podatke o raziskovalcih in projektih z metodami za analizo besedil. Vsak projekt smo predstavili z naslovom in povzetkom, če je le ta bil na voljo. Vsakega raziskovalca pa z unijo opisov projektov na katerih je sodeloval. Ob tem smo analizirali besedilo samo in njegove lastnosti, zgradili vsebinsko ontologijo projektov in klasificiraliraziskovalce v znanstvena področja glede na njihove projekte.

3. Modelirali smo časovni razvoj omrežij raziskovalcev v smislu evolucije skupin točk. Ob tem smo opazovali spreminjanje velikosti omrežja, gostote omrežja, premera omrežja in povezane komponente. Ugotovili smo, da velikost omrežja z leti narašča približno linearno, da omrežje z leti postaja vse bolj gosto (raziskovalci medsebojno sodelujejo) in da se premer omrežja krči.

4. Modelirali smo časovni razvoj tematik. Pri tem smo celotne podatke razbili na časovne rezine in jih obdelali za potrebe analize razvoja vsebin (angl., topic evolution) v smislu novih projektov z določeno vsebino. Podatki zajemajo obdobje 1994 do 2010. Opazovali smo tudi kako se množica najbolj pogostih besed omenjenih v naslovih in povzetkih projektov spreminja skozi čas.

5. Kot rezultat preteklih nalog smo modelirali razvoj znanosti skozi tri vidike hkratno: (a) skozi društveni vidik (npr. sodelovanje med ljudmi), (b) skozi tematski vidik (npr. vsebina sodelovanja) in (c) skozi čas (npr. intenzivnost v danem trenutku). Ob tem smo predlagali novo mero za izračun centralnosti vozlišča, ki upošteva tako strukturo

omrežja (graf) kot tudi vsebino projektov (besedilo).

Opisano v R 3.1.

6. Ovrednotili smo modeliranje podatkov. Pri tem smo za ovrednotenje rezultatov iz preteklih nalog uporabili tradicionalne statistične mere. Opisano v R.3.2

DS4 Vizualizacija in interaktivna analiza
1. Naredili smo interaktivno vizualizacijo velikih omrežij. Izdelali smo modul za vizualizacijo in interakcijo z velikimi omrežji, ki temelji na znanih algorimih. Pri tem smo izdelali programski modul, ki je vključen v končni projektni sistem. Opisano v R.4.1

2. Pri interaktivni vizualizaciji korpusov besedil smo prilagodili rešitve izdelane na IJS. Sistemu smo dodali funkcionalnost za izris hierarhije tematik in pripadajočo uporabniško interakcijo. Opisano v R 4.1.

3. Izdelali smo interaktivno vizualizacijo časovnih podatkov. Pri tem smo nadgradili programske module razvite v preteklih dveh nalogah s časovno komponento, ki deluje nad omrežji in besedili. Opisano v R 4.1.

4. Izdelali smo kratko analizo interakcij z uporabniki. Pri tem smo ovrednostili uporabnost in učinkovitost orodij za interakcijo in vizualizacijo. Opisano v R 4.2.

DS5 Izdelava sistema
1. Izdelali smo strežniški sistem, ki povezuje analitske komponente izdelane vsklopih DS2, DS3 in DS4. Opisano v R 5.1.

2. Izdelali smo spletni odjemalec, ki komunicira s strežnikom. Opisano v R 5.1.

3. Opravili smo testiranje strežniškega in odjemalskega dela sistema. Opisano v R5.2.

4. Izdelali smo končno verzijo sistema (R5.3)

DS6 Diseminacija, promocija in dolgoročna vizija

1. Pripravili smo podroben diseminacijski in promocijski plan. Pri njegovi pripravi smo se usmerili na učinkovito promocijsko strategijo za promocijo portala v slovenskem prostoru, ki je usmerjen v specifične skupine uporabnikov, visokošolske organizacije, srednje in osnovne šole, raziskovalce, raziskovalne institucije, razvojnike v industrijskih okoljih ter javne ustanove. Pripravljeni plan je bil osnova za izvajanje diseminacijske in promocijske aktivnosti tekom trajanja celotnega projekta. Pri tem smo se v glavne posluževali obstoječih diseminacijskih kanalov. Opisano v R6.1.

Z izvedbo diseminacijskega in promocijskega načrta smo vzpostavili sistem obveščanja, širjenja, in povezovanja različnih ciljnih skupin preko enotnega sistema http://scienceatlas.ijs.si/ za enostaven in odprt dostop do razvojno-raziskovalnih podatkov s ciljem spodbujanja novih idej, sodelovanja med domačimi institucijami in industrijo, promocije znanstveno-raziskovalnih dosežkov doma in v tujini ter vzpostavitvijo okolja inovativne in kreativne kulture.

2. V okviru tega smo izvedli različna predavanja za kadrovnike, knjižničarje, študente in srednješolce, ki je potekalo v Univerzitetni knižnici Maribor. Promocijsko predavanje smo izvedli tudi v okviru gibanja InCo v Državnem svetu, posneta je bila oddaja na tretjem programu Radia Slovenija, povezali smo se z Združenjem manager ter vzpostavili kanal, kjer predstavljamo raziskovalne in podjetniške dosežke. Na konferenci "Drugačna pot do znanja" smo predstavili Videolectures portal in dosežke znanosti osnovno in srednješolskim učiteljem s področja tehnike in tehnologije. Promocijsko predavanje smo izvedli tudi na posvetovanju "Prosti dostop do dosežkov

slovenskih znanstvenikov", ki sta ga organizirala Sekcija za specialne knjižnice in Sekcija za visokošolske knjižnice pri Zvezi bibliotekarskih društev Slovenije. Opisano v R6.2.

3. Za potrebe motivacijske vertikale smo v sodelovanju z ARRS in razvojno-raziskovalnimi oddelki in inštituti pripravljali gradivo za promocijske videe, ki raziskovalcem, gospodarstvu in širši javnosti predstavljajo dobre raziskovalne skupine, posameznike ter raziskovalne dosežke. Rezultat naloge je 70 kratkih promocijskih videov v slovenskem in angleškem jeziku na portalu, na naslovu http://videolectures.net/promogram/ (R6.3) .

6.Ocena stopnje realizacije programa dela na raziskovalnem projektu in zastavljenih raziskovalnih ciljev[4]

V projektu smo zastavili naslednje cilje v smislu analize znanosti v slovenskem prostoru, ki smo jih tudi uresničili. Omogočiti analizo slovenskega znanstvenega prostora s treh vidikov:

- družbeni vidik, ki vključuje povezanost posameznih raziskovalcev, institucij in projektov;
- vsebinski vidik, ki vključuje tematike publikacij in projektov;
- časovni vidik, ki analizira znanost skozi čas, trende in predikcije razvoja.

Raziskovalni problem, ki smo ga reševali v projektu sestoji iz več komponent:

- Uvoz, priprava in konsolidacija (v smislu povezovanja podatkov) podatkov iz različnih virov. Ključna problema sta (a) identifikacija enakih objektov zapisanih na različne (vendar podobne) načine v različnih bazah, in (b) razločevanje različnih objektov zapisanih na enak način (razdvoumljanje oz. disambiguacija podatkov). Rezultat te faze so očiščeni in konsolidirani podatki - postopki za obdelavo so avtomatski.
- Za konsolidirane podatke o procesu razvoja znanosti lahko izpostavimo tri glavne lastnosti: (a) podatki oblikujejo omrežja (npr. sodelovanje med institucijami, soavtorstva ipd.), (b) podatki imajo vsebinsko komponento (npr. tekstovne opise, povzetke, video predstavitve, ipd), in (c) podatki imajo časovno komponento (kar bo omogočalo analizo evolucije znanstvenega procesa). V smislu metodologije obdelave takih podatkov, smo razvili vrsto prijemov, ki omogočajo večdimenzionalno analizo in agregiranje zbranih podatkov. Rezultat so implementirani algoritmi v okviru sistema, ki interpretir večdimenzionalne zahtevke in vrača rezultate v obliki klasičnih analiz in v obliki interaktivnih vizualizacij.
- Prikaz podatkov, ki opisujejo znanost je zaradi kompleksne narave opazovanih podatkov netrivialen. Vsaka od treh glavnih dimenzij (omrežja, vsebine, čas) vnaša dodatno kompleksnost v prikaz. Poleg prikazovanja posameznih dimenzij smo se ukvarjali tudi s prikazom kombiniranih dimenzij (npr. prikaz razvoja znanstvenih tematik skozi čas).

Izdelana spletna aplikacija http://scienceatlas.ijs.si/ omogoča naslednje glavne funkcionalnosti:

- uvoz/izvoz podatkov iz relevantnih baz preko vmesnikov;
- čiščenje in povezovanje baz podatkov;
- analiza in agregiranje podatkov z modernimi analitskimi metodami;
- interaktiven vmesnik za sprotno analizo podatkov;
- učinkovitost rešitve v smislu obdelave velike količine podatkov;
- omogočen podroben vpogled v različne vidike znanosti;
- instalacijo sistema oz. delov sistema na institucijah za kreiranje znanstvene politike v Sloveniji

7.Utemeljitev morebitnih sprememb programa raziskovalnega projekta oziroma sprememb, povečanja ali zmanjšanja sestave projektne skupine[5]

Ni bilo bisvetnih sprememb.

8.Najpomembnejši znanstveni rezultati projektne skupine[6]

	Znanstveni dosežek			
1.	COBISS ID		25283367	Vir: COBISS.SI
	Naslov	*SLO*	Vizualizacija znanstvenih sodelovanj v Sloveniji	
		ANG	Visualizations of Slovenian scientific community	
	Opis	*SLO*	Z uporabo naprednih tehnik analize podatkov lahko pridobimo nove vpoglede v podatke o znanstvenih sodelovanjih na nacionalnem nivoju. Ob tem podatke predstavimo kot graf raziskovalcev in razoskovalnih vsebin. V članku smo uproabili dve obstoječi tehnikii: diagram sodelovanj in zemljevid kompetenc. Diagramom sodelovanj prikaže sodelovanja med raziskovalci za izbranega raziskovalca ali raziskovalko, medtem ko zemljevid kompetenc pokaže raziskovalne vsebine na katerih je posameznik delal.	
		ANG	Using advanced analysis techniques new useful insight into data can be achieved. This paper addresses a problem of gaining insights in the data on scientific collaboration on a National level, where data can be seen as a graph with researchers and research content. Two existing visualization techniques were applied on data about scientific community in Slovenia: collaboration diagram and competence map. Collaboration diagram gives a clear overview of collaborations for a selected researcher, while competence map shows semantically grouped research content the researcher has worked on.	
	Objavljeno v		Institut Jožef Stefan; Zbornik 14. mednarodne multikonference Informacijska družba - IS 2011, 10.-14. oktober 2011; 2011; Str. 129-132; Avtorji / Authors: Karlovčec Mario, Mladenić Dunja, Grobelnik Marko, Jermol Mitja	
	Tipologija		1.08 Objavljeni znanstveni prispevek na konferenci	
2.	COBISS ID		25858087	Vir: COBISS.SI
	Naslov	*SLO*	Prekojezično identifikacija in disambiguacija imenskih entitet	
		ANG	Cross-lingual named entity extraction and disambiguation	
			Predlagali smo metodo za identifikacijo in disambiguacijo imenskih entitet za primer, ko se jezik v kateremu je napisano besedilo razlikuje od jezika v	

	Opis	*SLO*	kateremu imamo zapisano bazo znanja. Pokazali smo delovanje predlagane metode za disambiguacijo angleških in slovenskih imenskih entitet.
		ANG	We propose a method for the task of identifying and disambiguation of named entities in a scenario where the language of the source text differs from the language of the knowledge base. We demonstrate this functionality on English and Slovene named entity disambiguation.
	Objavljeno v		Mednarodna podiplomska šola Jožefa Stefana; Zbornik; 2012; Str. 176-181; Avtorji / Authors: Štajner Tadej, Mladenić Dunja
	Tipologija		1.08 Objavljeni znanstveni prispevek na konferenci

9.Najpomembnejši družbeno-ekonomski rezultati projektne skupine[7]

	Družbeno-ekonomski dosežek			
1.	COBISS ID			Vir: vpis v poročilo
	Naslov	*SLO*	Atlas slovenske znanosti	
		ANG	Atlas of Slovenian Science	
	Opis	*SLO*	Izdelana spletna aplikacija http://scienceatlas.ijs.si/ omogoča naslednje glavne funkcionalnosti: • uvoz/izvoz podatkov iz relevantnih baz preko vmesnikov; • čiščenje in povezovanje baz podatkov; • analiza in agregiranje podatkov z modernimi analitskimi metodami; • interaktiven vmesnik za sprotno analizo podatkov; • učinkovitost rešitve v smislu obdelave velike količine podatkov; • omogočen podroben vpogled v različne vidike znanosti; • instalacijo sistema oz. delov sistema na institucijah za kreiranje znanstvene politike v Sloveniji	
		ANG	The developed Web portal enables the following functionality: • data import and export from the relevant databases using interfaces • data cleaning and integration • data analysis and aggregation • interactive use interface for data analysis • scalable solution handling large amount of data • enabled detailed view in different aspects of Slovenian science • final version of the system to be installed and/or publicly available	
	Šifra		F.11 Razvoj nove storitve	
	Objavljeno v		http://scienceatlas.ijs.si/	
	Tipologija		2.21 Programska oprema	

10.Drugi pomembni rezultati projektne skupine[8]

Postavitev baze, ki združuje podatke o slovenskih raziskovalcih, njihovih projektih in objavah iz določenega obdobja v sodelvoanju z Inštitutom za matematiko, fiziko in mehaniko in Fakulteto za informacijske študije v Novem mestu na aplikativnem projektu »Omrežja soavtorstev slovenskih raziskovalcev: teoretična analiza in razvoj uporabniškega vmesnika za vizualizacijo«.

11.Pomen raziskovalnih rezultatov projektne skupine[9]

11.1.Pomen za razvoj znanosti[10]

SLO

Raziskovalna komponenta projekta naslovi vrsto problemov, ki so trenutno aktualni v raznih raziskovalnih področjih, ki se ukvarjajo z analizo in obdelavo podatkov. Glavni raziskovalni

izzivi, ki smo jih reševali v projektu: (1) povezovanje raznovrstnih baz podatkov (ugotavljanje identitete podatkovnih objektov), (2) prikaz (vizualizacija) in interaktivna analiza (angl. OLAP) kompleksnih podatkov kot so omrežja, besedila, časovni razvoj, (3) učinkovitost (skalabilnost) metod za analizo velikih količin podatkov.

ANG

The research component of the project addresses a number of problems from different research areas dealing with data analysis. The main research challenges that have been addressed in the project are the following: (1) connecting relevant databases, (2) data visualization and interactive data analysis of complex data including text, social networks, time tagged data, (3) scalability of methods for data analysis.

11.2.Pomen za razvoj Slovenije[11]

SLO

Vzpostavitev enotnega sistema za enostaven dostop do razvojno-raziskovalnih podatkov s ciljem spodbujanja novih idej, sodelovanja med domačimi institucijami in industrijo, promocije znanstveno-raziskovalnih dosežkov doma in v tujini ter vzpostavljanje okolja inovativne in kreativne kulture. Razvit sistem je dostopen kot spletna informacijska točka (spletni portal) na osnovi podatkov in baz, ki beležijo dosežke slovenskih raziskovalcev. S tem omogočimo odprt dostop do razvojno raziskovalnih javnih podatkov. Ob tem spletna informacijska točka ponuja vrsto inovativnih storitev za celovito empirično analizo slovenske znanosti skozi vrsto analitskih orodij za analizo dogajanja v slovenskem razvojno raziskovalnem okolju, vključno s kompetenčnimi grafi in sodelovalnimi grafi. S tem ponudimo orodja za pomoč industriji in raziskovalnim organizacijam pri iskanju potencialnih partnerjev, potrebnih kompetenc in rezultatov. http://scienceatlas.ijs.si/

Za potrebe diseminacije in motivacije smo pripravili krajše promocijske filme, ki promovirajo znanstveno-raziskovalne dosežke. http://videolectures.net/promogram/

ANG

Establishment of a system for easy access to data on Slovenian researchers to stimulate new ideas, collaboration between research and industry, promotion of scientific results. The developed system is accessible as Web portal integrating data from several databases on Slovenian researchers. This enables easy access to publicly available data on activity of Slovenian researchers. At the same time the Web portal offers a number of innovative services for empirical analysis of Slovenian research space via analytics tools including competence graphs and collaboration graphs. The project offers tools to support industry and research organizations in searching for partners, the needed competences and results. http://scienceatlas.ijs.si/

Short promotional videos of research and development results will serve dissemination purpose of the project and its results. http://videolectures.net/promogram/

12.Vpetost raziskovalnih rezultatov projektne skupine.

12.1.Vpetost raziskave v domače okolje

Kje obstaja verjetnost, da bodo vaša znanstvena spoznanja deležna zaznavnega odziva?

- [x] v domačih znanstvenih krogih
- [ ] pri domačih uporabnikih

Kdo (poleg sofinancerjev) že izraža interes po vaših spoznanjih oziroma rezultatih?[12]

Raziskovalci in študenti z Univerze v Ljubljani, še posebej z Fakulteta za računalništvo ter Fakultete za matematiko in fiziko in Fakultete za informacijske študije v Novem mestu.

12.2.Vpetost raziskave v tuje okolje

Kje obstaja verjetnost, da bodo vaša znanstvena spoznanja deležna zaznavnega odziva?

- [x] v mednarodnih znanstvenih krogih
- [ ] pri mednarodnih uporabnikih

Navedite število in obliko formalnega raziskovalnega sodelovanja s tujini raziskovalnimi inštitucijami:[13]

Na odseku imamo letno kativnih okoli 15 evropskih projektov. Poleg tega imamo tudi domače projekte, sodelujemo v programski skupini in dveh centrih odličnosti. Raziskovalni rezultati projekta so bili predstavljeni na več delovnih srečanj, kjer obstaja potreba za analizo podatkov podanih v tekstovni obliki ali kot socialno omrežje (kot so evropski projekti RENDER, XLike, PlanetData) .

Kateri so rezultati tovrstnega sodelovanja:[14]

Rezultati sodelovanja vključujejo razvoj posplošene mere za oceno centralnosti posameznega vozlišča v omrežju, ki uporablja tako strukturo omrežja (n.pr., sodelujoče organizacije v raziskovalnem projektu) kot tudi besedilo podano za posamezna vozlišča (n.pr., naslov raziskovalnega projekta). Objava je v pripravi. Poleg tega si obetamo skupno prijavo novih mednarodnih projektov, ki bodo uporabljali in nadgradili rezultate projekta, tako v smeri vključitve podatkov o mednarodnih sodelovanjih raziskovalcev kot tudi v smeri evalvacije in nadgradnje raziskovalnih metod za analizo besedil in omrežji.

13.Izjemni dosežek v letu 2012[15]

13.1. Izjemni znanstveni dosežek

Nova metoda za ocenjevanje centralnosti točke v omrežju, ki upošteva besedilo podano kot opis vozlišča v omrežju. Objava v pripravi: Centrality measure incorporating textual context.

13.2. Izjemni družbeno-ekonomski dosežek

Vzpostavitev enotnega sistema za enostaven dostop do razvojno-raziskovalnih podatkov s ciljem spodbujanja novih idej, sodelovanja med domačimi institucijami in industrijo, promocije znanstveno-raziskovalnih dosežkov doma in v tujini ter vzpostavljanje okolja inovativne in kreativne kulture. Razvit sistem je dostopen kot spletna informacijska točka (spletni portal) na osnovi podatkov in baz, ki beležijo dosežke slovenskih raziskovalcev: http://scienceatlas.ijs.si/

C. IZJAVE

Podpisani izjavljam/o, da:

- so vsi podatki, ki jih navajamo v poročilu, resnični in točni
- se strinjamo z obdelavo podatkov v skladu z zakonodajo o varstvu osebnih podatkov za potrebe ocenjevanja in obdelavo teh podatkov za evidence ARRS
- so vsi podatki v obrazcu v elektronski obliki identični podatkom v obrazcu v pisni obliki
- so z vsebino zaključnega poročila seznanjeni in se strinjajo vsi soizvajalci projekta
- bomo sofinancerjem istočasno z zaključnim poročilom predložili tudi elaborat na zgoščenki (CD), ki ga bomo posredovali po pošti, skladno z zahtevami sofinancerjev.

Podpisi:

zastopnik oz. pooblaščena oseba raziskovalne organizacije:	in	*vodja raziskovalnega projekta:*
Institut "Jožef Stefan"		Dunja Mladenić
_____		_____

ŽIG

Kraj in datum:	Ljubljana	2.4.2013

Oznaka prijave: ARRS-CRP-ZP-2013-02/7

---

[1] Opredelite raziskovalno področje po klasifikaciji FOS 2007 (Fields of Science). Prevajalna tabela med raziskovalnimi področji po klasifikaciji ARRS ter po klasifikaciji FOS 2007 (Fields of Science) s kategorijami WOS (Web of Science) kot podpodročji je dostopna na spletni strani agencije (http://www.arrs.gov.si/sl/gradivo/sifranti/preslik-vpp-fos-wos.asp). Nazaj

[2] Napišite povzetek raziskovalnega projekta (največ 3.000 znakov v slovenskem in angleškem jeziku). Nazaj

[3] Napišite kratko vsebinsko poročilo, kjer boste predstavili raziskovalno hipotezo in opis raziskovanja. Navedite ključne ugotovitve, znanstvena spoznanja, rezultate in učinke raziskovalnega projekta in njihovo uporabo ter sodelovanje s tujimi partnerji. Največ 12.000 znakov vključno s presledki (približno dve strani, velikost pisave 11). Nazaj

[4] Realizacija raziskovalne hipoteze. Največ 3.000 znakov vključno s presledki (približno pol strani, velikost pisave 11). Nazaj

[5] V primeru bistvenih odstopanj in sprememb od predvidenega programa raziskovalnega projekta, kot je bil zapisan v predlogu raziskovalnega projekta oziroma v primeru sprememb, povečanja ali zmanjšanja sestave projektne skupine v zadnjem letu izvajanja projekta, napišite obrazložitev. V primeru, da sprememb ni bilo, to navedite. Največ 6.000 znakov vključno s presledki (približno ena stran, velikosti pisave 11). Nazaj

[6] Navedite znanstvene dosežke, ki so nastali v okviru tega projekta.
Raziskovalni dosežek iz obdobja izvajanja projekta (do oddaje zaključnega poročila) vpišete tako, da izpolnite COBISS kodo dosežka – sistem nato sam izpolni naslov objave, naziv, IF in srednjo vrednost revije, naziv FOS področja ter podatek, ali je dosežek uvrščen v A'' ali A'. Nazaj

[7] Navedite družbeno-ekonomske dosežke, ki so nastali v okviru tega projekta.
Družbeno-ekonomski rezultat iz obdobja izvajanja projekta (do oddaje zaključnega poročila) vpišete tako, da izpolnite COBISS kodo dosežka – sistem nato sam izpolni naslov objave, naziv, IF in srednjo vrednost revije, naziv FOS področja ter podatek, ali je dosežek uvrščen v A'' ali A'.

Družbeno-ekonomski dosežek je po svoji strukturi drugačen kot znanstveni dosežek. Povzetek znanstvenega dosežka je praviloma povzetek bibliografske enote (članka, knjige), v kateri je dosežek objavljen.

Povzetek družbeno-ekonomskega dosežka praviloma ni povzetek bibliografske enote, ki ta dosežek dokumentira, ker je dosežek sklop več rezultatov raziskovanja, ki je lahko dokumentiran v različnih bibliografskih enotah. COBISS ID zato ni enoznačen izjemoma pa ga lahko tudi ni (npr. prehod mlajših sodelavcev v gospodarstvo na pomembnih raziskovalnih nalogah, ali ustanovitev podjetja kot rezultat projekta … - v obeh primerih ni COBISS ID). Nazaj

[8] Navedite rezultate raziskovalnega projekta iz obdobja izvajanja projekta (do oddaje zaključnega poročila) v primeru, da katerega od rezultatov ni mogoče navesti v točkah 8 in 9 (npr. ker se ga v sistemu COBISS ne vodi). Največ 2.000 znakov, vključno s presledki. Nazaj

[9] Pomen raziskovalnih rezultatov za razvoj znanosti in za razvoj Slovenije bo objavljen na spletni strani: http://sicris.izum.si/ za posamezen projekt, ki je predmet poročanja. Nazaj

[10] Največ 4.000 znakov, vključno s presledki. Nazaj

[11] Največ 4.000 znakov, vključno s presledki. Nazaj

[12] Največ 500 znakov, vključno s presledki. Nazaj

[13] Največ 500 znakov, vključno s presledki. Nazaj

[14] Največ 1.000 znakov, vključno s presledki. Nazaj

[15] Navedite en izjemni znanstveni dosežek in/ali en izjemni družbeno-ekonomski dosežek raziskovalnega projekta v letu 2012 (največ 1000 znakov, vključno s presledki). Za dosežek pripravite diapozitiv, ki vsebuje sliko ali drugo slikovno gradivo v zvezi z izjemnim dosežkom (velikost pisave najmanj 16, približno pol strani) in opis izjemnega dosežka (velikost pisave 12, približno pol strani). Diapozitiv/-a priložite kot priponko/-i k temu poročilu.
Vzorec diapozitiva je objavljen na spletni strani ARRS http://www.arrs.gov.si/sl/gradivo/, predstavitve dosežkov za pretekla leta pa so objavljena na spletni strani http://www.arrs.gov.si/sl/analize/dosez/ Nazaj

Obrazec: ARRS-CRP-ZP/2013-02 v1.00
BB-B0-49-25-B5-B0-CC-41-10-29-F1-96-44-05-B5-45-E0-86-58-89

Atlas slovenske znanosti

RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

R1 - Analiza zahtev in podrobna specifikacija sistema

Avtorji: Mitja Jermol (IJS)

Ljubljana, 13.3.2011

Vsebina

1 Uvod

Ta dokument zajema celovito specifikacijo sistema Atlas Slovenske Znanosti, ki je nastala na podlagi podrobne analize zahtev naročnika ter operaterjev treh ključnih portalov: SICRIS, VideoLectures.NET ter IST-WORLD. Dokument je namenjen razvijalcem sistema in služi kot podlaga za vse nadaljnje razvojne aktivnosti. Dokument je sestavljen iz uvoda, drugega poglavja, ki definira cilje projekta ASZ ter opisom podrobnih strateških, vsebinskih in tehničnih zahtev ter omejitev. Tretje poglavje podrobno analizira osnovne gradnike sistema. Četrto poglavje opisuje strukturo aplikacije in modulov ASZ. Peto poglavje je podrobna specifikacija sistema z opisom logične strukture, komponent sistema ter fizične strukture. V zaključku dokumenta so podana kratka razmišljanja ter reference.

2 Cilji projekta in aplikacije Atlas Slovenske Znanosti

Vizija projekta Atlas Slovenske Znanosti (ASZ) je vzpostavitev enotnega sistema za enostaven in odprt dostop do razvojno-raziskovalnih podatkov s ciljem spodbujanja novih idej, sodelovanja med domačimi inštitucijami in industrijo, promocije znanstveno-raziskovalnih dosežkov doma in v tujini ter vzpostavljanje okolja inovativne in kreativne kulture. V ta namen bomo uporabljali obstoječe podatkovne zbirke:

- Osnovne podatkovne zbirke: SICRIS podatkovna zbirka slovenskih raziskav, IST-WORLD podatkovna zbirka mednarodnih raziskav v katerem so vključeni tudi podatki CORDIS podatkovne zbirke Evropske Unije ter VideoLectures.NET podatkovna zbirka.
- Pomožne podatkovne zbirke: COBISS podatkovna zbirka objav in publikacij slovenskih raziskovalcev, Google Schoolar podatkovna zbirka objav in publikacij svetovnih raziskovalcev, IST-WORLD/publications podatkovna zbirka objav in publikacij Evropskih raziskovalcev.
- Dodatne podatkovne zbirke, ki v sklopu projekta še ne bodo vključene, bo pa infrastruktura rešitve upoštevala specifičnosti teh podatkovnih zbirk za kasnejše vključevanje: spletne predstavitve raziskav in dosežkov posamezne organizacije, podatkovna zbirka Evropskih patentov, potencialne nove podatkovne zbirke, ki se bodo pojavile v toku projekta.

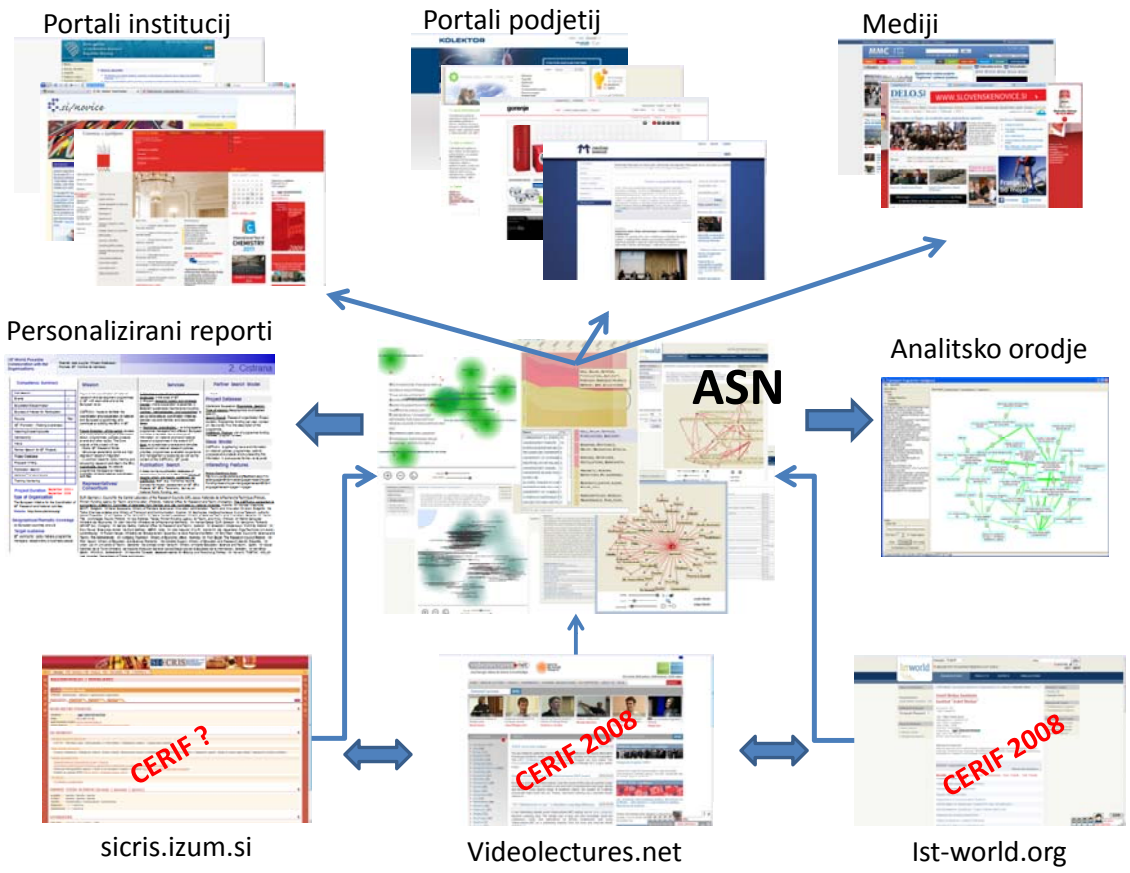
V ta namen bo projekt AZS vzpostavil enotno spletno informacijsko točko na osnovi obstoječih podatkov in baz, ki beležijo dosežke slovenskih raziskovalcev, ki:

- bo omogočala odprt dostop do vseh razvojno raziskovalnih javnih podatkov različnih modalnosti (tekstovni, numerični, video prezentacije, grafi, strukturirano znanje v obliki taksonomij),
- bo ponujala vrsto inovativnih storitev na bazi semantičnih tehnologij za celovito empirično analizo slovenske znanosti:
    - poglobljeno iskanje po tekstu in videih,
    - vrsto analitskih orodij za analizo dogajanja v slovenskem razvojno raziskovalnem okolju, vključno s kompetenčnimi grafi, sodelovalnimi grafi, predikcijami, trendi in simulacijami,
    - orodja za pomoč industriji in raziskovalnim organizacijam pri iskanju potencialnih partnerjev, potrebnih kompetenc in rezultatov, orodje za sestavljanje optimalnega konzorcija bodočih raziskovalnih projektov,

- bo omogočala dvosmerne prenose podatkov med tremi osnovnimi bazami podatkov:
  - SICRIS baza podatkov (z domačimi raziskovalnimi podatki)
  - IST-World baza podatkov (z mednarodnimi raziskovalnimi podatki) za potrebe primerjav, iskanja zunanjih kompetenc, mednarodnih analiz ipd.
  - Videolectures.net baza video predavanj in dosežkov domačih in tujih znanstvenikov in vzpostavitev znanstvene videografije v obliki videoCRIS repositorija.
- bo omogočala enostavno vključitev posameznih storitev v druge spletne portale kot npr. portal ARRS, portali raziskovalnih institucij, portali razvojnih oddelkov ipd. s ciljem širše promocije slovenske znanosti.

2.1 Definicija osnovnih zahtev

Osnovne zahteve AZS so bile definirane na osnovi pogovorov in specifikacij z naročnikom ter lastniki in upravljalci posameznih podatkovnih zbirk. Na osnovi dobljenih rezultatov smo zgradili shematsko strukturo rešitve na sliki 1, ki prikazuje glavne skupine zahtev:



Slika 1: Shematska struktura rešitve AZS

Osnova rešitve je enoten spletni portal za pregled, preiskovanje, analize ter simulacije na osnovi sledečih podatkov (na sliki 1 označen kot ASN):

- raziskovalcev in raziskovalnih skupin,

- raziskovalnih organizacij kamor štejemo podatke o raziskovalnih skupinah, ki so formalno prijavljene na Ministrstvu za visoko šolstvo, znanost in tehnologijo ter raziskovalnih oddelkov ostalih organizacij, ki so dosegljivi preko spleta,
- raziskovalnih projektih,
- raziskovalnih dosežkov, kamor štejemo projektne rezultate in znanstvene publikacije v tradicionalni in video obliki.

Funkcije portala bodo morale omogočati celovit pregled, enostavne in kompleksne analize ter simulacije. V ta namen bo potrebno izdelati uporabniške scenarije in aplikacije, ki bodo omogočale prereze preko vsebin, omrežij in dimenzije časa. Zato bo npr. potrebno poleg prikazovanja posameznih dimenzij, omogočati tudi prikaz kombiniranih dimenzij (npr. prikaz evolucije društvenega omrežja, prikaz razvoja znanstvenih tematik skozi čas, prikaz znanstvenih tematik na društvenih omrežjih).

Naslednja osnovna zahteva je dvosmeren prenos podatkov med tremi osnovnimi podatkovnimi zbirkami. Zaradi doseganja sinergičnih učinkov je interes naročnika in upravljalca navedenih spletnih portalov avtomatsko in dvosmerno prenašanje vsebin oz. povezovanje na nivoju vsebin. Zato bo potrebno na nivoju podatkovnih zbirk izdelati ustrezne vmesnike (spodaj na sliki 1):

- vmesnik med SICRIS in VideoLectures.NET,
- vmesnik med SICRIS in IST-WORLD,
- vmesnik med videolectures.NET in IST-WORLD, ki že obstaja.

Tretja pomembna zahteva je povezana s ciljem čimširše diseminacije rezultatov v slovensko poslovno okolje. Poleg diseminacijskih in promocijskih aktivnosti, ki so opisane v R61 Promocijski in diseminacijski plan je potreba po izdelavi ustreznih rešitev, ki bi tesneje povezovale informacijske kanale in posamezne aktivnosti v širši družbi. Ker bo celotna rešitev temeljila na spletnih servisih, bomo za vsak analitski modul izdelali takoimenovane »snipets« oz aplikacije, ki bodo enostavno integrabilne v druge spletne portale. V ta namen smo identificirali sledeče skupine spletnih portalov (gornji del slike 1):

- portali institucij kot so portal MVZT, portal ARRS, portal GZS, portal slovenia.info, portal IJS, portali univerz, portali fakultet, in drugi možni portali,
- portali podjetij do katerih bomo dostopali preko direktnih povezav IJS, ARRS ter povezav GZS,
- Mediji kot so finance-on.net, 24ur.com, dnevnik.si, delo.si, rtvslo.si in podobno.
- Socialna omrežja, kot so Facebook, Linkedin, Twitter, Orkut in podobno.

Posebej, z namenom dostopanja do podjetij in organizacij, predvidevamo tudi možnost izdelave in pošiljanja personaliziranih analiz in poročil posameznim odjemalcem (na sliki označeno kot personalizirani reporti).

Četrta pomembna zahteva je vezana na interes ARRS in MVZT po podrobnem vpogledu v raziskovalne podatke. V ta namen bomo predlagali tudi izdelavo ločenega analitskega okolja, kjer bo ARRS lahko vključeval podatke, ki niso javnega značaja kot npr finančne podatke, zaupna poročila ipd. Tukaj predvidevamo izdelavo tudi dodatnih analitskih metod in predvsem metod za planiranje in simulacije vplivov raziskovalnih aktivnosti ARRS in MVZT s ciljem načrtovanja raziskovalnih politik.

Naslednji nivo zahtev so tehnične narave in so bile podane iz treh osnovnih podatkovnih zbirk.

IST-WORLD: portal za analitiko evropske znanosti je dosegljiv na http://www.ist-world.org in je rezultat 6OP projekta »IST World: Knowledge Base for RTD Competencies in IST« Project No: FP6-2004-IST-3 – 015823. Portal upravljamo na Institutu »Jožef Stefan«. Osnovna zahteva portala je, da so podatki v drugih podatkovnih bazah zapisani v katerikoli verziji CERIF formata.

VideoLectures.NET: je portal, ki je ravno tako v upravljanju na Institutu »Jožef Stefan«. Osnovne zahteve portala za dvosmeren prenos podatkov so:

- File level: FLV, MPEG-4, WMF, MP3, AAC
- Content level metadata: Extended DC

Poleg tega VideoLectures.NET postavlja posebne kvalitativne zahteve za primer prenosa posnetega videa na VideoLectures.NET. Prvi pogoj je tehnična ustreznost formata ter tehnične zahteve na nivoju datoteke. Drugi pogoj je ustreznost posnetka, ki mora zadostovati kvaliteti zvoka ter ustrezni kvaliteti slike. Tretji pogoj je kvaliteta in primernost vsebine predavanja. To mora biti znanstveno predavanje oz. predstavitev z znanstveno vsebino.

SICRIS: trenutno od upravljalca SICRIS portala še nismo pridobili zahtev.

2.1.1 Definicija funkcionalnih zahtev analitskega modula

Funkcionalne zahteve za portal AZS so:

- Preiskovanje baze podatkov mora biti omogočeno preko enostavne in uporabniku razumljive aplikacije.
- Pregledovanje vsebine mora omogočati enostavno preklapljanje od entitete do entitete (organizacija, projekt, raziskovalec, rezultat) preko GUI-a, ki bo prikazal celovit pregled in specifične podrobne podatke v osnovni in agregirani obliki.
- Pregledovanje podatkov preko zgrajene taksonomije. Taksonomija (Ontologija) bo del podatkovne baze. GUI bo omogočal pregledovanje po taksonomiji. Za osnovo bomo prevzeli DMOZ (http://dmoz.org) taksonomijo, ki jo bomo nadgrajevali z novimi kategorijami s pomočjo servisov ontogen-a.
- Orodje za analizo raziskovalnih združb (posameznikov in organizacij) mora prikazovati na grafični način socialne mreže, povezave, uteži povezav ter tudi lokalne socialne mreže.
- Orodje za analize ekspertiz in kompetenc mora prikazovati naračunane kompetence in ekspertize posamezne organizacije, posameznike, skupine ter tudi geografskega področja. Uporabnik bo lahko spremljal kompetence izbrane entitete, trenutno in preteklo delo ter tudi izračunal predikcije kompetenc na osnovi raziskovalnih smernic.
- Analiza potencialnega konzorcija oz partnerja na raziskovalnem/razvojnem projektu mora prikazati rangirano listo potencialnih partnerjev. Rangiranje bo temeljilo na kombinaciji cenilk kompetenc, kvaliteti preteklega sodelovanja, kvaliteti rezultatov ter zaupanju med partnerji.
- Moduli za identifikacijo trendov in izdelavo predikcij morajo izračunati na osnovi preteklih podatkov, detektiranih trendov in naučenega trenutnega modela, možne bodoče smeri raziskav, področij, razvoj posamezne entitete, razvoj socialnih in sodelovalnih mrež. Smeri morajo biti prikazane grafično in zajemati faktorje zaupanja v napovedi.

- Rešitev mora omogočati največjo možno stopnjo integracije z izbranimi socialnimi omrežji kot so Facebook, Linkedin, Twitter, OpenBC ter socialnim omrežjem VideoLectures.NET in IST-WORLD.

2.1.2 Definicija drugih, nefunkcionalnih zahtev analitskega modula

Druge nefunkcionalne zahteve so:

- Učinkovita analiza podatkov: glede na veliko količino podatkov ter pričakovan obisk uporabnikov morata biti tako offline kot realtime analitika ustrezno učinkovita, odzivna in prikazovati rezultate kompleksnih matematičnih analiz v čimkrajšem času. Ključno pri tem bo integracija analitskih metod za analitiko v realnem času ter ustrezne vizualizacijske metode.
- Uporabniku prijazen, naraven in grafično bogat GUI.
- Učinkovit in neškodljiv prenos podatkov med posameznimi podatkovnimi zbirkami. Potrebno je paziti na integriteto in funkcionalnost vseh zunanjih portalov. Podatki, na katerih bomo izvajali analize morajo biti up-to-date.
- Skalabilnost. Storitvi bomo sčasoma dodajali nove podatke, nove podatkovne zbirke ter nove analitske funkcionalnosti. Zato mora celovita vertikala rešitve omogočati skalabilnost.

2.2 Opis osnovnih omejitev

Prva in najpomembnejša omejitev je ustreznost podatkov. Le ti morajo vsebovati informacije o vsebini (opis organizacije, projekta, ljudi, opis rezultatov ipd.), povezavah (povezava med organizacijami, med ljudmi, med ljudmi in organizacijami, ipd.) ter časovni dimenziji (kdaj se je projekt začel, kdaj končal, kdaj se je raziskovalec zaposlil pri določeni organizaciji, kdaj je bil članek objavljen ipd.).

Pomembna tehnična omejitev je direktno dostopanje do posameznih podatkovnih zbirk. V ta namen bomo v okviru projekta razvili ustrezne adapterje. Za dostop do SICRIS baze podatkov trenutno testno uporabljamo API, ki pa omogoča samo enosmerni prenos podatkov.

Omejitve dostopanja do drugih portalov in informacijskih kanalov bodo znane, ko bomo z njimi dosegali dogovore o sodelovanju.

2.3 Opis osnovnih funkcij uporabe

Rezultati projekta bodo:

- spletni portal s funkcionalnostmi:
  - pregleda vsebin na enem mestu,
  - semantično podprtega preiskovanja vsebin,
  - analitska orodja za:
    - analizo družtvenih omrežij in besedil skozi čas,
    - računanje in analizo kompetenc posameznikov, organizacij,

    - sprotno analizo znanosti in sorodnih procesov,
    - vpogled v in vzpostavljanje preglednosti delovanja in razvoja slovenske znanosti,
    - pripravo publikacij na temo empirične analize slovenske znanosti,
  - orodja za pomoč pri iskanju ustreznih partnerjev in izračun najbolj optimalnega konzorcija,
  - spletnih strani osebne in organizacijske biblio- in videografije,
- vmesniki za izmenjavo podatkov med tremi bazami,
- prenosljivost in uporabnost sistema oz. delov sistema (npr. za potrebe ARRS) za analizo drugih (ne nujno slovenskih) podatkov o znanstvenih področjih, bibliografskih baz, baz projektov itd.
- Z namenom online povezovanja strokovnjakov iz industrije in raziskovalnih organizacij bomo kot neodvisni servis, ki bo instaliran na različnih portalih postavili tudi LifeNetLive aplikacijo, ki v realnem času povezuje obiskovalce več spletnih strani glede na semantično podobnost tematike.

Zato bo morala aplikacija omogočati naslednje funkcionalnosti:

- Stalen uvoz/izvoz podatkov iz relevantnih baz (podrobneje naštetih v naslednjih sekcijah) preko vmesnikov.
- Čiščenje in povezovanje baz podatkov.
- Analizo in agregiranje podatkov z modernimi analitskimi metodami in dodatnimi funkcionalnostmi kot so:
  - iskanje znanj, kompetenc in ustreznih partnerjev,
  - iskanje ekspertov,
  - iskanje in analiza RTD rezultatov,
  - iskanje in identifikacija potencialnih raziskovalnih tematik,
  - iskanje komplementarnih in kompetitivnih projektov,
  - gradnja in vzdrževanje osebnih profesionalnih mrež preko obstoječih socialnih mrež,
  - poglobljene analize raziskav in gibanj v raziskavah,
  - monitoring in opozarjanje na detektirane pomembne dogodke, aktivnosti in rezultate,
  - napovedovanje razvoja znanosti in raziskav glede na različne kriterije,
  - detekcija trendov v znanosti, raziskavah in raziskovalni socialni skupnosti,
  - detekcija neformalnih raziskovalnih skupnosti na določeno tematiko, uspešnih raziskovalcev,
  - ocenjevanje uspešnosti posameznikov, skupin in organizacij,
  - ...
- Aplikacijo za online združevanje raziskovalcev s podobnimi interesi.
- Interaktiven vmesnik za sprotno analizo podatkov.
- Učinkovitost (skalabilnost) rešitve v smislu obdelave velike količine podatkov.
- Omogočen podroben vpogled v različne vidike znanosti.
- Instalacijo sistema oz. delov sistema na institucijah za kreiranje znanstvene politike v Sloveniji in v perspektivi tudi v svetu.

3 Analiza osnovnih gradnikov ASZ

V tem poglavju so opisani gradniki, ki predstavljajo osnovo ASZ. Prvi trije so neodvisne in samostojne spletne storitve VideoLectures.NET, SICRIS in IST-WORLD. Druga dva sta knjižnici metod, algoritmov in aplikacij, ki bodo sestavljala analitski del portala ASZ.

3.1 VideoLectures.NET

Portal VideoLectures.NET je trenutno največji referenčni portal z izobraževalnimi video vsebinami na svetu. Uporabnikom s področja znanosti in raziskav, izobraževanja ter gospodarstva, ponuja največjo zbirko visokokakovostnih, recenziranih, videov obogatenih s prezentacijami in dodatnimi gradivi. Z uporabo naprednih semantičnih tehnologij in različnih tehnik vizualiziranja podatkov, omogoča edinstveno izobraževalno izkušnjo. VideoLectures.NET ne ponuja samo zbirke video posnetkov, temveč tudi strukturirane izobraževalne tečaje in kurikule, ki jih pripravljajo priznani profesorji iz različnih področij znanosti. Portal tako služi kot odprta video izobraževalna platforma za študente, profesorje, raziskovalce, akademike in gospodarstvenike ter vso javnost. Z omogočanjem prostega dostopa do vsebin priznanih univerz in strokovnih konferenc, skuša premostiti rastoči ekonomski in izobraževalni prepad med razvitimi in nerazvitimi deli sveta. Izobraževalni videi na portalu pokrivajo različna področja od računalniških ved, naravoslovja, družboslovja in humanistike, do novih in nastajajočih znanstvenih disciplin kot so kompleksna znanost, interoperabilnost, ipd. VideoLectures.NET poskuša z arhiviranjem posnetih predavanji iz univerz, konferenc, delavnic, in drugih znanstvenih dogodkov, ohranjati bogato in dinamično raziskovalno in izobraževalno dogajanje ter na ta način nadgrajevati tradicionalne učne zbirke in repozitorije znanja. VideoLectures.NET predstavlja prosto dostopno izobraževanje vsakomur, ne glede na njegovo socialno-ekonomsko poreklo. S svojo odprtostjo in prosto dostopnostjo nudi možnost izmenjave znanja na vseh stopnjah, in na ta način dolgoročno koristi družbi ter spodbuja razvoj gospodarstva.

Na naslovu http://videolectures.net je bilo januarja 2011 arhiviranih 543 dogodkov, 10261 predavanj in 12657 video posnetkov, ki jih je prispevalo 7946 avtorjev. Na portal dodamo, tedensko, povprečno 58 novih video vsebin, dnevno ga obišče do 8000 obiskovalcev.

Vsebina portala je zasnovana tako, da s svojimi inovativnimi pristopi ponuja celovito podporo izobraževalnemu procesu, in sicer: istočasno prikazovanje videa, prosojnic in komentarjev, možnost gradnje in upravljanje individualnih kurikulov ter sočasno distribucijo videa preko svetovnega spleta.

V letu 2010 je bil portal nadgrajen z **blogom**, uporabnikom pa so bile poslane tudi prve e-novice. Z uporabo e-novic, prejemajo registrirani uporabniki na svoj elektronski naslov redna obvestila o najnovejših dogodkih na portalu. Z e-novicami želimo povečati oglede predavanj in okrepiti medsebojne odnose z uporabniki. VideoLectures.NET ima v svoji spletni bazi podatkov vpisanih 10229 registriranih uporabnikov. Prve e-novice so bile poslane na 10321 spletnih naslovov, od tega je novice odprlo in sprejelo 2622 uporabnikov. V letu 2011 bomo e-novice pošiljati 4 krat mesečno.

Z željo po gradnji socialnega omrežja z uporabniki smo na Videolectures.NET začeli s pisanjem in urejanjem blog prispevkov v angleškem in slovenskem jeziku. Prispevki so večinoma neformalna poročila uredništva portala, ki zajemajo dogajanja na portalu, med drugim predstavljajo kritike

predavanj ter ostala vsebinska sporočila s področja izobraževanja, zanimiva za gledalce. Način objavljanja prispevkov je takšen, da je večina besedila prispevka vidnega na domači strani, klik na naslov prispevka bralca odpelje na podstran, kjer se nahaja prispevek v celoti, zato imata največ ogledov ravno domači strani. Na slovenskem blogu (http://blogslo.videolectures.net/) je trenutno objavljenih 181 prispevkov, kjer se nahaja 83 komentarjev, prispevki so kategorizirani v 22 kategorij in 751 iskalnih ključnih besed. Najbolj prometni dan je bil 8. december 2010, s 70. ogledi, skupnih ogledov pa je, od postavitve bloga julija 2010 do januarja 2010, 4.003.

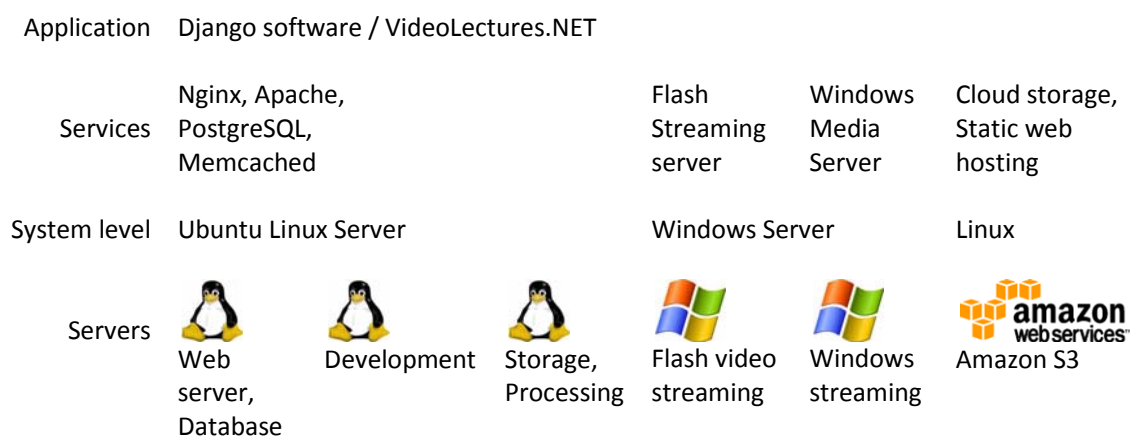
Angleški blog ima 4 podstrani, 305 objav, 84 komentarjev, 1336 označenih iskalnih ključnih besed in 29 kategorij. Najbolj prometni dan je bil 14. september 2010, z 236. ogledi, skupno je le teh 21.740. Na blogu je v vsakem trenutku približno 9 bralcev. Uporabljena platforma je WordPress, verzija 2.9.2.

Prav tako smo začeli graditi socialno mrežo uporabnikov portala preko facebooka http://www.facebook.com/videolectures in twitterja http://twitter.com/videolectures/ .

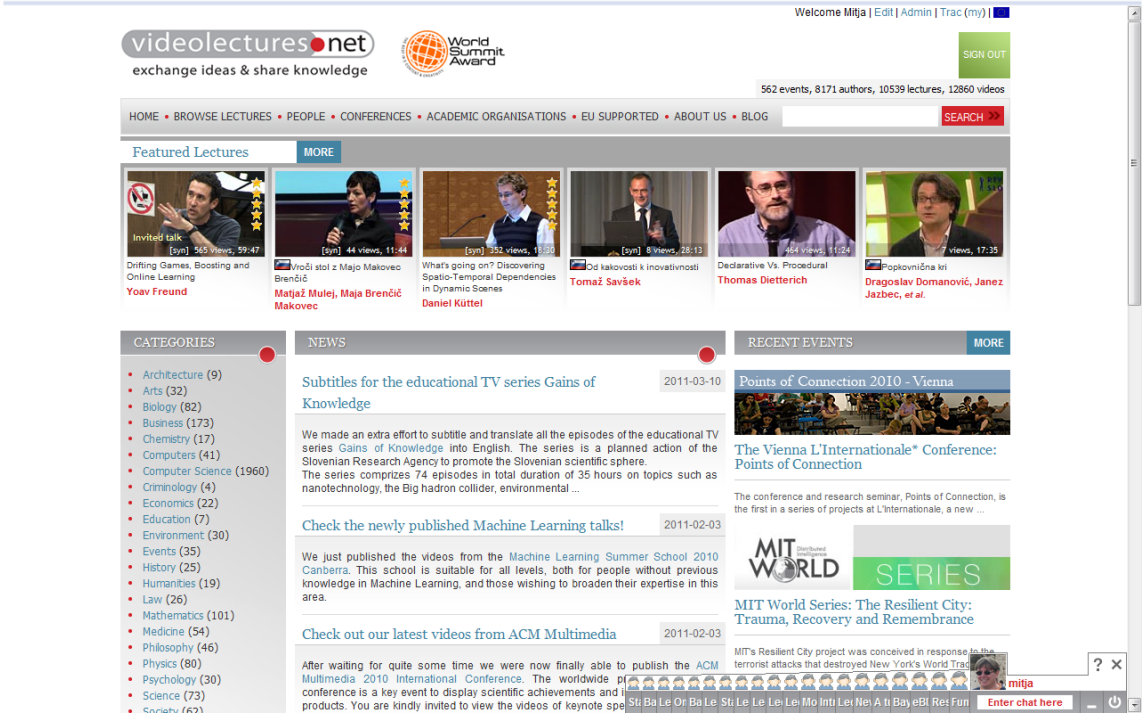
Zaradi kakovostnih vsebin in prepoznavnosti portala, ter ponujanja neomejenega brezplačnega dostopa do izobraževalnih vsebin smo že sklenili dogovore o sodelovanju in gostovanju predavanj s sledečimi priznanimi univerzami in instituti: MIT – Massachusetts Institute of Technology (ZDA), University of Cambridge (VB), Carnegie Mellon University (ZDA), Yale (ZDA), OpenCourseWare Consortium, CERN (CH) in ETH Zürich (The Swiss Federal Institute of Technology Zurich).

V Sloveniji sodelujemo z nekaterimi članicami Univerze v Ljubljani ter Javno agencijo za raziskovalno dejavnost republike Slovenije (ARRS).

Spodnja slika prikazuje strukturo tehnične rešitve portala.



Slika 2: Tehnična struktura portala VideoLectures.NET



Slika 3: VideoLectures.NET

3.2 SICRIS

Opis SICRIS-a smo povzeli po opisu, ki je dostopen na http://sicris.izum.si/about/cris.aspx?lang=slv.

Informacijski sistem SICRIS razvijata in vzdržujeta Institut informacijskih znanosti v Mariboru (IZUM) in Agencija za raziskovalno dejavnost republike Slovenije (ARRS). Trenutno so v SICRIS-u predstavljene naslednje entitete:

- 881 raziskovalnih organizacij,
- 1408 raziskovalnih skupin,
- 13994 raziskovalcev,
- 5384 raziskovalnih projektov,
- 438 raziskovalnih programov.

Podatki o projektih v sistemu SICRIS so integrirani v evropski informacijski sistem za raziskovalno dejavnost z imenom ERGO, ki povezuje vse evropske informacijske sisteme, ki izpolnjujejo predpisane pogoje. Informacijski sistem SICRIS omogoča tudi pregled predstavitvenih strani več kot 500 projektov okvirnih programov EU neposredno iz baze podatkov Projects sistema CORDIS.

Pri pripravi strukture baz podatkov so upoštevani veljavni mednarodni standardi, klasifikacije in šifranti, priporočila EU (CERIF - Common European Research project Information Format) ter zakonska določila in predpisi, ki veljajo v Sloveniji. Baze podatkov so med seboj povezane, večina podatkov pa je v slovenskem in angleškem jeziku. Omogočeno je iskanje po vseh ključnih poljih.

SICRIS je povezan z informacijskim sistemom COBISS.SI oziroma z njegovo bibliografsko bazo podatkov COBIB.SI, kar omogoča uporabnikom tudi neposreden vpogled v bibliografije raziskovalcev.

Baza podatkov ORGANIZACIJE naj bi vsebovala podatke o vseh raziskovalnih organizacijah, ki od leta 1995 dalje izvajajo projekte, (so)financirane s strani ARRS-ja (do leta 2001 - Ministrstva za znanost in tehnologijo, do leta 2004 - Ministrstva za šolstvo znanost in šport), vendar so popolni le podatki za organizacije, ki so posredovale zahtevane podatke.

V SICRIS se vključujejo tudi organizacije, ki sicer ne sodelujejo pri izvajanju projektov ARRS, če izvajajo raziskovalno - razvojno dejavnost, o kateri poročajo Statističnemu uradu RS in želijo biti z dejavnostjo svojih raziskovalcev/ekspertov predstavljene v SICRIS-u.

Baza podatkov SKUPINE vsebuje osnovne podatke (šifra, naziv, raziskovalno področje) o vseh raziskovalnih skupinah, ki od leta 1998 dalje izvajajo projekte, (so)financirane s strani ARRS-ja (do leta 2001 - Ministrstva za znanost in tehnologijo, do leta 2004 - Ministrstva za šolstvo znanost in šport), druge predstavitvene podatke pa le za trenutno aktualne raziskovalne skupine, če so raziskovalne organizacije posredovale zahtevane podatke. V sestav skupine so vključeni le raziskovalci in strokovni ali tehnični sodelavci, za katere so bili prejeti zahtevani podatki.

Baza podatkov RAZISKOVALCI vsebuje osnovne podatke (šifra, ime in priimek, raziskovalno področje) o vseh raziskovalcih, ki od leta 1998 dalje sodelujejo pri izvajanju projektov ARRS (do leta 2001 - Ministrstva za znanost in tehnologijo, do leta 2004 - Ministrstva za šolstvo znanost in šport) ali pa so njihov aktivni status prijavile raziskovalne organizacije, druge predstavitvene podatke pa le za raziskovalce, ki so zahtevane podatke posredovali in soglašali z njihovo objavo.

Slika 4: SICRIS

V SICRIS se lahko vključijo tudi raziskovalci/eksperti, ki sicer ne sodelujejo pri izvajanju projektov ARRS, če to želijo in posredujejo zahtevane podatke.

Baza podatkov PROJEKTI vsebuje podatke o projektih, ki jih (so)financira ARRS od leta 1998 dalje (do leta 2001 - Ministrstvo za znanost in tehnologijo, do leta 2004 - Ministrstva za šolstvo znanost in šport), vanjo pa bodo vključeni tudi podatki o drugih raziskovalnih projektih, ki jih bodo izvajalci želeli predstaviti.

3.3 IST-World

Portal IST World je rezultat evropskega SSA ("Specific Support Action") projekta IST World, ki je financiran s strani Šestega okvirnega programa (Tehnologija informacijske družbe) Evropske komisije. Številka pogodbe: FP6-2004-IST-3 - 015823.

Portal IST World (http://www.ist-world.org) nudi informacije o ekspertih, raziskovalnih skupinah, centrih in podjetjih, ki so povezana v ustvarjanje tehnologij v okviru informacijske družbe. Pozornost servisa je usmerjena na ekspertizo in izkušnje, ki so relevantne v Evropski Uniji.

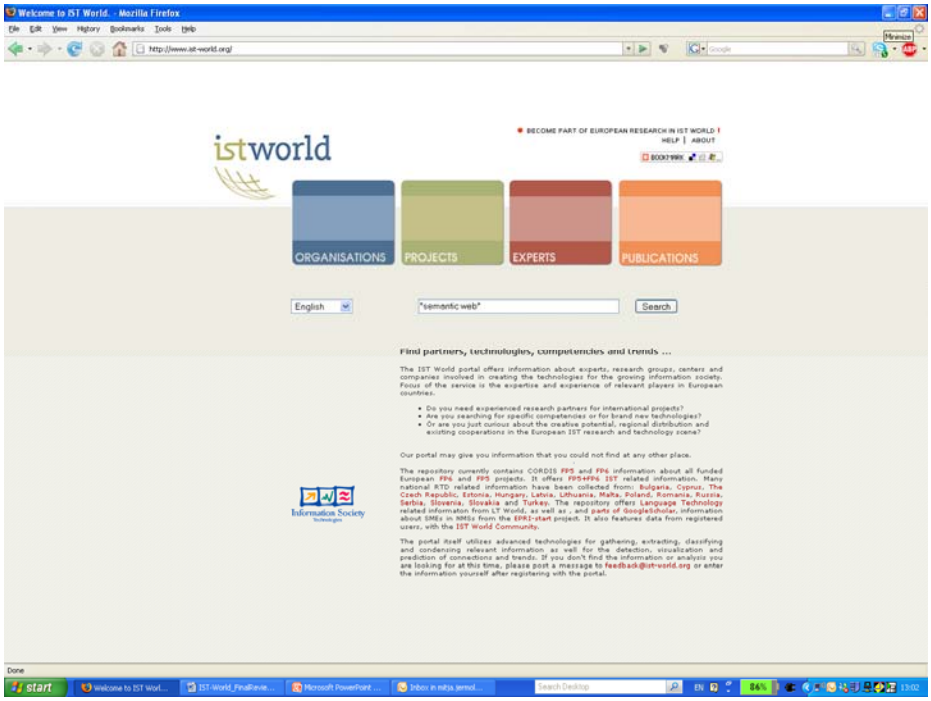
Repozitorij trenutno vsebuje informacije CORDIS OP5, OP6 in OP7, o vseh evropskih projektih financiranih v okviru OP7, OP6 in OP5. Ponuja informacije nanašajoče se na OP5+OP6+IST. Na portalu je bila zbrana velika količina državnih RTD podatkov držav kot so: Bolgarija, Ciper, Češka, Estonija, Madžarska, Latvija, Litva, Malta, Poljska, Romunija, Rusija, Srbija, Slovenija, Slovaška in Turčija. Repozitorij ponuja podatke o Jezikovnih Tehnologijah iz projekta LT World. Prav tako posamezne dele GoogleScholar-ja, informacije o majhnih in srednje velikih podjetjih (SME) in novih državah članicah (NMS) iz EPRI-start projekta. Prav tako vsebuje podatke, ki so jih posredovali uporabniki IST World skupnosti.

Portal uporablja napredne tehnologije za zbiranje, izpisovanje, klasificiranje ter povzemanje relevantnih podatkov, ter detekcijo, vizualizacijo in napovedovanje sodelovanj in trendov. Portal je razdeljen na dva dela:

- na osnovne iskalne in preiskovalne metode ter analitike,
- na kompleksne analitske metode, detekcije trendov ter predikcije.

Portal bo služil kot primer uspešne infrastrukture, ki jo bo uporabljal tudi ASZ. V večji meri že vsebuje vse analitske metode, ki bodo implementirane tudi v AZS. Trenutno portal oskrbi na dan več kot 15000 obiskovalcev in opravlja analize nad podatki o:

- 104893 raziskovalnih organizacijah,
- 43154 projektih,
- 423455 ekspertih oz. raziskovalcih in
- 2032493 publikacijah.



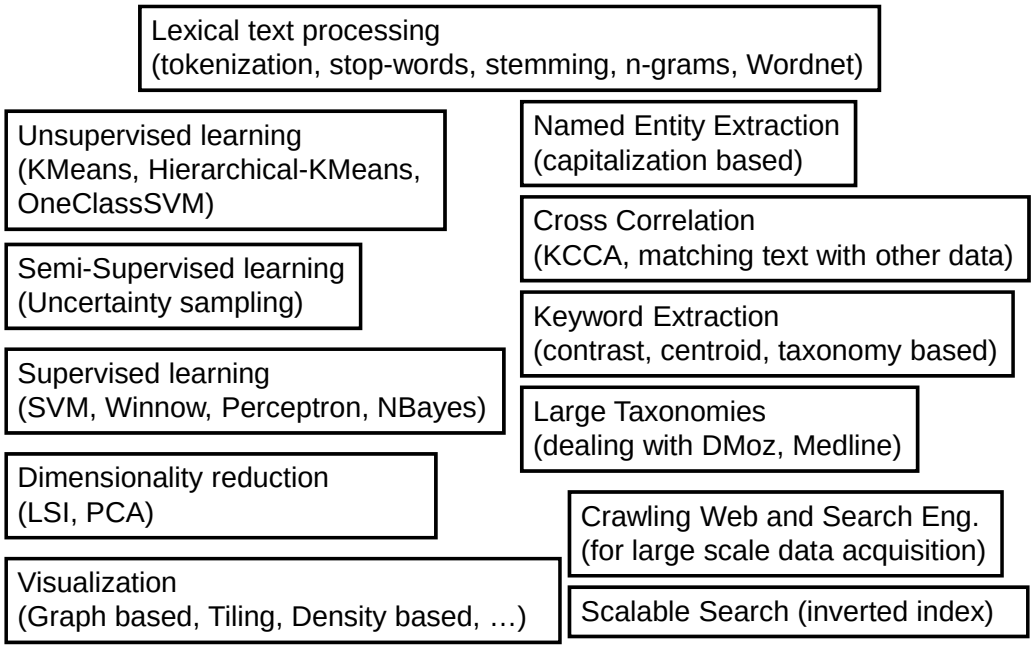
Slika 5: Vstopna stran v IST-WORLD

3.4 TextGarden in Graphgarden

TextGarden (http://ailab.ijs.si/dunja/textgarden/) je knjižnica programskih modulov za obvladovanje strukturiranih, polstrukturiranih in nestrukturiranih podatkov kot so podatki, teksti, slike, podatkovni streami, socialne mreže, ipd. Moduli predvsem pokrivajo naslednje skupine funkcionalnosti (1) analize večmodalnih podatkov kot, (2) medmodalne analize, (3) vizualizacijske metode za kompleksne tekstovne in mrežne podatke, (4) najnovejše analitske metode s področja strojnega učenja, rudarjenja po tekstu, kernel methods, ipd. za primere velikih količin podatkov. Vse metode spadajo pod LGPL licenco in se jih lahko prevede za okolje MS Windows, Linux in ima vmesnike za delovanje v okolju Matlab, Java, Phyton in R.

Razvoj knjižnice se je začel leta 1996 in je rezultiral v set C++ metod za učenje teksta. Pravi razvoj knjižnice se je pričel leta 2003, ko je TextGarden postal centralna knjižnice razvoja skupine več kot 10 raziskovalcev.

Spodnja slika prikazuje osnovne gradnike knjižnice:



Slika 6: Osnovni bloki metod TextGarden knjižnice

Tehnične karakteristike knjižnice so sledeče:

- TextGarden je v celoti napisan v C++ kodi in ga je mogoče prevesti v okolju MS Windows (Microsoft Visual C++, Borland C++) in Unix/Linux (GNU C).
- Teče tako na 32 kot tudi 64bitni platformi.
- Vsebuje več kot 200.000 vrstic kode.
- Do TextGarden knjižnic lahko dostopamo preko:
  - osnovnih C++ razredov,
  - kot DLL knjižnice z več kot 250 funkcionalnostmi,
  - ukazne vrstice z več kot 60. funkcionalnostmi,
  - preko GUI orodij kot sta DocAtlas, OntoGen, ipd.,
  - preko vmesnikov do različnih platform kot sta Java, Python, Matlab, Mathematica, R, Prolog, ipd.

GraphGarden je podobna knjižnica metod za potrebe rudarjenja in modeliranja grafov. Metode so sposobne obdelovati velike grafe, ki imajo do 200 milijonov vozlišč ter več kot 2 milijardi povezav. Tehnične specifikacije so iste kot pri TextGardnu. Dosegljive so na http://agava.ijs.si/~jure/GG/.

Tabela spodaj prikazuje trenutno stanje knjižnic GraphGardna:

alg.h	**basic algorithms for manipulating graphs**
anf.h	Approximate Neighborhood Function for measuring graph diameter. Avoids node sampling and scales to large graphs
bigalg.h	some algorithms for TBigNet
bigg.h	TBigGraph -- big disk based graphs that do not fit into memory
bignet.h	TBigNet -- memory efficient implementation of TNodeNet (avoids memory

	fragmentation)
blognet.h	blog network -- posts on blogs linking each other
casc.h	cascade analysis and counting
cga.h	Community Guided Attachment (see our KDD '05 paper)
clust.h	graph clustering and community finding
cncom.h	extracting connected components
emailnet.h	email network
ff.h	Forest Fire model (see our KDD '05 paper)
ggen.h	basic graph generation models
ghash.h	hash table where key is a graph. Used for counting graphs.
GMine.h	main file
gnet.h	networks (TNodeNet, TNodeEdgeNet)
gproj.h	graph projections (see our WWW '07 paper)
graph.h	graphs (TNGraph, TNEGraph)
gstat.h	calculates various statistics of graphs
gsvd.h	spectral analysis of graphs (singular value decomposition)
gviz.h	interface to GraphViz for plotting small graphs
imdbnet.h	IMDB network
kronecker.h	Kronecker graphs (see PKDD '05 and ICML '07)
layout.h	positions the nodes on the plane for drawing
plots.h	plots graph properties (degree distributions, etc.)
sampl.h	graph sampling (see KDD '06)
timenet.h	time evolving networks

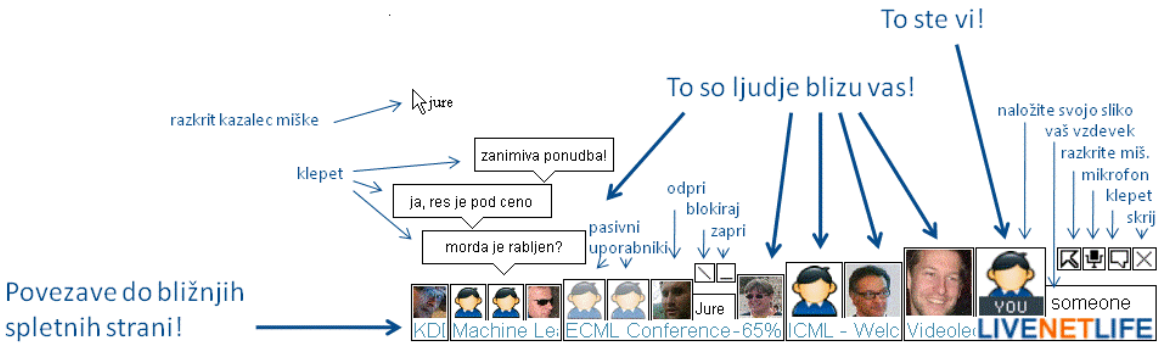
Tabela 1: Knjižnice GraphGarden

3.5 LifeNetLive

Storitev LiveNetLife omogoča kontekstualno določeno povezovanje in sodelovanje v realnem času. Storitev predstavlja novi kanal komunikacije med ljudmi na spletu. Semantični algoritmi v realnem času odločijo, katere spletne strani, dokumenti, e-pošta ali predstavitve, naj vključujejo tudi katere online ljudi, ki tako sestavljajo ad hoc interesno skupino. To je revolucionaren pristop k spletnem komuniciranju in odkrivanju informacij.

Cilj je ponuditi nov način navezovanja stikov prek spleta in oblikovati novo dimenzijo sodelovanja izven trenutno prevladujočih socialnih okvirov. Povezovanje, sodelovanje in medsebojno podporo med uporabniki se ustvari s pomočjo povezave, ki temelji na skupnih interesih. Količina in kakovost odnosov se lahko bistveno povečata kar omogoča izboljšanje zasebnega kot tudi komercialnega sodelovanja. Tradicionalne panoge kot so založništvo, mediji in spletne trgovine ter spletne tržnice lahko s pomočjo LiveNetLife tehnologije razvijejo nove ponudbe in poslovne priložnosti. Podjetja lahko uporabijo LiveNetLife tudi za izboljšanje interne komunikacije in sodelovanja. LiveNetLife je mogoče razumeti kot samostojno storitev, v obliki dodatne aplikacije za določeno okolje (npr. socialne mreže) ali pa kot vgrajeni modul za različne aplikacije.

LiveNetlife storitev je trenutno na voljo kot beta prototip (private beta release). Nameščena je na 3. spletnih mestih in presega 10.000 obiskov na dan. Slika 7 prikazuje ključne komponente grafičnega vmesnika sistema LiveNetLife.



Slika 7: Ključni elementi grafičnega uporabniškega vmesnika LiveNetlife

4 Osnovna struktura modulov aplikacije ASZ

4.1 Analiza razredov konceptov analitskega modula

Na osnovi zahtev in predstavljenih omejitev smo najprej zgradili strukturo domenskih razredov in konceptov, ki jih bomo razgradili v serijo spletnih storitev.

Analiza razredov konceptov (Concept Class Analysis) prikazuje statično strukturo modela sistema. Razred predstavlja množico objektov, ki imajo podobno strukturo, funkcionalnost, obnašanje in relacije. Diagram razredov objektov je metoda, ki tako predstavlja razrede, strukturo razredov, metode, atribute in relacije med razredi. Ker je to statični diagram seveda ne more prikazovati dinamičnega obnašanja sistema.

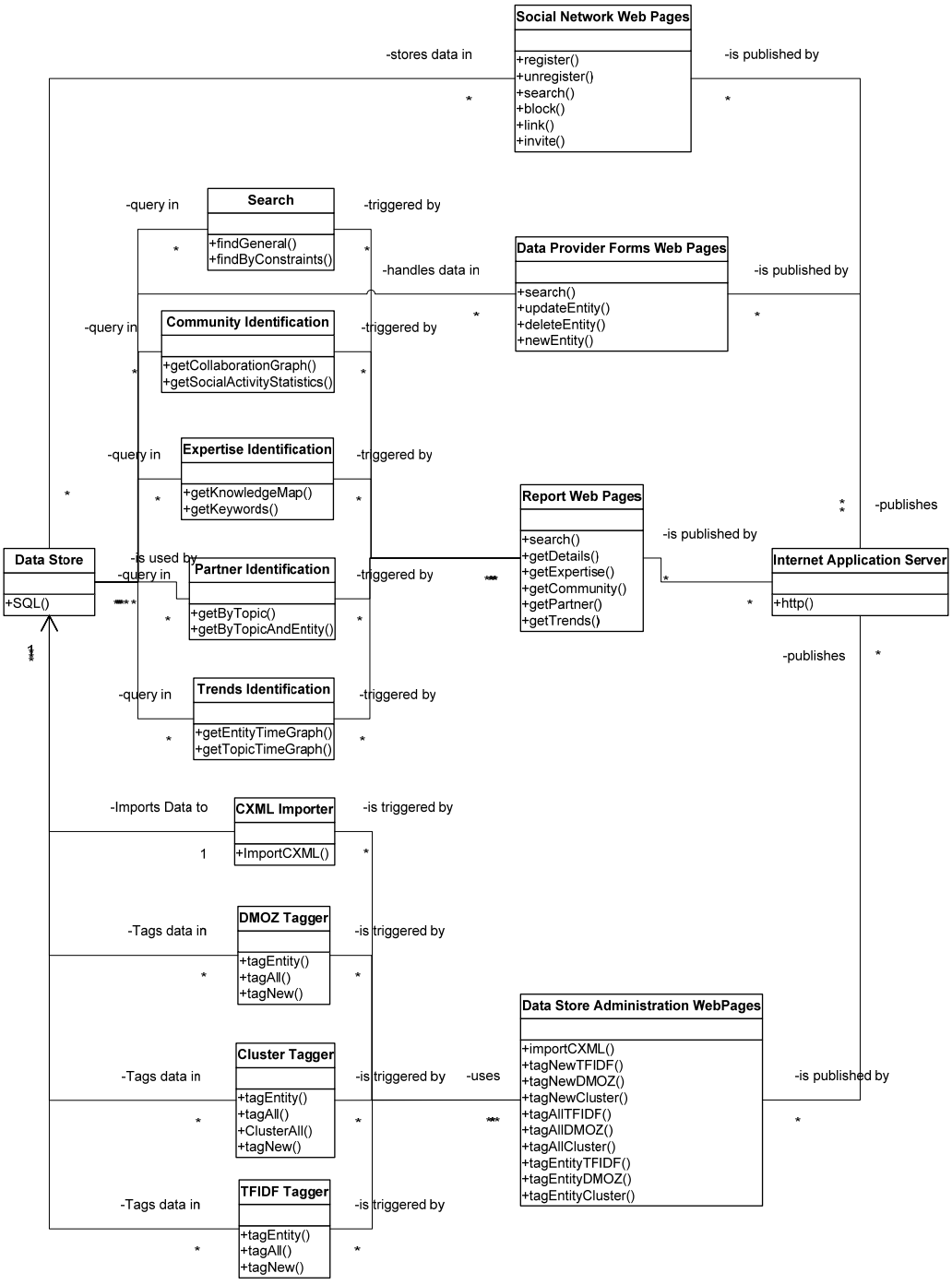
ASZ razredi so podrobneje predstavljeni v Tabeli 1. Diagram na Sliki 2 potem prikazuje funkcionalne dele ASZ rešitve in relacije med njimi.

Št.	Razred	Opis
1	Data Store	Shranjevanje in omogočanje dostopa do podatkov ASZ. Omogoča enostavno, hitro, učinkovito in varno hranjenje podatkov.
2	CXML Importer	Uvaža podatke v ASZ iz XML formata. Podatke parsa in preverja glede na konsistentnost preden vključi Data Store za dejansko shranjevanje podatkov.
3	DMOZ Tagger	Označi tekstovne vsebine glede na DMOZ[1] kategorije.
4	TFIDF Tagger	Označi tekstovne vsebine z ustreznim TFIDF vektorji s ciljem učinkovitejšega in hitrejšega procesiranja.
5	Cluster Tagger	Izgradi samo organizirano hierarhijo tekstovnih vsebin in označi ustrezno vsebino s podatki o mestu v hierarhiji taksonomije.
6	Search	Online optimizacija preiskovalnih stavkov nad hranjenimi podatki.
7	Community Identification	Izvaja optimizirane data mining izračune in preiskovalne stavke na sodelovalnih grafih, ki so shranjeni v podatkovni bazi.
8	Expertise Identification	Izvaja optimizirane data mining izračune in preiskovalne stavke na tekstovnih vsebinah, ki so shranjene v podatkovni bazi.
9	Partner Identification	Izvaja iskanje ustreznih partnerjev ter rangiranje partnerjev glede na text mining izračune.
10	Trends Identification	Izvaja časovno odvisne izračune in evalvacije nad sodelovalnimi

[1] DMOZ – Open Directory Project: The Open Directory Project is the largest, most comprehensive human-edited directory of the Web: http://www.dmoz.com/

		grafi in tekstovnimi vsebinami.
11	Data Administration Web Pages	Omogoča dostop do administracijskih funkcionalnosti administratorju ASZ portala.
12	Report Web Pages	Pripravlja analize in omogoča dostop do preiskovalnih funkcionalnosti ASZ portala do WWW.
13	Data Provider Forms Web Pages	Orodja za omogočanje dostopa preko spleta in upravljanja podatkovne baze vsem trem operaterjem podatkovnih baz (VideoLectures.NET, IST-WORLD in SICRIS) in potencialnih novih podatkovnih baz.
14	Social Network Web Pages	Omogoča funkcionalnosti in integracije s socialnimi omrežji preko spleta.
15	Internet Application Server	Omogočanje dostopa preko http protokola do servisov ASZ portala in omogočanje učinkovitega produkcijskega okolja.

Tabela 2: Razredi konceptov AZS



Slika 8: Analiza razredov

4.2 Definicija osnovnih aplikacij

Med osnovne aplikacije ASZ štejemo predvsem:

- aplikacije pregledovanja in preiskovanja podatkovne baze ASZ,
- vmesnike za prenos podatkov,
- izdelava poročil.

V nadaljevanju bomo vsako od teh na kratko opisali.

4.2.1 Aplikacije pregledovanja in preiskovanja podatkovne baze ASZ

- **Preiskovanje baze podatkov ASZ**: ASZ preiskovalnik bo intuitivna in učinkovita funkcionalnost za preiskovanje ASZ baze podatkov. Omogočala bo:
    - **General Search:** Funkcionalnost predstavlja preiskovanje z enostavnimi preiskovalnimi stavki na bazi ključnih besed. Podobne rešitve poznamo na vseh najbolj popularnih spletnih straneh kot npr. Google.
    - **Advanced Constraint Search:** Ta tip preiskovanja bo temeljil na preiskovalnem mehanizmu, ki bo omogočal definiranje omejitev s katerimi uporabnik omeji prostor iskanja. Tako napredno preiskovanje je namenjeno globljemu preiskovanju in analitiki in je uporabno predvsem za uporabnike z globljim poznavanjem tematike. Omejitvi, ki jih bo metoda upoštevala sta predvsem:
        - **Entity type:** projekti, organizacije oz. raziskovalne skupine, strokovnjaki in publikacije,
        - **Entity description constraints according to the specified entity type:** semantično področje, čas in datum, avtor, geografsko področje, ipd.
- **Pregledovanje vsebine preko taksonomije in klasificiranih objektov.** Za ustrezno taksonomijo bomo na začetku uporabili trenutno največjo javno dostopno taksonomijo DMOZ. V nadaljevanju bomo taksonomijo nadgrajevali glede na nove tematike. GUI bo omogočal tudi enostavno pregledovanje podatkov preko taksonomije.
- **Prikaz podrobnosti in pregledovanje po grafu entitet.** Portal bo omogočal podroben vpogled v vsebino na nivoju različnih granularnosti. Zato bo moral prikazovati osnovne podatke ter različne nivoje agregiranih podatkov. Poleg tega bo moral vse rezultate opremiti z aktivnimi hiperlinki, ki bodo omogočali prehajanje po grafu entitet. Vsaka entiteta je lahko osnovna entiteta (oranizacija, posameznik, objava, projekt) ali pa agregirana entiteta kot npr. tematika, skupina, področje, regija, ...

4.2.2 Vmesniki za prenos podatkov

Tukaj gre predvsem za vmesnike za dvosmerni prenos podatkov med tremi osnovnimi podatkovnimi zbirkami:

- **SICRIS <-> VideoLectures.NET,**
- **SICRIS <-> IST-WORLD,**
- **IST-WORLD <-> VideoLectures.NET.**

IST-WORLD in SICRIS temeljita na CERIF podatkovnem modelu. IST-WORLD je na CERIF 2009 medtem ko je SICRIS na CERIF 2006. Vmesnik med SICRIS <-> IST-WORLD je torej lahko izveden tudi na nivoju

povezav med dvema bazama podatkov. Trenutno nam je IZUM dostavil ustrezen API preko katerega sicer lahko dostopamo direktno do SICRIS baze, ne moremo pa vzpostaviti ustreznega avtomatizma. Podobno poteka enosmerna povezava med SICRIS <-> VideoLectures.NET in sicer kot uporaba API ja za potrebe VideoLectures.NET. Vnos v SICRIS iz VideoLectures.NET poteka ročno. Vmesnik za dvosmerno povezavo med IST-WORLD <-> VideoLectures.NET je že implementiran.

Cilj projekta je vzpostaviti ustrezen režim med tremi bazami podatkov, ki bo omogočal avtomatsko izmenjavo podatkov. Predlagamo izdelavo ustreznih vmesnikov.

4.2.3 Izdelava poročil

Tukaj je predvsem mišljena avtomatska izdelava personaliziranih poročil za vsako entiteto v bazi. S to storitvijo nameravamo predvsem podpreti širše diseminacijske aktivnosti predvsem do poslovnega okolja. Izdelava poročil bo potekala z uporabo MS SQL reportinga.

4.3 Definicija analitskih aplikacij

Analitske aplikacije vključujejo oboje off-line analitiko in online analitiko. Tukaj opisujemo samo najbolj pomembne sklope analitskih aplikacij. Analitike, ki bi jih implementirali za potrebe ARRS in MVZT v ločeno aplikacijo bodo opisane bolj podrobno v naslednjih poročilih.

- **Community Identification:** Avtomatska analitika bo omogočala vpogled v podgrafe, kjer so relacije med entitetami socialne povezave (sodelovanja med entitetami na projektu, publikaciji, raziskavi, ipd). Tako bo uporabnik dobil vpogled v socialne skupnosti raziskovalcev in organizacij. Prav tako bo možno pregledovati socialna omrežja posameznih entitet. Z uporabo analize časovne dimenzije bomo prikazovali grafe zaupanja med entitetami ter s tem uteževali potencialne stabilne socialne grafe. Ti rezultati bodo vhod v storitev definiranja ustreznih konzorcijev oz. skupin, kjer je zaupanje med partnerji ena od cenilk. Za potrebe prikaza bomo uporabili tehnike dinamičnega prikaza grafov.
- **Collaboration Diagram:** To bo eden od pomembnih rezultatov pregleda v modelirane socialne grafe. Graf sodelovanj bo prikazoval socialno mrežo entitet glede na trenutne in pretekla sodelovanja. Vizualizacijska metoda bo omogočala interaktivnost, zoomiranje, predvsem pa bom morala avtomatsko preračunavati najboljše možne postavitve grafov, ki bodo lahko imeli tudi po 100 in več vozlov.
- **Competence garphs:** Na osnovi ekstrahiranih kompetenc iz projektov, publikacij in dosežkov, bo ta aplikacija izračunala kompetenčne grafe za posamezne entitete ter poljubne skupine entitet. Poleg tega bo prikazala relevantne statistike izbranih entitet ali skupin. Entitete, ki so si kompetenčno podobne bodo na grafih pozicionirane skupaj, entitete, ki so si daleč bodo tudi na grafih oddaljene. Entitete bodo na grafu pozicionirane glede na predhodno izračunano pokrajino kompetenc. Taka analiza bo omogočala vrsto različnih interpretacij od iskanja podobnih entitet, razumevanja intenzivnosti raziskav na določenem področju, detektiranja neizkoriščenih raziskovalnih področij do predikcij raziskovalnih gibanj glede na pretekla raziskovanja in raziskovalne politike. Vizualizacijska metoda bo slonela na dinamičnem in interaktivnem prikazu 2D grafa v obliki kompetenčne pokrajine.
- **Expertise Identification:** Na osnovi generiranega kompetenčnega grafa bo ta metoda naračunala in prikazala competence in ekspertize posamezne entitete. To bo utežena lista kompetenc, kjer bo najbolj izrazita kompetenca ocenjena najvišje najmanj izrazita pa

najnižje. S pomočjo sumarizacijskih metod bomo lahko prikazali kratke opise entitet ter poglobljeno študijo razvoja kompetenc. Z aktivnimi povezavami na ostale analitike (predvsem sodelovalne grafe in predikcije) bo lahko uporabnik dobil poglobljen vpogled v posamezno entiteto, ter njen socialni krog.

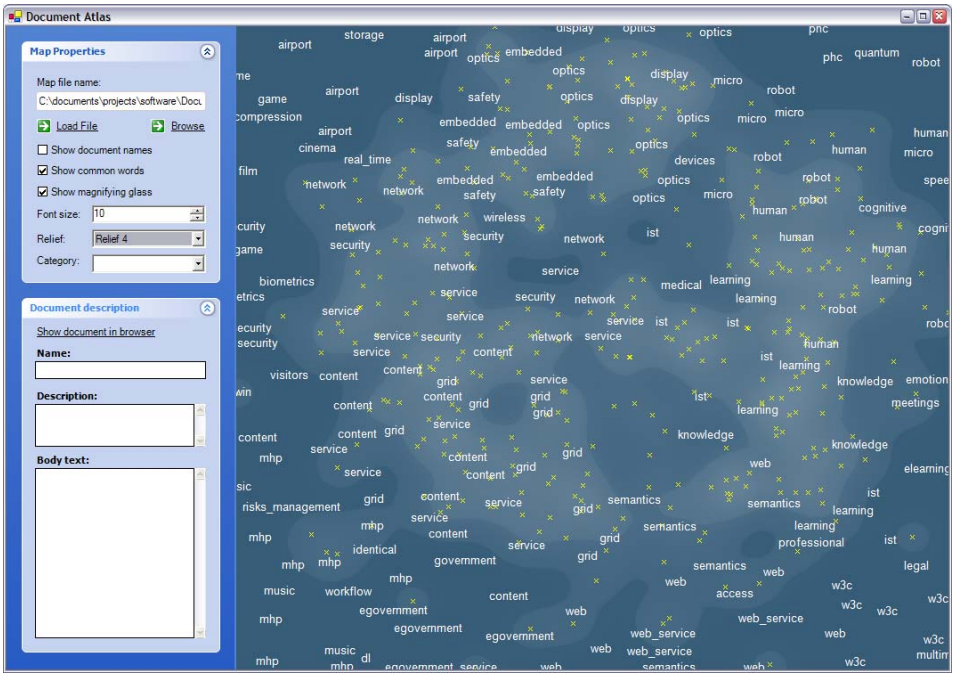
- **Partner Identification:** Ta aplikacija bo s pomočjo avtomatske analize predlagala rangirano listo ustreznih entitet. Utež, ki bo uporabljena za rangiranje bo sestavljena iz več spremenljivk kot so ekspertiza iz iskanega področja, zaupanje v entiteto, pretekli rezultati in kvaliteta rezultatov. Iskanje oz. kriterij iskanja bo definiral uporabnik preko seznama ključnih besed, ki najbolje definirajo ustrezno raziskovalno področje. Ta aplikacija je pomembna predvsem kot pomoč pri iskanju ustreznih partnerjev, ekspertiz in potencialnih konzorcijev za uspešno sestavo ekipe razvojnega ali raziskovalnega projekta.
- **Trends Identification:** Ker vse tri podatkovne zbirke vsebujejo tudi časovno komponento, bomo z ustreznimi metodami analize trendov detektirali pojavljajoče trende in jih predvsem skušali ločiti od trenutnih skokov. Detektirani trendi bodo prikazovali trende v raziskavah, pojavljanje novih raziskovalnih področij, trende v razvoju posameznika, skupin in organizacij, trende na nacionalnem nivoju ipd. Zgodnje odkrivanje trendov je predvsem pomembno za hitro reagiranje, na eni strani preprečevanje problemov in na drugi spodbujanje trendov. Orodje za detekcijo trendov bo prikazovalo tako trende, ki so globalnega značaja in so po vsej verjetnosti že znani kot tudi trende lokalnega značaja, ki so težko sledljivi s klasičnimi orodji in so predvsem tisti, ki lahko povzročijo ključne preskoke v določeni domeni. Trendi bodo prikazani kot časovni grafi odvisnosti domen na različnih nivojih granularnosti.
- **Forecasting:** podobno kot za trende bomo za potrebe napovedovanj gibanj uporabili časovno dimenzijo podatkov. Napovedovati je mogoče gibanja v znanosti, razvoj novih tematik, uspešnost konzorcijev, spremembe v strategijah raziskovalnih organizacijah, ipd. Napovedovanje seveda pomeni, da bo metoda predstavila listo potencialnih napovedi in jih opremila z ustrezno stopnjo zaupanja. Vizualizacija napovedi bo uporabljala podobno metodo kot vizualizacija trendov.

4.4 Definicija vizualizacijskih metod

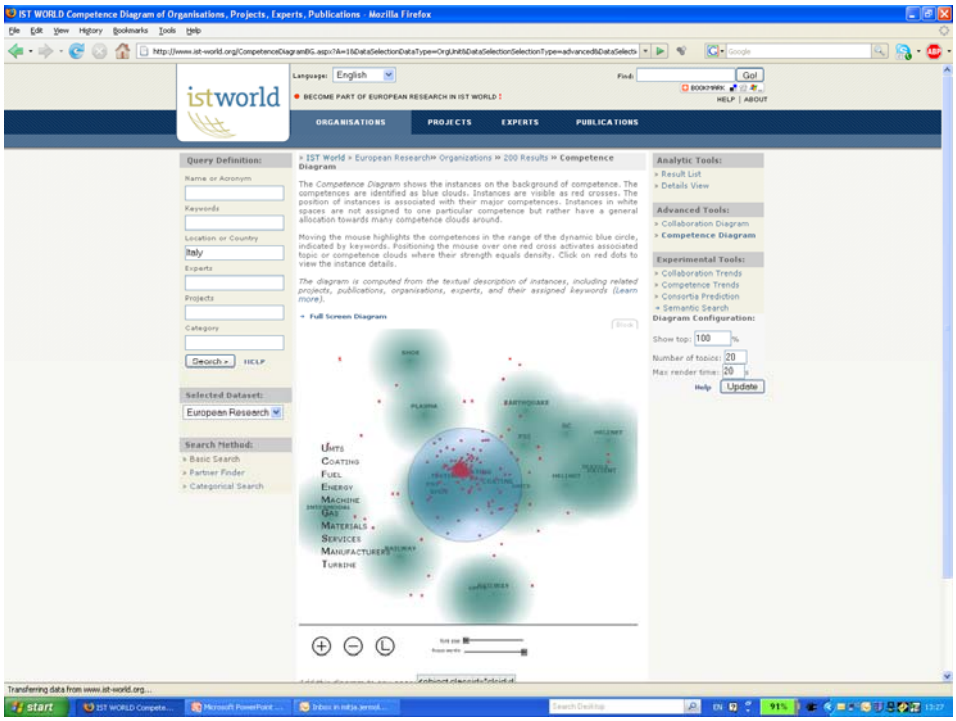
V tem poglavju bomo nanizali nekaj vizualizacijskih metod, ki jih bomo uporabili pri razvoju ASZ. Večina teh metod že deluje na portalu IST-WORLD.org. Za primer ASZ bomo te metode še izboljšali ter uvedli nekaj novih.

4.4.1 Vizualizacija kompetenčnih grafov

V ta namen bomo uporabili metode, ki so bile razvite za potrebe aplikacije DocumentAtlas (http://docatlas.ijs.si). Spodaj prikazujemo dve implementaciji:



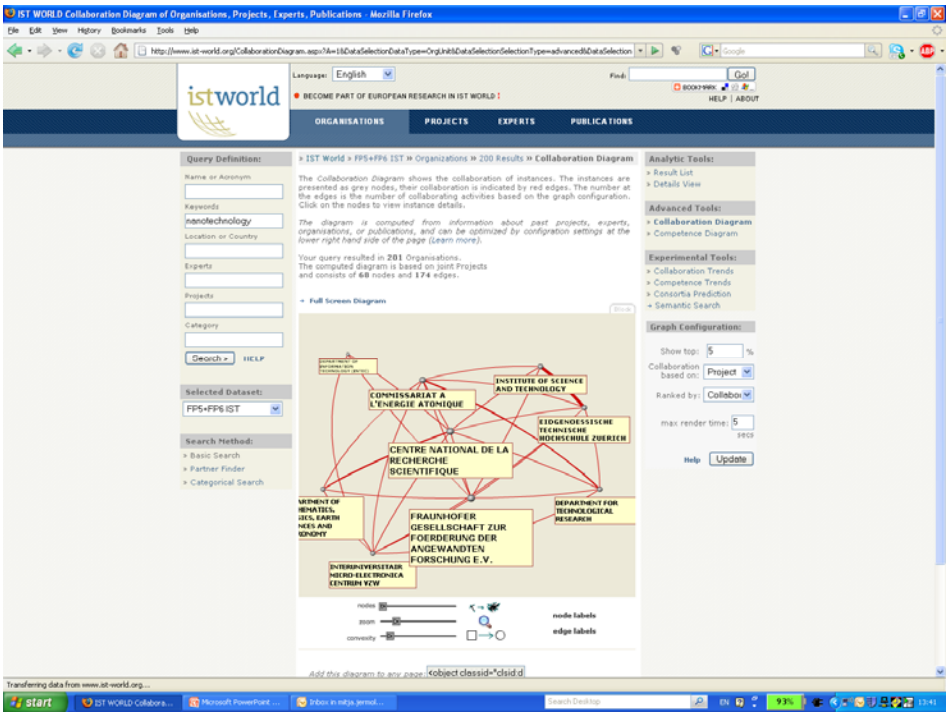
Slika 9: Kompetenčni graf z DocumentAtlasom



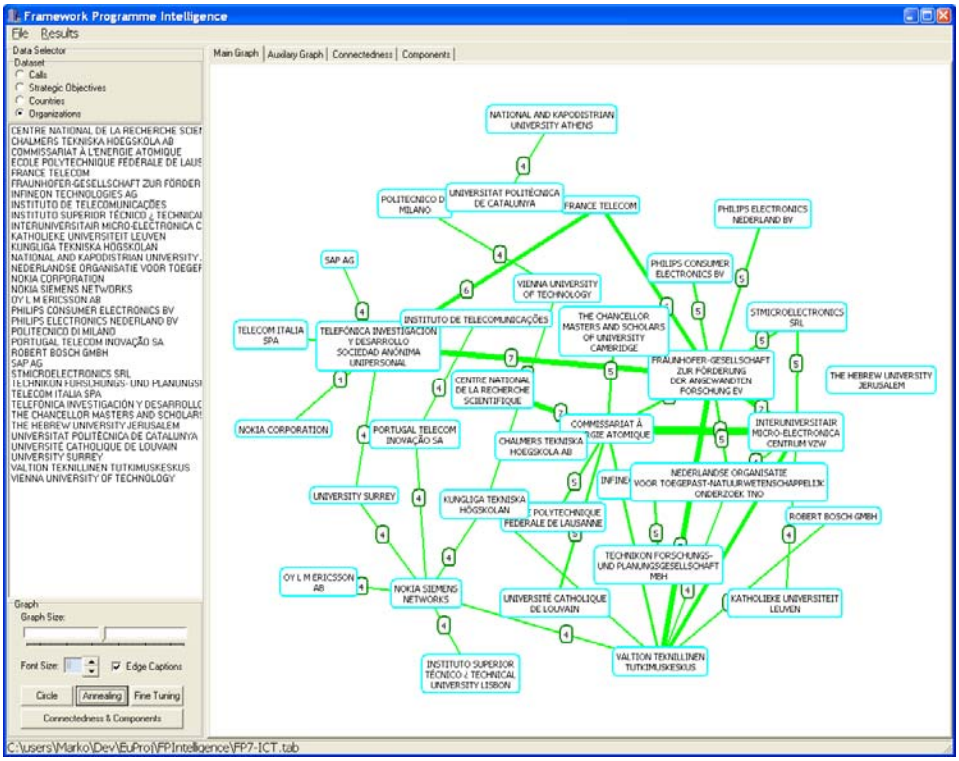
Slika 10: Kompetenčni graf v IST-WORLD

4.4.2 Vizualizacija sodelovalnih grafov

Vizualizacije sodelovalnih grafov slonijo na metodah za modeliranje in analizo grafov. V tem primeru gre za socialne grafe, kjer so vozlišča grafa entitete, povezave med njimi pa sodelovanja.



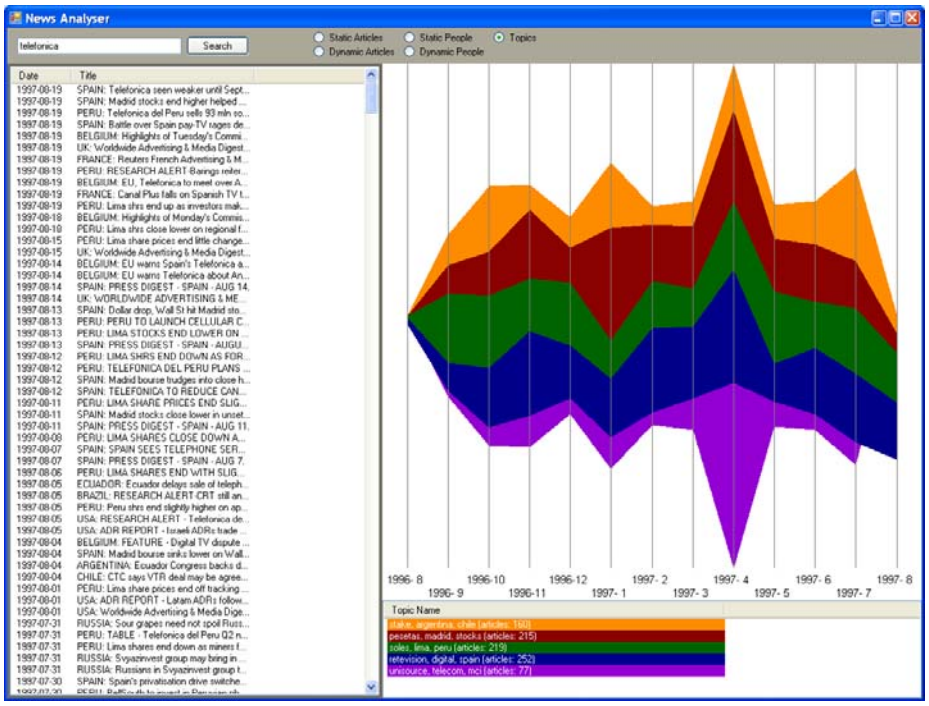
Slika 11: Sodelovalni graf v IST-WORLD



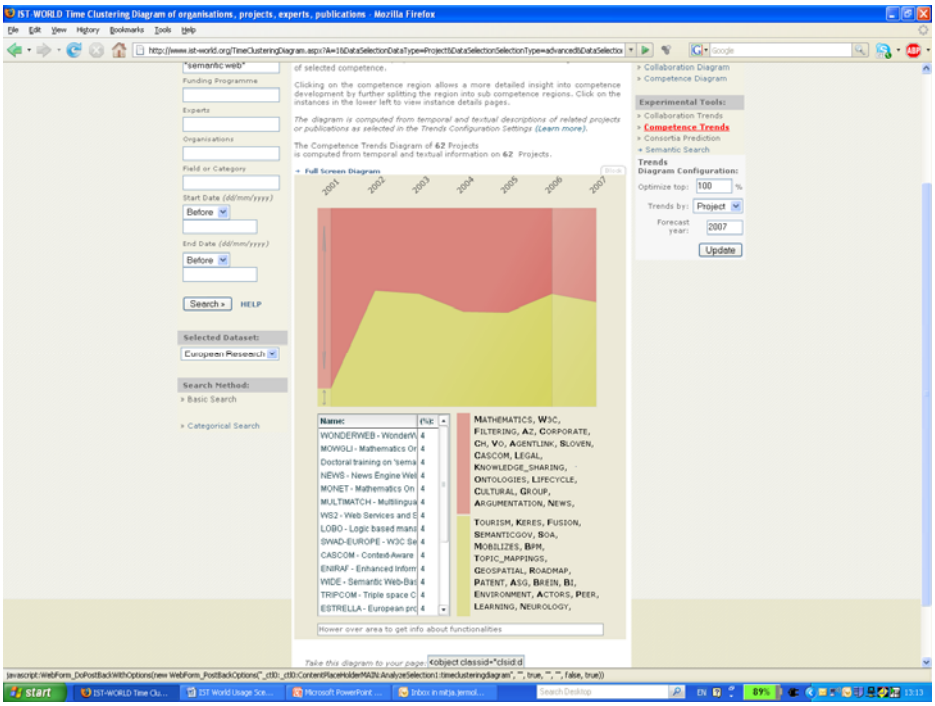
Slika 12: Sodelovalni graf v analitskem orodju za EC

4.4.3 Vizualizacija trendov in predikcij

Orodja za vizualizacijo trendov in predikcij temeljijo na prikazu časovne soodvisnosti vsebin oz. posameznih entitet.



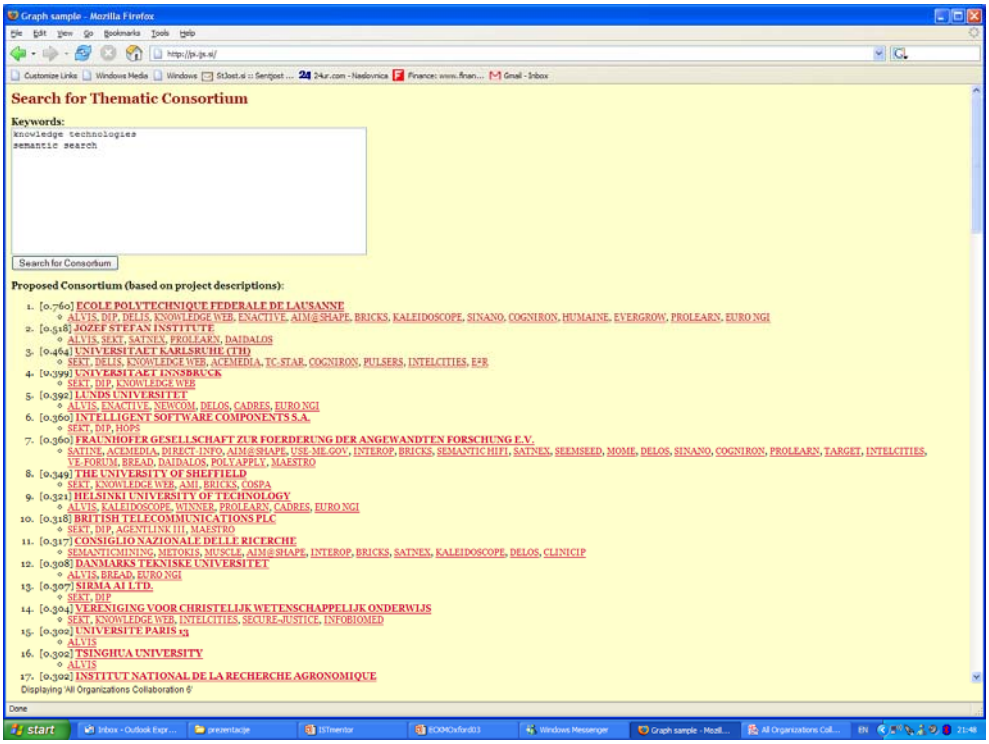
Slika 13: Vizualizacija trendov za NewYorkTimes



Slika 14: Vizualizacija predikcij na IST-WORLD

4.4.4 Rangiranje partnerjev glede na ustreznost

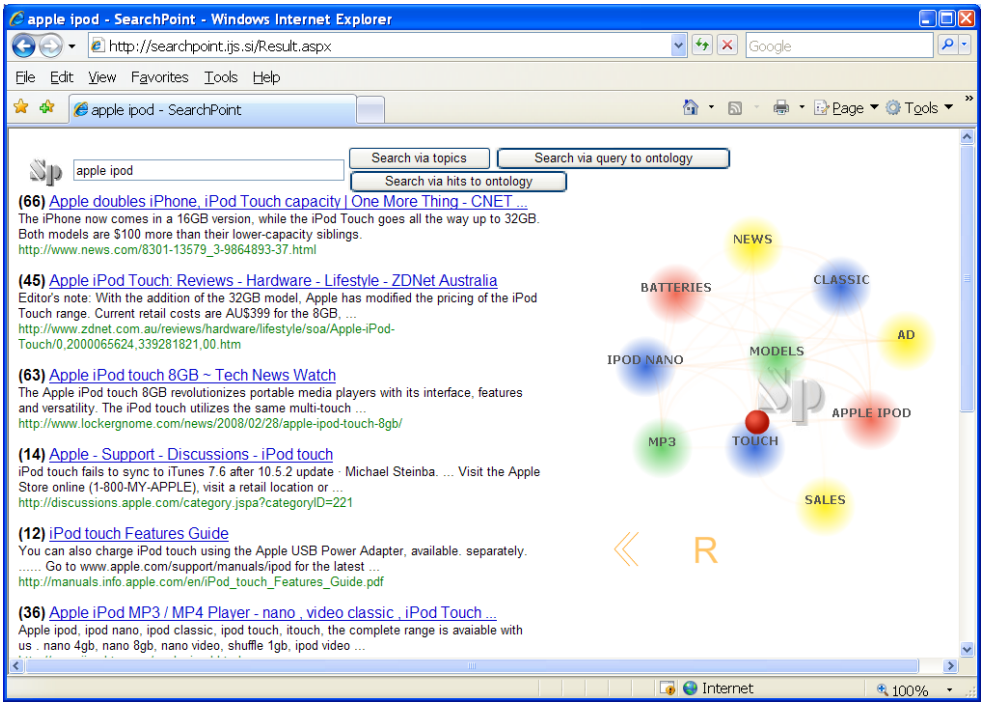
Slika 15: Rangiranje entitet glede na ustreznost sodelovanja



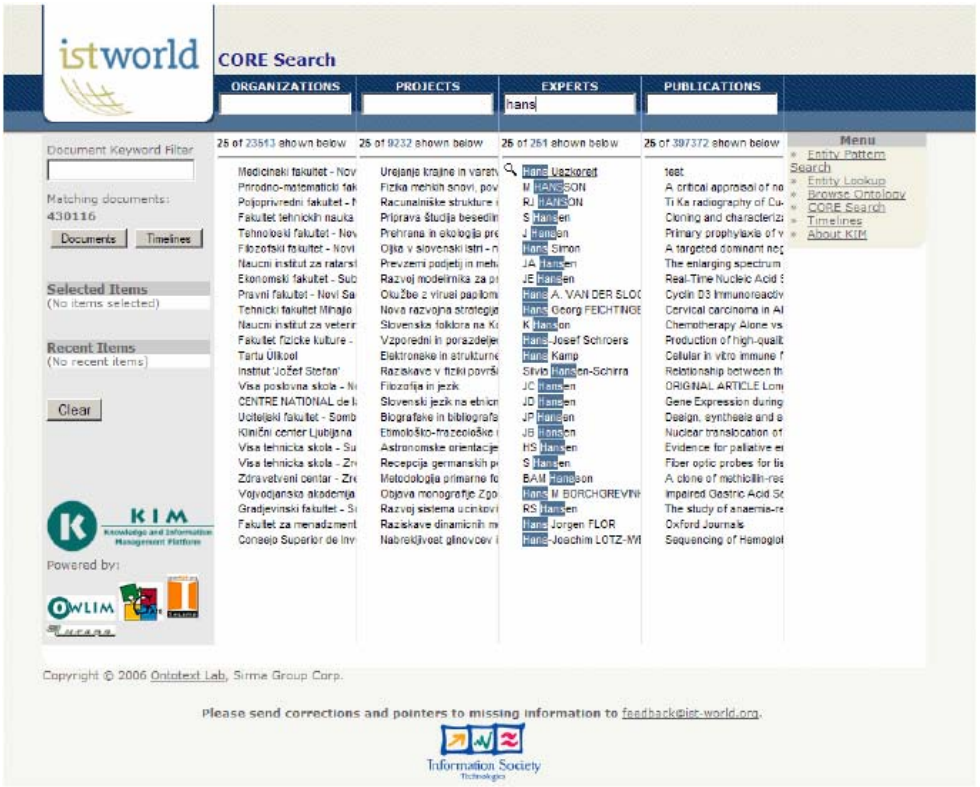
Slika 16: Rangiranje partnerjev za EC

4.4.5 Semantično preiskovanje (SearchPoint)

Semantično preiskovanje z uporabo aplikacije SearchPoint (http://searchpoint.ijs.si)



Slika 17: Semantično preiskovanje s SearchPoint



Slika 18: Preiskovanje preko taksonomije v IST-WORLD

4.5 Virtualni svet slovenske znanosti v katerem se, v živo, srečujejo RTD akterji (raziskovalci, inovatorji, podjetniki, itd.) glede na njihove trenutne interese.

Storitev LiveNetLife omogoča, da v novi enotni informacijski točki, obiskovalce le-te pokažemo eden drugemu, takrat ko iščejo ali se ukvarjajo s podobnimi informacijami o raziskovalnih dejavnosti.

Na primer, v spletnem portalu nove enotne informacijske točke se objavi stran s podatki o projektu ASZ ter več strani s podatki o sodelavcih-znanstvenikih v okviru projekta ASZ. Čeprav so to različne strani znotraj enotne informacijske točke, se s pomočjo storitve LiveNetLife obiskovalci teh strani zaradi sorodnih informacij le-teh začnejo zavedati prisotnosti eden drugega, lahko pričnejo z medsebojnim komuniciranjem in obiščejo spletne strani drug drugega. Tako se spodbudi ozaveščenost in komuniciranje s podobno mislečimi raziskovalci, inovatorji in podjetniki. Prikaz akterjev v obliki grafičnih avatarjev, neposredno v vseh pomensko podobnih informacijskih virih, omogoča uporabniku informacijske točke, da se zave drugih ljudi, ki uporabljajo ali iščejo podobne raziskovalne informacije. Na ta način je ustvarjena ad hoc skupina oz. ekipa relevantnih akterjev. Končni uporabnik tako lahko reši trenutno nalogo kolektivno in zato hitreje.

Storitev poleg omenjenega prikaže tudi spletne povezave do omenjenih informacijskih virov. Te povezave kažejo na vire, ki so uporabljeni v istem trenutku in so pomensko podobni trenutno uporabljenemu viru informacij. Končni uporabnik zato lahko hitreje odkrije relevantne, vendar drugače skrite informacije o svojem problemu.

Zavedanje uporabnika informacijske točke o drugih obiskovalcih, ki uporabljajo podobne vire informacij, v istem trenutku, omogoča uporabniku, da bolj zaupa v informacijsko točko. Prav tako povečuje verjetnost pozitivnega izida reševanja uporabnikovega problema, saj dodaja nove *kolektivne možnosti* reševanja uporabnikovih težav.

Storitev LiveNetLife omogoča, da več virtualnih svetov na podlagi spletnih portalov povežemo v enotni svet, v katerem se obiskovalci portalov začno zavedati drug drugega kadar iščejo ali se ukvarjajo s podobnimi informacijami o raziskovalnih dejavnostih, ne glede na to na katerem spletnem portalu se obiskovalci nahajajo.

Na ta način obiskovalci večkrat obiščejo in dalj časa uporabljajo spletna mesta. Saj so obiskovalci teh spletnih strani deležni izkušnje večjega zaupanja in varnosti, kontekstualne možnosti navigacije in ozaveščenosti o podobno mislečih ljudeh, ta spletna mesta učinkujejo bolj varno, koristno, ustrezno, bolj socialno in manj predvidljivo, torej bolj zanimivo.

Izkušnje ozaveščenosti in komuniciranja med podobno mislečimi znanstveniki in gospodarstveniki omogoča obiskovalcem na enem spletnem portalu, da opazijo podobno misleče obiskovalce na drugem spletnem portalu. Uporaba kontekstualnih opcij navigacije jim omogoča, da obiskujejo drug drugega in na ta način prispevajo k povečanju prometa v obeh spletnih portalih.

Povezovanje spletnih portalov v okviru enotnega virtualnega sveta slovenske znanosti torej spodbuja mednarodno, intergeneracijsko in interdisciplinarno povezovanje uporabnikov portalov.

Storitev LiveNetLife omogoča tudi analizo uporabnikov virtualnega sveta slovenske znanosti za namen iskanja in prepoznavanja reprezentativnih vzorcev skupin uporabnikov, ki delijo zanimanje za avtomatsko identificirane RTD interese.

Porabniki teh informacij bodo tako posamezni RTD akterji (npr. znanstveniki, inovatorji, podjetniki, itd.), ki bodo želeli povečati svoj marketinški domet s ciljnim naslavljanjem identificirane skupine, kot tudi akterji-institucije (npr.: ARRS in MVZT), ki jih bo zanimalo spremljanje učinkov sistemskih ukrepov glede na dinamiko sodelovanja in produkcijo znanstvenih rezultatov.

5 Specifikacija sistema ASZ

5.1 Logična arhitektura

V tem poglavju je prikazana logična arhitektura ASZ. Najprej so navedene logične komponente, ki so nato uporabljene z namenom grupiranja razredov analitik glede na njihove funkcionalnosti. Tabela 2 prikazuje seznam logičnih komponent skupaj z opisom, skupinami razredov in vmesniki. Slika 3 prikazuje logično arhitekturo portala ASZ.

5.1.1 Logične komponente

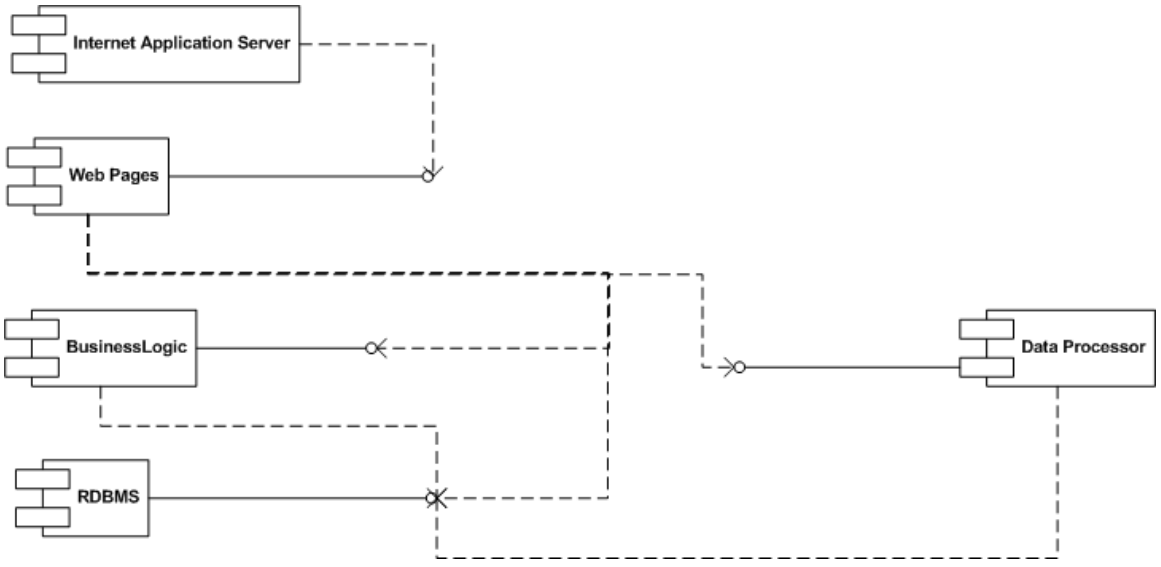
Št.	Komponenta	Opis	Vsebuje razrede	Vmesnik
1	RDBMS	RDBMS izvaja Data Store Class in je zato odgovorna za shranjevanje podatkov ter omogočanje dostopa do podatkov ASZ. Omogoča hiter, učinkovit dostop ter varno hranjenje podatkov.	Aquarius Server	SQL
2	Data Processor	Data Processor komponenta združuje vse razrede, ki so ključni za offline procesiranje in analizo podatkov. Predvsem je odgovorna za uvoz velike količine podatkov ter data in text mining funkcionalnosti na hranjenih podatkih. Biti mora tesno povezana/integrirana z RDBMS komponento.	CXML Importer Cluster Tagger DMOZ Tagger TFIDF Tagger	CXML Importer: ImportCXML Cluster Tagger: cluster TFIDF/DMOZ/Cluster tagger: tagAll tagNew tagEntity
3	Business Logic	Business Logic komponenta vsebuje večino ASZ funkcionalnosti. Uporablja preiskovanje po tekstu, text mining in data mining tehnike na originalnih in predprocesiranih podatkih.	Search Community Identification Expertise	Search: findGeneral findByConstraints Community Identification: getCollaborationGraph GetSocialActivityStatistics Expertise Identification:

			Identification Partner Identification Trends Identification	getKnowledgeMap getKeywords Parter Identification: getByTopic getByTopicAndEntity Trend Identification: getEntityTimeGraph getTopicTimeGraph
4	Web Pages	Web Pages komponenta generira html strani, ki so vidne preko spleta. Uporablja najnovejše razvijalne tehnike in ostalo .NET, APS.NET, ADO.NET z namenom enostavne integracije ter učinkovitega izvajanja operacij.	Data Administration Web Pages Report Web Pages Data Provider Forms Web Pages	Data Administration importCXML tagNewTFIDF tagNewDMOZ tagNewCluster tagAllTFIDF tagAllDMOZ tagAllCluster tagEntityTFIDF tagEntityDMOZ tagEntityCluster Report Search getDetails getExpertise getCommunity getPartner getTrends Data Provider Forms: Search newEntity

			Social Network Web Pages	updateEntity deleteEntity Social Network: Register Unregister Block Link Invite
5	Internet Application Server	Internet Application Server komponenta vključuje razred Internet Application Server. Je odgovorna za stabilno in učinkovito okolje za objavljanje rezultatov preko Web Pages komponente.	Internet Application Server	http

Tabela 3: Logične komponente

5.1.2 Diagram logične strukture



Slika 19: Logična arhitektura ASZ portala

5.1.3 Opis logične strukture

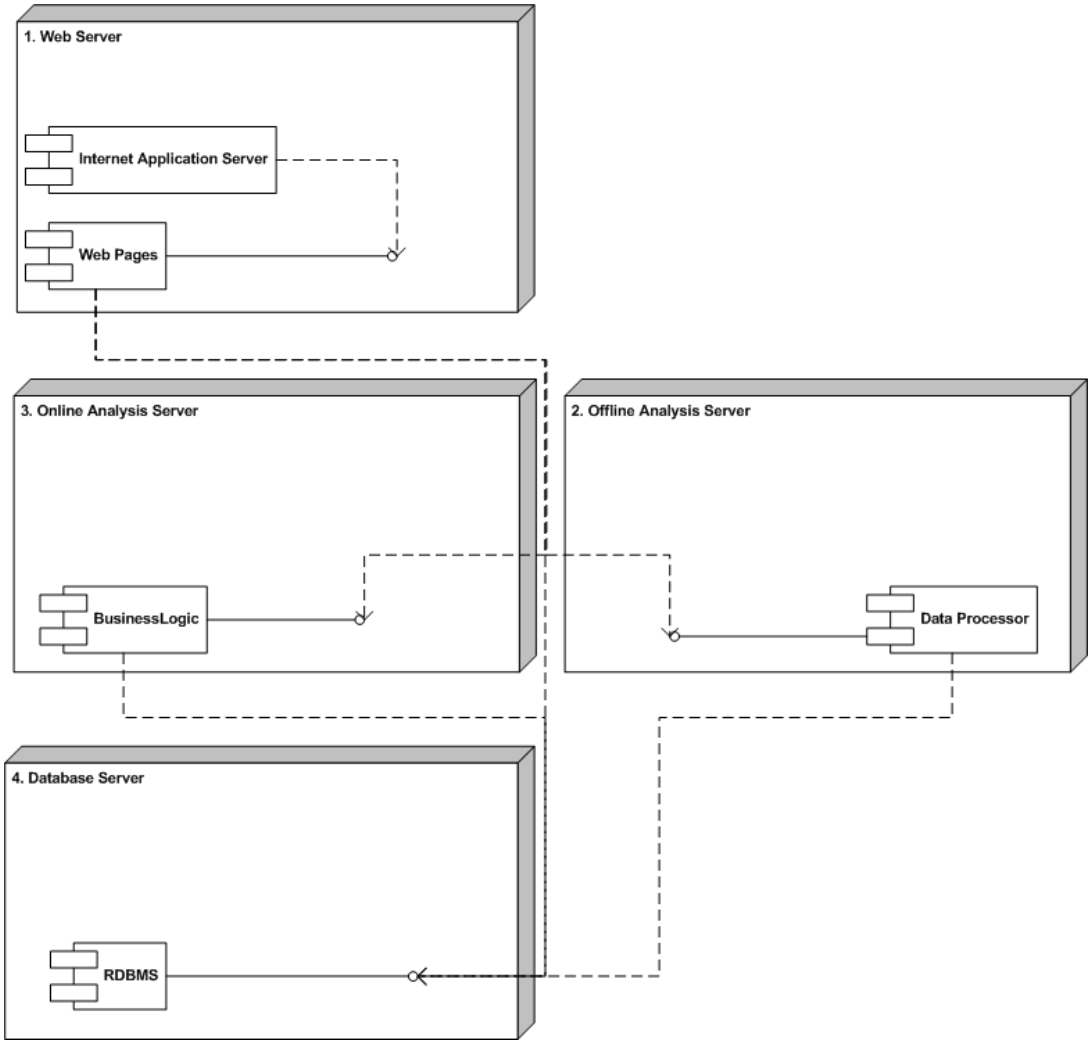
ASZ informacijski sistem je strukturirana kot tri nivojska arhitektura, ki omogoča zahtevano robustnost, zanesljivost, razširljivost, modularnost in centralni nadzor. Izbrana arhitektura omogoča, da je sistem učinkovit in varen. Izbrana arhitektura mora omogočati procesiranje in analizo velikih količin različnih tipov podatkov v realnem času in hkrati varno shranjevati vse te podatke. Predlagana arhitektura temelji na tesni integraciji komponent DataProcessor in BusinessLogic z učinkovitimi mehanizmi RDBMS za izvajanje preiskovalnih stavkov, izvajanje varnosti in omogočanje skalabilnosti. Zaradi dokaj standardne arhitekture omogoča enostavno vzdrževanje in nadgrajevanje z drugimi okolji.

Arhitektura sledi IST-WORLD arhitekturi, ki uspešno servisira več kot 15.000 uporabnikov dnevno z več kot 80.000 projekti, 60.000 organizacijami in več kot 2.000.000 članki.

5.2 Fizična arhitektura sistema

V tem poglavju je prikazana fizična arhitektura ASZ rešitve preko prikaza fizičnih komponent. Vsaka škatlica na sliki 4 prikazuje proces, ki mora podpirati izbrano funkcionalnost. Z namenom doseganja najboljših možnih performanc, bo predstavljena arhitektura izvedena na štirih strežnikih, kot je prikazano na škatlicah na sliki 4. Tabela 3 prikazuje tehnologije na katerih bo slonela arhitektura portala AZS.

5.2.1 Arhitekturni diagram



Slika 20: Fizična arhitektura ASY

5.2.2 Tehnološke osnove

Glede na sliko 4, ki predstavlja fizično arhitekturo ASZ, bodo posamezne komponente slonele na sledečih tehnologijah (Tabela 3):

No	**Component Technology**
1	Microsoft Internet Information Server on MS Windows 7 Enterprise
2	SOAP Web Service inside MS Internet Information Server on MS Windows 7 Enterprise

3	SOAP Web Service inside MS Internet Information Server on MS Windows 7 Enterprise
4	RDBMS MS SQL Server on MS Windows 7 Enterprise 64bit

Tabela 4: Osnovne tehnologije na katerih bo slonel ASZ

5.2.3 Opis fizične strukture

Definirali smo štiri osnovne fizične procese, ki vsebujejo logične komponente ASZ portala. Vse komponente so namenoma izbrane tako, da so del okolja istega proizvajalca programske opreme, da so bile že testirane v realnem okolju intenzivnih analitskih metod in so predvsem del razvojnega okolja IJS. Te komponente so:

1. MS SQL Server. Microsoftova RDBMS je preverjena rešitev, ki vključuje mnogo osnovnih servisov, ki so potrebni za osnovne in napredne analitske metode. Predvsem bomo za potrebe ASZ uporabljali MS preiskovanje po tekstu ter osnovne data mining servise.

2. MS Internet Information Server. Je Microsoftov Web Application server, ki ponuja učinkovito, hitro in enostavno okolje za razvoj aplikacij z uporabo tehnologij kot so ASP.NET in C#. Pomembno je predvsem to, da okolje ponuja tesno integracijo z MS SQL Serverjem preko uporabe tehnologij kot je ADO.NET.

3. On-Line Analysis Server. Process online analitike vsebuje logične komponente za izvajanje vseh online analitik in modeliranja. Tukaj bomo uporabili tehnike in metode, ki so bile razvite in so še v razvoju na IJS. Vse metode so intregrirane kot odprta koda (LGPL) v dve knjižnici za analizo podatkov in tekstov (Textgarden) in analitiko in analizo grafov (Graphgarden). Za integracijo bomo uporabljali standardne spletne servise (Web Service), kar bo omogočalo fleksibilnost, modularnost in integracijo novih komponent. Prav tako bodo storitve iz tega sklopa enostavno dosegljive za katerokoli drugo aplikacijo tudi izven ASZ. Za doseganje performančnih zahtev (zmanjšanje prenosa podatkov) bo ta proces instaliran na istem računalniku kot MS SQL Server.

4. Off-Line Analyis Server. Ta proces bo uporabljal računsko visoko intenzivne operacije kot so npr. supervised and unsupervised data mining. Tudi te metode so bile razvite na IJS in so dostopne v knjižnicah Text Garden in GraphGarden. Vse metode bodo dosegljive kot standardni spletni servisi.

6 Zaključek

Ta dokument predstavlja osnovo za gradnjo sistema Atlas Slovenske Znanosti, ki ima več namenov:

- vzpostaviti enotno točko pregleda razvojno raziskovalnih rezultatov, objav, projektov, kompetenc, organizacij, posameznikov,
- vzpostaviti portal z naprednimi analitskimi metodami za potrebe analize stanja, trendov ter bodočih smernic slovenske znanosti,
- vzpostaviti dvosmerni pretok med tremi osnovnimi podatkovnimi bazami: SICRIS, VideoLectures.NET ter IST-WORLD,
- vzpostaviti okolje za združevanje znanosti in industrije,
- vzpostaviti mrežo portalov z analitskimi komponentami iz AZS.

7 Reference

[1] http://www.ist-world.org

[2] http://sicris.izum.si

[3] http://videolectures.net

[4] http://agava.ijs.si/~jure/GG/

[5] http://ailab.ijs.si/dunja/textgarden/

[6] http://livenetlife.com

[7] Ferlez, J.: Deliverable D2.3, Portal Architecture Specification, IST World: Knowledge Base for RTD Competencies in IST Project No: FP6-2004-IST-3 – 015823, Apr 2005 – Sep 2007, 2006

[8] Ferlez, J., Joerg, B., Yankova, M., Deliverable D5.4, Portal with Advanced Functionalities, IST World: Knowledge Base for RTD Competencies in IST Project No: FP6-2004-IST-3 – 015823, Apr 2005 – Sep 2007, 2007

Atlas slovenske znanosti

RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

R2.1 - Data model and modules for importing, filtering and cleaning data

Ljubljana, 14.9.2011

Table of content

1 Introduction

This document includes three main parts: (1) Definition of possible sources and implementation of CERIF 2008 data model, (2) Building a module for importing data from various sources, and (3) Building a module for filtering and cleaning data.

In the first part new identified data source is introduced, this is LinkedIn professional network on the internet. Next is description of implementing data model of the Atlas of Slovenian Science project. Second part which is about module for importing data into Atlas of Slovenian Science database, describes GetData application which is the implementation of this module. It describes web service used for obtaining the data, details of getting the main concepts, associate concepts and connections between concepts. Third and final part of this document describes the process of cleaning and filtering the data.

2 Definition of possible sources of the data and implementation of CERIF 2008 data model

2.1 Identification of possible sources of data

There are three main sources of data for Atlas project, these are: SICRIS, IST-World and VideoLectures.NET. Additional identified source of data that is LinkedIn professional network on the internet.

2.1.1 SICRIS

SICRIS is the first source of data that will be used in the project. It is an Information System developed by the Institute of Information Science in Maribor and Slovenian Research Agency. The system contains following entities: 901 research organizations, 1429 research groups, 13905 researchers, 5389 research projects, and 438 research programs. (IZUM in ARRS)

2.1.2 IST WORLD

IST World is a portal that offers information about experts, research groups, centers and companies involved in creating the technologies for the growing information society. The repository currently contains CORDIS FP5, FP6 and FP7 information about funded European FP7, FP6 and FP5 projects; as well as some national repositories and collections. (IST World Consortium)

2.1.3 VideoLectures.NET

VideoLectures.NET is a free and open access educational video lectures repository. The lectures are given by distinguished scholars and scientists at the most important and prominent events like conferences, summer schools, workshops and science promotional events from many fields of Science. Repository contains 629 events, 8943 authors, 11531 lectures and 13765 videos. (Center for Knowledge Transfer, Jozef Stefan Institute)

2.1.4 LinkedIn

LinkedIn is a social networking web application that enables publishing of professional information of individuals. »As of August 4, 2011, LinkedIn operates the world's largest professional network on the Internet with more than 120 million members in over 200 countries and territories.« (LinkedIn) LinkedIn offers API (application programming interface) which can be used in Atlas of Slovenian Science project (LinkedIn Developers). Some examples of potential use of LinkedIn network are: using profile information (for e.g. current employment, past employment, specialties, experience, education) for better modeling competences of researchers; using connections of researchers with colleagues, schoolmates and friends to aid modeling of collaboration; using profile picture, company website, emails, etc., to enrich the profile preview of a researcher on the Atlas of Slovenian Science Portal.

2.2 Implementation of CERIF 2008 data model

Data model for this project is in accordance with CERIF 2008 specification. The design objectives for the CERIF 2008 data model are to provide a full CRIS data model with flexibility to allow the majority of existing CRIS to accommodate their own database structure (euroCRIS, 2010, p. 2). CRIS implementations may choose the entities and attributes required for their purpose from formed data model template.

2.2.1 Brief history of CERIF

The Common European Research Information Format (CERIF) was developed under the co-ordination of the European Commission. In its attempt to harmonize national Current Research Information Systems (CRIS) the European Commission funded work on the CERIF 1991 standard. In 2000, the European Commission transferred the custodianship of the CERIF standard to euroCRIS. (euroCRIS)

2.2.2 CERIF features

What follows is a brief, high level description of the latest CERIF release. CERIF has the following design features:

1. Supports people, organizations, projects, funding programs, publications, patents, products, services, facilities and equipment;
2. Provides a fully connected relational data model with powerful, flexible role-based relationships, including recursive relationships to represent hierarchies of people, organizations, projects and funding programs;
3. Supports multiple language attributes;
4. Supports the latest Dublin Core standard.

2.2.3 Implementation of CERIF 2008 within 5 levels of abstraction

The data model template consists of 5 levels of abstraction. These 5 levels with implementation on Atlas of Slovenian Science data model are given below.

Level 1 explains top level entities. These top level entities are: Person, Organization Unit and Project. Atlas data model implements these three main entities with the tables: tblResearchers, tblOrganisations and tblProjects. In addition to these three main concepts described in CERIF, tblLectures id added due to connecting with VideoLectures.NET data source with has lectures as a main concept.

Level 2 includes secondary entities, which are the main associative entities with the top-level entities. In Atlas associative entities are tblRsrEducation (describing education of researcher) and tblRsrClassification (describing field of science in which researcher is working in).

Level 3 includes multilingual support of the data model. Some project may have titles in different languages, or i.e. title in one language and description in several other languages. Atlas data model fully supports entities for translation.

Level 4 contains lookup entities. These are entities with defined set of values. In Atlas data model level 4 entities are tblScienceCodes, tblFieldCodes and tblSubfieldCodes, which contain predefined ARRS (Classifications) classification codes for sciences, science fields and subfields.

Level 5 deals with the many-to-many relationships between entities. Atlas implements this level with entities: tblRsrHasPrj (researchers involved in projects), tblPrjOfOrg (projects of organizations) and tblRsrInOrg (researchers in organizations). There is an additional many-to-many connection between researchers and lectures. This connection is implemented with table tblRsrHasLecture. Figure 1 shows the design data model for the Atlas of Slovenian Science project.

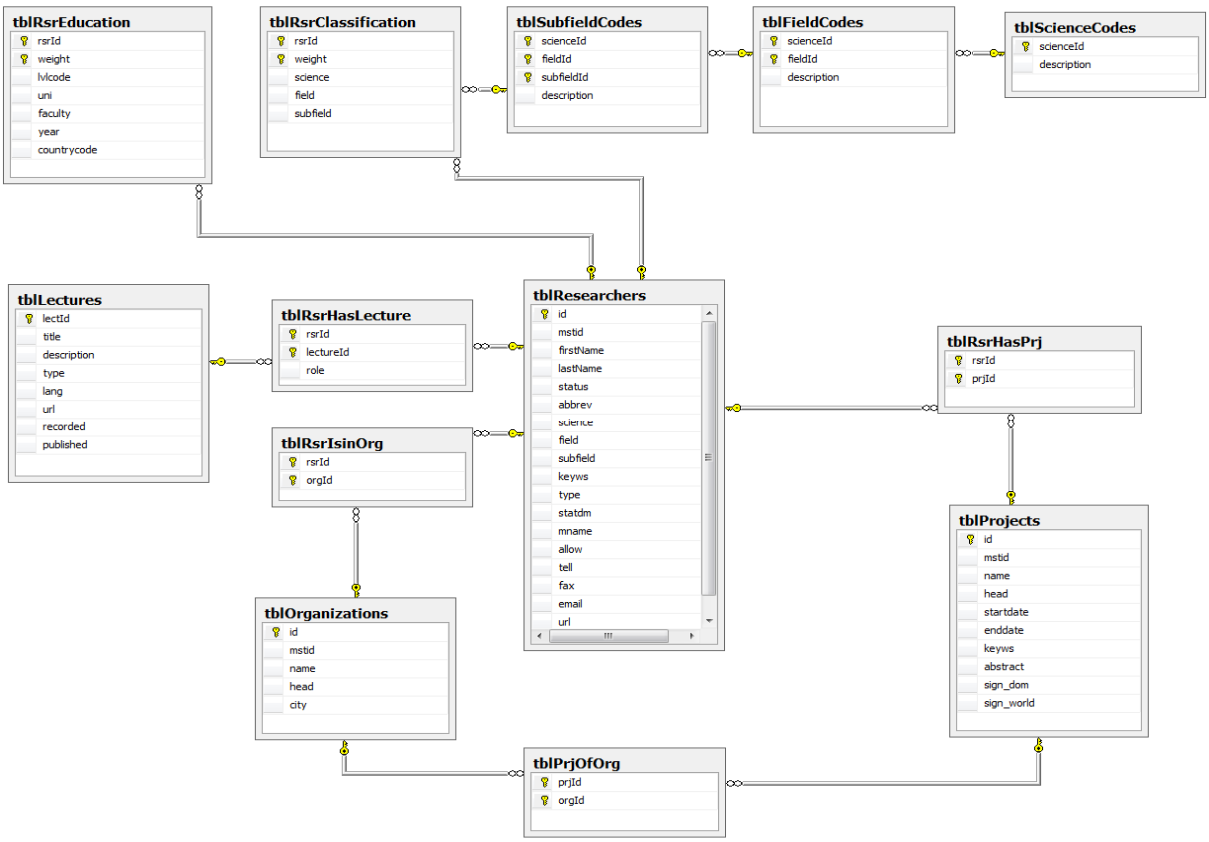


Figure 1 - Atlas data model

3 Building the module for importing data from various sources

Module for importing data from various sources consists of specially build application called GetData that gets the data and prepares it in the format ready for importing into Atlas data repository. To get the data from SICRIS data source, SICRIS web service is used. Process of obtain all the data has three parts. First part is getting the complete lists for the main entities. Second is getting associated concepts of the main ones. Finally, third part is getting concepts which represent many-to-many connections between concepts. Data from VideoLectures.NET portal are obtained in the form of JSON object. Elements of this object are parsed and transformed into SQL commands, ready for importing into Atlas of Slovenian Science database.

3.1 SICRIS Web Service

SICRIS web service enables access to Slovenian research work data. It is located on web address: http://sicris.izum.si/CrisXMLWebServis/CrisXMLWebServis.asmx and it can be accessed only with allowed IP address. Web method used for retrieving results is called - Retrieve. Input parameters of the method Retrieve are: country, entity, methodCall, and fields. Country is a parameter that accepts two letter value label for country. For Slovenia this value is: "SI". Parameter Entity defines the name of the XML element that contains the results. Reserved values for parameter Entity are: "RSR" – for researchers, "ORG" – for organizations, "GRP" – for groups, "PRJ" – for projects, and "PRG" – for programs. MethodCall is a parameter which is used to pass the name of the method to be used for retrieving data. Parameter Fields is used to specify fields which should be included in the results. If this parameter is empty, all fields of the result will be returned.

SICRIS Web Service is included as a Web Reference in a C# application with name – CrisWebService, so it could be used to get the data from SICRIS. Two objects are created: `cd` of type CrisWebService.CrisData used to call the methods of web service and `sr` of type CrisWebService.SearchResults used to receive the results:

```
CrisWebService.CrisData cd = new CrisWebService.CrisData();

CrisWebService.SearchResults sr;
```

These two objects are used in various calls of the Web Service in order to get all the data.

3.2 Getting the main concepts from SICRIS

With calls of the web service, lists of main concepts (researchers, organizations and projects) can be obtained. Since these lists are needed several times in order to get all the data, their identification numbers can be saved in a plain textual file, so they can be quickly loaded and used in the later process. Getting the lists of researchers, organizations and projects is the first step. When these lists are obtained, they can be saved into files with .SQL extension and imported into Atlas database.

To obtain lists of main concepts with GetData application, classes for three main concepts are created (Figure 2). Class Researcher has these 14 properties:

- (int) id – unique identification number used in SICRIS database
- (int) mstid – Slovenian Research Agency identification number
- (string) first_name – first name of the researcher
- (string) last_name – last name of the researcher
- (string) status – status of the researcher. There are four possible values: (1) ACT – active, (2) NAC – unactive, RIP – deceased, RET – retired.
- (string) abbrev – code for the highest degree of education of researcher. There are three possible values: BCD – undergraduate degree, MSD - graduate master's degree, DOD - PhD.
- (string) science – category of scientific work classification
- (string) field – field of category of scientific work classification
- (string) subfield – subfield of category of scientific work classification
- (string) keyws – keywords associated to the researcher
- (string) tel – telephone number of researcher
- (string) fax – fax number of researcher
- (string) email – email of the researcher
- (string) url – web page address of the researcher

Properties of the class Project:

- (int) id – unique identification number for project used in SICRIS database
- (string) mstid - Slovenian Research Agency identification number for project
- (string) name – name of the project
- (int) head – identification number of the head researcher of the project
- (string) startdate – starting date of the project
- (string) enddate – ending date of the project
- (string) keyws – keywords associated to the project
- (string) abst – abstract of the project
- (string) sign_dom – significance of the project for domestic community
- (string) sign_world – significance of the project for world community

Properties of the class Organization:

- (int) id – unique identification number for organization used in SICRIS database
- (int) mstid – Slovenian Research Agency identification number for organization
- (string) name – name of the organization
- (int) head – identification number of the head researcher of the organization
- (string) city – location of the organization

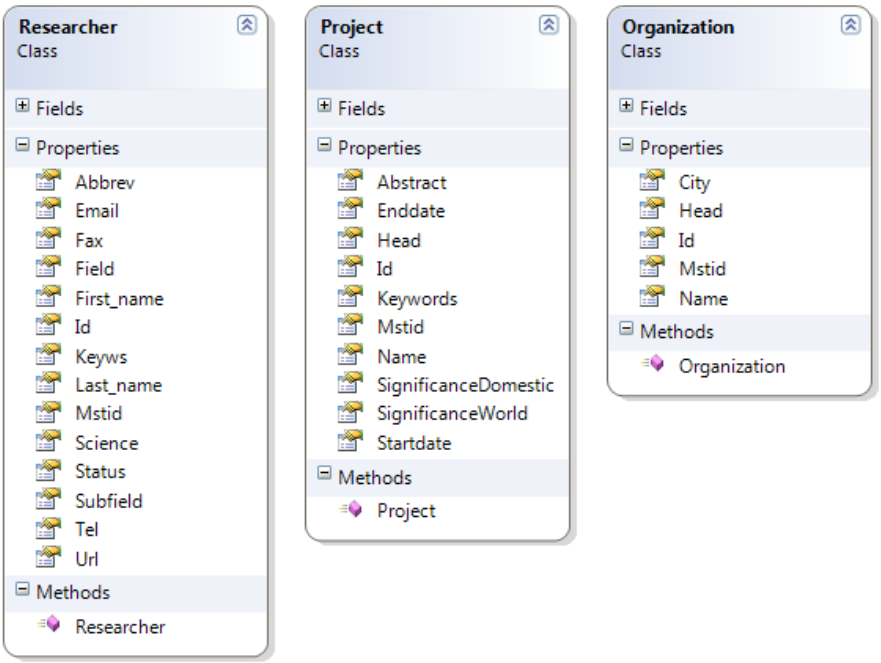


Figure 2 - Researcher, Project and Organization class in GetData application

3.2.1 Obtaining lists of id numbers

List of researcher's id numbers can be obtain using simple SQL query directly in the DBMS (data base management system). In case of this application SQL Server 2008 R2 Management Studio DBMS is used. SQL query for obtaining list of researcher's ids is:

```
select id from tblResearchers;
```

The result is then saved as a tab delimited text file, which has an id number per line. The lists of id numbers of projects and organizations are obtain in the same way using these commands:

```
select id from tblProjects;
```

```
select id from tblOrganizations;
```

Textual files with identification numbers of researchers, projects and organizations are loaded using GetData application and stored in the three lists of integers:

```
List<Int32> rsr_ = new List<Int32>(); List<Int32> prj_ = new List<Int32>();
List<Int32> org_ = new List<Int32>();
```

To obtain researcher, projects and organizations data with GetData application, three lists are created. One of class Researcher, Project and Organization:

```
List<Researcher> rsr = new List<Researcher>();
List<Organization> org = new List<Organization>;
List<Project> prj = new List<Project>();
```

These are the dataobject which will handle the results of method calls, after they are parsed.

3.2.2 Obtaining main attributes of the concept: Researcher

`SearchRetrieve` method of the web service can be called to get researchers with their main attributes. This is an example call:

```
sr = cd.SearchRetrieve("si", "slv", "RSR", "mstid=%", 1, 40000, "", "");
```

First parameter is country, it defines country of the researchers (`si` is Slovenia). Second parameter is language and defines language of the results (`slv` is Slovenian). Third parameter is entity and defines name of the main node which will contain results (the node will be named `RSR` in the above example). Fourth parameter is query (in this case it returns all researchers). Fifth parameter is called currentPage and sets the page of the returned result. Sixth parameter is pageSize and it sets the maximum number of rows per page. Seventh parameter is sort, this parameter defines sorting of the results (if it is empty no additional sorting of the query results is performed). Finally, eighth parameter is fields and defines which fields will be returned as result (if it is empty, all fields will be returned).

`sr` is the instance of class `SearchResults` and it is used to receive the results of the function call. The results are in the XML format.

This XML result is then parsed in order to extract values of attributes and elements. These values are attributes: id, mstid and stat of RSR node; nodes: fname and lname nodes; and attribute code of elements: abbrev, science, field and subfield. Using these vales a new object of the class Researcher is added into the list of researchers:

```
rsr.Add(new Researcher(id, mstid, fname,
lname,stat,abbrev,science,field,subfield," "," "," "," "," "));
```

Even though the last five parameters of the researcher object are missing, the existing ones are sufficient for creating initial inserts into the database. By performing Save Researchers operation in the GetData application, list of researchers is saved as a file containing SQL insert commands. Following example shows two lines of the file for importing list of researchers:

```
INSERT INTO tblResearchers VALUES('1','19334','John','Smith',
'ACT','DOD','N','02','07');

INSERT INTO tblResearchers VALUES('2','21654','Joan','Jett',
'ACT','DOD','D','01','01');
```

3.2.3 Obtaining main attributes of the concept: Project

Projects are retrieved using the same method (SearchRetrieve), with different value of entity parameter (PRJ):

```
sr = cd.SearchRetrieve("si", "slv", "PRJ", "name=\"\"", 1, 40000, "", "");
```

Similar to researchers, project are not obtained with all of attributes, therefor project objects are created with some missing empty parameters, which are obtained with other method calls:

```
prj.Add(new Project(id, mstid, name, head, startdate,
enddate,"","","","",""));
```

Example of SQL commands to insert projects into database table tblProjects are:

```
INSERT INTO tblProjects VALUES('6668','J1-3608','Preslikave na
algebrah','5615','1.5.2010','30.4.2013');

INSERT INTO tblProjects VALUES('6859','V5-1027','Odnos do znanja v družbi
znanja','9080','1.10.2010','30.9.2012');
```

3.2.1 Obtaining main attributes of the concept: Organization

Organizations are retrieved using the same SearchRetrieve, with `ORG` as a value of entity parameter:

```
sr = cd.SearchRetrieve("si", "slv", "ORG", "name=%", 1, 40000, "", "");
```

In contrast to researchers and projects, organizations are obtained with complete set of attributes, which means there is no need for additional method calls. Organization object are created with all parameters in the following way:

```
org.Add(new Organization(id,mstid,name,head,city));
```

These are exemples from generated SQL file used to insert organizations into database:

```
INSERT INTO tblOrganizations VALUES('2340','500','Institut informacijskih
znanosti','24011','Maribor');

INSERT INTO tblOrganizations VALUES('559','106','Institut ''Jožef
Stefan''','6589','Ljubljana');
```

3.2.2 Obtaining additional attributes of the researchers

Five attributes of researcher object are missing, because they are obtained with the different method call of the web service. One of these missing attributes is Keyws. Keywords are obtained by calling `GetKeywords` method of the web service:

```
sr = cd.Retrieve("si", "RSR_DATA",
"Sicris_app_UI.Researcher.GetKeywords.slv." + rsr_[i], "");
```

The example shows that the method is called for every researcher, using previously obtained list of researcher identification numbers. Obtained keywords are exported into file as SQL UPDATE commands:

```
UPDATE tblResearchers SET keyws='motorna vozila, pogonski sistemi motornih
vozil, pretvorniki fizikalnih veličin' WHERE id = '6576';

UPDATE tblResearchers SET keyws='Mehanika tal, zemeljska dela, terenske
preiskave, laboratorijske preiskave' WHERE id = '5033';
```

Other four missing attributes are contact information of a researcher. This attributes are obtained using GetContacts method:

```
sr = cd.Retrieve("si", "RSR_DATA",
"Sicris_app_UI.Researcher.GetContacts.slv." + rsr_[i], "");
```

Contact information is added to researchers in the database table tblResearchers with the SQL UPDATE commands as in following examples:

```
UPDATE tblResearchers SET
tel='12',fax='124',email='john.smith@email.si',url='' WHERE id= '1';

UPDATE tblResearchers SET
tel='223',fax='224',email='joan.jett@email.si',url='' WHERE id= '2';
```

3.2.3 Obtaining additional attributes of the projects

Attributes of projects which need to be obtained with additional function calls are: keywords, abstract, domestic significance and world significance. Project abstracts are obtained with following function call:

```
sr = cd.Retrieve("si", "PRJ_DATA", "Sicris_app_UI.Project.GetKeywords.slv."
+ prj_[i], "");
```

Table tblProjects is updated with keywords with UPDATE commands as in following examples:

```
UPDATE tblProjects SET keyws='zmanjšana zmožnost, merjenje, zdravje,
mednarodna klasifikacija' WHERE id= '4327';

UPDATE tblProjects SET keyws='nitrat, rastlina, talna voda, podtalnica,
gnojenje, namakanje, izotopi' WHERE id= '4284';
```

To obtain abstracts of projects, this function call is performed:

```
sr = cd.Retrieve("si", "PRJ_DATA", "Sicris_app_UI.Project.GetAbstract.slv."
+ prj_[i], "");
```

Abstracts of projects are then updated with UPDATE commands like this:

```
UPDATE tblProjects SET abstract='Zasnova principov enotnega metajezika, ki
opisuje skladenjske, semantične...' WHERE id= '4410';

UPDATE tblProjects SET abstract='Kakovost tablet je predvsem odvisna od
njihove sestave, izgleda in dimenzij...' WHERE id= '4407';
```

Last two attributes which need to be obtained are domestic and world significance of the project. These are obtained with method call:

```
sr = cd.Retrieve("si", "PRJ_DATA",
"Sicris_app_UI.Project.GetSignificance.slv." + prj_[i], "");
```

The generated UPDATE commands for significance look like this:

```
UPDATE tblProjects SET sign_dom='Pomen naših raziskovalnih rezultatov je
takšen, kot ga imajo rezultati domače...' sign_world= 'Pri modeliranju
kolonije bakterij gre za pionirski poskus opisa rasti bakterij kot aktivnih
sistemov...' WHERE id= '5672';
```

3.3 Getting the associated concepts from SICRIS

Associated concepts are those connected with main concepts (typically with one-to-many type of relation). Associated concepts in Atlas data model are: Education and Classification, therefore two classes for these concepts need to be created in the GetData application (Figure 3). Class Education has these 7 properties:

- (int) rsrId – unique identification number of researcher
- (string) weight – consecutive number of gained education
- (string) lvlcode – code for level of education. It can be: BCD – undergraduate diploma, MSD master degree, DOD – PhD, SPC - specialization
- (string) uni – name of the university
- (string) faculty – name of the faculty
- (int) year – year when education level was accomplished
- (string) countrycode – code of the country of educational institution

Class Classification has these five properties:

- (int) rsrId – unique identification number of researcher
- (string) science – category of scientific work classification
- (string) field – field of category of scientific work classification
- (string) subfield – subfield of category of scientific work classification
- (string) weight – number of researchers classification (some researchers have more classification of their scientific work)

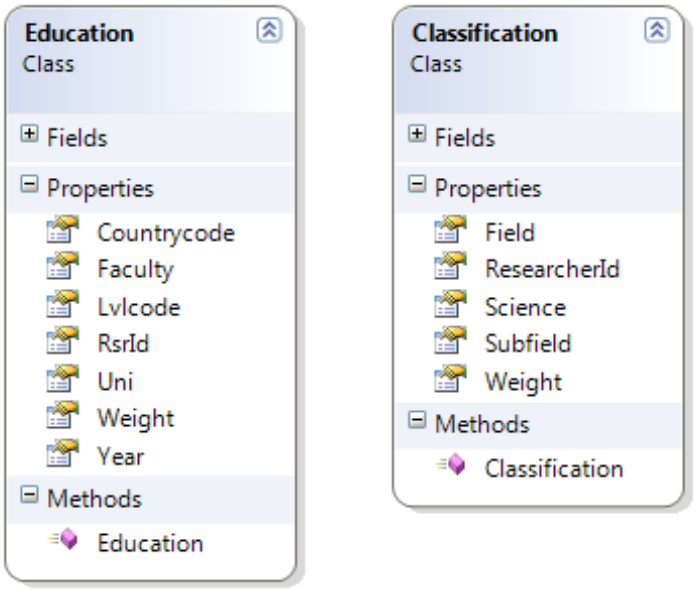


Figure 3 - Eduaction and Classification class in GetData application

To handle education and classification results, two lists are created. One called classi of the class Classification for the classification data, and one called edu of class Education for the education data:

```
List<Classification> classi = new List<Classification>();
List<Education> edu = new List<Education>();
```

3.3.1 Obtaining attributes of the concept: Education

Attributes of the class Education are obtained using the method GetEducation from SICRIS Web Service. This method is in GetData application is called in the following way:

```
sr = cd.Retrieve("si", "RSR_DATA",
"Sicris_app_UI.Researcher.GetEducation.slv." + rsr_[i]+".%", "");
```

After the XML result is paresed, new instances of the class Education are added into the list `edu`:

```
edu.Add(new Education(baseid, weight, edulvl, uni, faculty, year,
countrycode));
```

Using the edu list, SQL INSERT commands can be generated and exported into file, which can be used to import the data into the table tblRsrEducation. These are examples of the INSERT commands:

```
INSERT into tblRsrEducation values ( '14107','2','MSD','Univerza
Harvard','Harvard Law School','2001','US');'
```

```
INSERT into tblRsrEducation values ( '4580','3','DOD','Univerza v
Ljubljani','BF - Gozdarski oddelek','1977','SI');
```

3.3.2 Obtaining attributes of the concept: Classification

Classification data are obtained using the method call GetClassification, in the following way:

```
sr = cd.Retrieve("si", "RSR_DATA",
"Sicris_app_UI.Researcher.GetMSTClassification.slv." + rsr_[i], "");
```

The results are parsed and stored into `classi` list:

```
classi.Add(new Classification(baseid, science, field, subfield, weight));
```

Finnaly they are exported into SQL file in form of INSERT commands like these:

```
INSERT into tblRsrClassification values ('30415','1','4','02','01');
```

```
INSERT into tblRsrClassification values ('14996','1','5','02','02');
```

3.4 Getting the connections between concepts from SICRIS

Last step in obtaining the complete data set is getting the many-to-many relations between concepts. Many-to-many type of relation is implemented in the way that it is divided into two relations one-to-many, with adding a new concept which connects these two relations. First step is creating classes for the concepts that implement the connections. These concepts are: (1) researcher has project – class RsrPrj is created, (2) researcher is in organization – class RsrOrg is created, and project of organization – class PrjOrg is created. Created classes are shown in Figure 4. Each of these classes has only two properties. These properties are identification numbers of concepts which are connected.

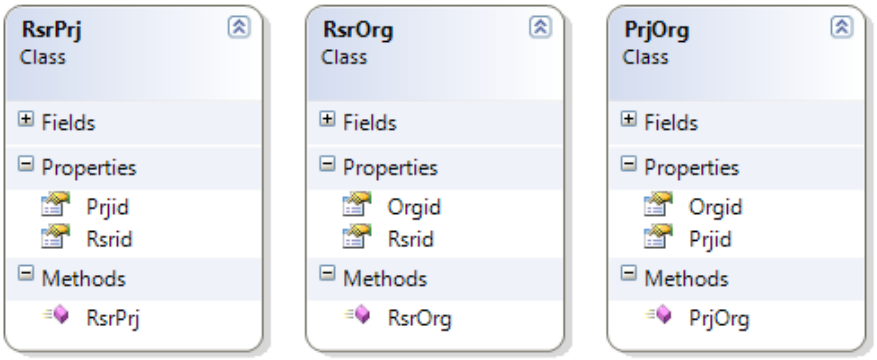


Figure 4 - Classes for relations between main concepts in GetData application

After the classes are created, the instances of these classes are made in the form of list, in order to capture the results:

```
List<RsrPrj> rsr_prj = new List<RsrPrj>();
List<RsrOrg> rsr_org = new List<RsrOrg>();
List<PrjOrg> prj_org = new List<PrjOrg>();
```

3.4.1 Obtaining the connections between researchers and projects

Connections between researchers and project are obtained by calling a method to get all the projects for each researcher. All researchers are listed using previously obtained list of researcher's ids. Implemented method call looks like this:

```
sr = cd.Retrieve("si", "RSR-PRJ",
"Sicris_app_UI.Researcher.GetProjects.PRJ.slv." + rsr[i].Id, "");
```

Parsed results are added into list `rsr_prj`:

```
rsr_prj.Add(new RsrPrj(baseid,
Convert.ToInt16(node.Attributes["prjid"].Value)));
```

These are some of the lines of resulting file for inserting connections into the database:

```
INSERT INTO tblRsrHasPrj VALUES('5134','6040');

INSERT INTO tblRsrHasPrj VALUES('4291','5224');

INSERT INTO tblRsrHasPrj VALUES('4291','5212');

INSERT INTO tblRsrHasPrj VALUES('4291','6040');
```

It can be seen from the above example that researcher with id: 5134, is on the project with id: 6040; and researcher with id 4291 is on projects with ids: 5224, 5212 and 6040 (project on which first researcher is also on).

3.4.2 Obtaining the connections between researchers and organizations

Method call that returns all researchers of an organization is called for every organization, using previously obtained list of all organizations. Method call is implemented in this way:

```
sr = cd.Retrieve("si", "RSR-ORG",
"Sicris_app_UI.Organization.GetResearchers.slv." + org[i].Id, "");
```

The results are then inserted into list `rsr_org`:

```
rsr_org.Add(new
RsrOrg(Convert.ToInt32(node.Attributes["rsrid"].Value),baseid));
```

SQL commands for inserting connections between organizations and researchers look like this:

```
INSERT INTO tblRsrIsinOrg VALUES('3964','559');

INSERT INTO tblRsrIsinOrg VALUES('3955','559');

INSERT INTO tblRsrIsinOrg VALUES('3955','3060');
```

We can see that researchers with id numbers 3964 and 3955 are in the organization with id number 559, but the researcher 3955 is also in organization 3060.

3.4.3 Obtaining the connections between organizations and projects

Final connection to obtain is the one between organizations and projects. These connections tell us which projects have each organization. Again, each organization can have more projects and each project can be assigned to more organization. Therefore, this is many-to-many type of connection which is stored using additional concept PrjOrg. The connections are obtained using this method call:

```
sr = cd.Retrieve("si", "PRJ-ORG",
"Sicris_app_UI.Organization.GetProjects.PRJ.slv." + org[i].Id, "");
```

The results are stored in the list `prj_org`:

```
prj_org.Add(new PrjOrg(Convert.ToInt32(node.Attributes["prjid"].Value),
baseid));
```

Finnaly, INSERT commands are generated, like in these example:

```
INSERT INTO tblPrjOfOrg VALUES(7,'698');

INSERT INTO tblPrjOfOrg VALUES(8,'585');

INSERT INTO tblPrjOfOrg VALUES(8,'586');

INSERT INTO tblPrjOfOrg VALUES(8,'785');
```

3.5 Getting data from VideoLectures.NET

Data from VideoLecuters.NET source are obtained in form of JSON objects. Three file were obtained: (1) collection of authors of videos, (2) collections of videos and (3) collection of connections between authors and lectures.

Atlas of Slovene Science analyses Slovenian scientific community, therefor only Slovenian authors should be imported into the database. Authors in the VideoLectures.NET database do not have attribute country, so it is need to identify only Slovenian authors and lectures. There are different

strategies for filtering the data which are described in chapter 4 of this report: Building module for filtering and cleaning data.

There are other issues that occur with importing the data from VideoLectures.NET source into Atlas of Slovenian Science database, for instance, even thou authors and researchers come from two different tables form two different datasets, they represent the same concept. Those two tables have also some different attributes and in some cases the same attributes have different attribute values. These types of issues are discussed in chapter 2: Building module for connecting data modules of the report R22: Module for connecting Data objects and module for exporting and data indexing.

3.5.1 Collection of authors

Collection of authors is given as a file in JSON format. This is an example of file content:

```
{
"A1":{"url":"izidor_golob",
      "name":"Izidor Golob",
      "gender":"M",
      "organization":"Institute of Informatics, Faculty of Electrical
Engineering and Computer Science, University of Maribor",
      "email":"izidor.golob@uni-mb.si",
      "refs":{
            "homepage":"http://lisa.uni-mb.si/osebje/IzidorGolob.htm"
      }
},
"A2":{"url":"mirko_golobic",
      "name":"Mirko Golobič",
      "gender":"M",
      "organization":"Association for Near Death Studies",
      "email":"bogomir.golobic@krka.si",
      "refs":{
            "homepage":"http://lkm.fri.uni-
lj.si/xaigor/slo/znanclanki/obsmrtna.htm"
      }
},
"A3":{"url":"marko_grobelnik",
      "name":"Marko Grobelnik",
      "gender":"M",
      "organization":"Artificial Intelligence Laboratory, Jožef Stefan
Institute",
      "email":"marko.grobelnik@ijs.si",
      "refs":{
            "homepage":"http://ailab.ijs.si/marko_grobelnik/"
      }
…
}
```

As it can be seen from the above example each author is a one JSON object, which has properties: url – defines internal url for the VideoLectures.NET portal; name – includes both first and last name of a author; gender – it can be character F for female or M for male authors; organization – full name of author's organization, which includes department; email and refs – containing home page url.

In order to parse this collection in GetData application and generate SQL commands for importing the data into Atlas of Slovenian Science database, Json.NET framework was used (CodePlex).

3.5.2 Collection of lectures

Collection of lectures is given in a JSON file format, which looks like in this example:

```
"L5160":{   "lang":"en",
      "url":"iswc07_daquin_watson",
      "type":"lecture",
      "views":191,
      "enabled":1,
      "public":1,
      "recorded":"2007-11-13T18:15:00",
      "published":"2008-06-27",
      "text":{
            "title":"The Watson plugin for the Neon toolkit",
            "desc":"Watson is a Semantic Web gateway…
      }
}
```

Lectures have these properties: lang – language of the lecture; url – internal url on VideoLectures.NET portal; type – defines if this is a lecture, event or project, views – number of views of the lecture; enabled – flag defining if the lecture is enable; public – flag defining if the lecture is publicly available; recorded – date and time when the lecture was recorded; published – date when the lecture was published, text – has two child properties: title – title of the lecture and desc – description of the lecture.

This data is parsed in the GetData using Json.Net framework, just like in case of collections of authors.

3.5.3 Collection of connections between authors and lectures

Third and final file of the VideoLectures.NET source is a file that contains connections between authors and their lectures. These connections are also given in JSON format. This is an example of file's content:

```
"L46":{     "parent":"L2",
      "authors":{"A10":"author"}
},
"L47":{     "parent":"L3",
      "authors":{"A29":"author"}
},
"L48":{     "parent":"L2",
      "authors":{"A30":"author"}
},
"L49":{     "parent":"L5",
      "authors":{"A1224":"coauthor","A31":"author","A1264":"coauthor"}
},
```

From the above example, it can be seen that each lecture can have a parent event which can be event or a project. This is defined by the property parent of each connection. Second property is authors, this property defines who the author is and which type of connection he has with this lecture.

3.6 Getting data from IST World

IST World is a web portal that integrates many data sources. These include automated data collection of: CORDIS FP7, CORDIS FP6, CORDIS FP5, Slovenian Dataset, GoogleScholar, and Pascal; national repositories or collections: Bulgarian repository, Estonian CERIF-based repository, International Language Technology repository (LT World), Czech Republic collection, Cyprus collection, Hungarian collection, Latvian collection, Lithuanian collection, Polish collection, Romanian collection, Slovak collection, Turkish collection. Focus of Atlas of Slovenian Science project is on Slovenian scientific community, therefore Slovenian Dataset is data source of interest, from the IST World data collections. Slovenian Dataset is obtained by automatic crawling of the pubic SICRIS website. (IST World Consortium) Atlas of Slovenian Science project has a direct link to the data from SICRIS portal. These data can be automatically obtained and update to the Atlas database, using the procedures described in chapters 3.2, 3.3 and 3.4 of this report. Since the data in the IST World is currently not updated and there is a better source for the same data available (namely SICRIS and COBISS), data from IST World is currently not imported into Atlas of Slovenian Science. Nevertheless, IST World is a valuable data source, because it integrates data from such a large variety of sources. Moreover, data is modeled in according to CERIF data model, which makes importing of data into Atlas of Slovenian science repository a simple task if desired in the future.

4 Building module for filtering and cleaning data

4.1.1 Filtering Slovenian authors VideoLectures.NET collection

Data obtained from VideLectures.NET data source contain full set of authors and lectures. These records are not dived into Slovene authors and lectures, and those from other countries. Atlas of Slovene Science analyses Slovenian scientific community, therefor only Slovenian authors should be imported into the database. Different approaches are employed to automatically filter the data to obtain only Slovene authors and their lectures.

First approach is using the lists of Slovenian authors generated for the purpose of Slovene Research Agency (Slovenian Research Agency). This list is publicly available on the URL: http://videolectures.net/site/list/authors/?language=sl. The list is used in the GetData application with which data are imported and connected into Atlas of Slovenian Science database. After each author is parsed, his or her name is searched on the list of Slovene authors. If no matches are found the author is considered a foreign author.

Second applied approach is comparing the values of attributes from VideLectures.NET and SICRIS data sources. Data from both sources is loaded with GetData application (this is explained in chapter 3 of this report), in which is filtering also performed. Values of email, first name and last name attributes of authors from VideoLectures.NET source are compared with values of the same attributes for researchers from SICRIS source. If the values match on email, the author is accepted as a Slovene, since to persons cannot have two identical emails. If the values match on first and last name, but do not match on email, author has to be checked manually. If the emails do not match, it can still be the same person with changed or different email. But, the fact that the person matches in

first and last name is not reliable enough evidence that it is the same person, since the cases of matching in the name are not very often, it is feasible to check that cases manually.

5 Conclusion

This report covers the definition of possible sources and implementation of CERIF 2008 data model. SICRIS, IST World and VideoLectures.NET are shortly described and LinkedIn is introduced as a new potential source of data. CERIF 2008 and its implementation for the Atlas of Slovenian Science data model are given next.

Second part describes building of a module for importing data from various sources. This module is implemented with the GetData application. Web service used for obtaining the data from SICRIS is described with details of implementing methods to retrieve data from SICRIS data source. Import of data fom IST World and VideoLectures.NET is also described in this part.

Third part of this document describes the process of cleaning and filtering data, why this process is needed and how the problem of filtering Slovene authors is performed.

6 References

[1] Center for Knowledge Transfer, Jozef Stefan Institute . (n.d.). *VideoLectures.NET*. Retrieved from http://videolectures.net/

[2] *Classifications*. (n.d.). Retrieved 8 29, 2011, from Slovenian Research Agency: http://www.arrs.gov.si/en/gradivo/sifranti/sif-vpp.asp

[3] CodePlex. (n.d.). *Json.NET*. Retrieved 09 10, 2011, from http://json.codeplex.com/

[4] euroCRIS . (n.d.). *euroCris*. Retrieved 09 10, 2011, from http://www.eurocris.org/

[5] euroCRIS. (2010). *CERIF 2008 - 1.2 Full Data Model (FDM).*

[6] IST World Consortium. (n.d.). *IST-World* . Retrieved from http://www.ist-world.org/

[7] IZUM in ARRS. (n.d.). Retrieved 03 20, 2011, from SICRIS: http://sicris.izum.si/

[8] LinkedIn. (n.d.). *LinkedIn*. Retrieved 09 2011, 10, from http://press.linkedin.com/about

[9] LinkedIn. (n.d.). *LinkedIn Developers*. Retrieved 09 2011, 10, from https://developer.linkedin.com/

[10] SICRIS. (2011). *CERIF 2008 - 1.2 Full Data Model (FDM).*

[11] Slovenian Research Agency. (n.d.). *ARRS*. Retrieved 8 4, 2011, from http://www.arrs.gov.si

[12] The Apache Software Foundation. (n.d.). *Lucene.NET*. Retrieved 09 10, 2011, from http://incubator.apache.org/lucene.net/

Atlas slovenske znanosti

RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

R2.2 - Module for connecting Data objects and module for exporting and data indexing

Ljubljana, 14.9.2011

Table of content

1 Introduction

This document has two main parts. The first part is about a module for connecting different data modules. Reasons for building of this module are explained. Process of generating connections between concepts from two different sources is also given. The second part describes a module for exporting and indexing data. Exporting is done via web service, with implementation of various methods which export data in xml format. Indexing is performed using Lucene.NET framework. Process of creating an index is given, together with description of index structure.

2 Building module for connecting data modules

Connecting data modules is performed in the GetData application with which data is imported from various sources. To import the data obtained from VideoLectures.NET source into the Atlas of Slovenian science database, connecting of modules is needed.

One reason for this is that VideoLectures.NET gives data about authors of lectures and the SICRIS gives data about researcher, even thou researcher and author are the same person in many cases. The connection of modules is performed in such a way that the authors are considered to be researchers which have a connection to one or more lectures. Therefore there is no need for a new table in the Atlas data model.

Another reason for module for connecting data modules are situations in which the attributes of the same real world concept, coming from different sources, have different values. Module for connecting data modules has to decide which value of the attribute is valid and which will be imported into database. Similar like in previous case, SICRIS is dominant source over VideoLectures.NET when metadata of researcher is taken into account.

2.1.1 Connecting researchers and lectures

Here is described how researchers which are imported from SICRIS get connected with lectures, using the connections between authors and lectures from VideoLectures.NET.

All the steps are performed in the GetData application. Loaded are: list of all researcher from SICRIS and lists of all authors, lectures and connections from VideoLectures.NET. Each connection between author and lecture is parsed. It is examined if the author exists in list of researchers (this is explained in the chapter 4: Building module for filtering and cleaning data, of the report R2.1: Data model and modules for importing, filtering and cleaning data). If the author is a Slovene researcher that is in the table of researcher tblResearchers, new connection between researcher and lecture is created. Before creating the connection, it is check if the lecture from the connection already is in the table of lectures. In case it is not, first an SQL INSERT command is generated which imports the lecture from the connection in the table of lectures. After both researcher and lecture are in the tables, conection between them is established by generating a new SQL INSERT query, which takes id of researcher, id of lecture, and the connection between them.

3 Building of export and data indexing module

3.1 Data exporting module

Data exporting module is implemented using web service called – AtlasWebService, which enables exporting data in XML format. This type of design of data exporting module enables remote access to data via Internet protocols (like HTTP) and receive the data formatted as XML documents. This type of design provides an API (application programming interface), which other application can implement and use the data in the standardize way. Access to the web service can be controlled and restricted to particular IP address.

There are five web methods implemented on the AtlasWebService, they provide export of five different types of data. These methods are implemented according to predicted needs of the different visualization and data analysis techniques, which will be performed on the Atlas of Slovenian Science web portal. If the need for other data exports occurs during the development, they will be added additionally. Five web methods are:

- AllRsr
- RsrByName
- CollaborationOfRsrOnPrj
- PrjByName
- PrjOfRsr

3.1.1 Exporting all researchers data

To export all researchers data, web method `AllRsr()` is implemented in the AtlasWebService. This method has no input parameters. As a result it returns XML document containing all researchers from the Atlas of Slovenian Science database.

Firstly, method establishes connection with the database. Next it creates new XML document, which will carry the results. Next step is defining an SQL query which will obtain the data from the database. Results of this query will be obtained from the main table for the researchers – tblResearchers, but also from lookup tables connected to this table: tblScienceCodes, tblFieldCodes and tblSubfieldCodes. This is the SQL query used in this method:

```
SELECT id, mstid, firstName, lastName, status, keyws,

      (select tblScienceCodes.description from tblScienceCodes where
      tblResearchers.science = tblScienceCodes.scienceId) as science,

      (select tblFieldCodes.description from tblFieldCodes where
      tblResearchers.science = tblFieldCodes.scienceId and
      tblFieldCodes.fieldId = field) as field,

      (select tblSubfieldCodes.description from tblSubfieldCodes where
      science = tblSubfieldCodes.scienceId and tblSubfieldCodes.fieldId =
      field and tblSubfieldCodes.subfieldId = subfield) as subfield,

tell, fax, email, url FROM dbo.tblResearchers;
```

The query returns: id, mstid, last name, first name, status, keywords, science, field, subfield, telephone number, fax number, email and webpage of every researcher. Since names of researcher's classification information are in lookup tables: tblScience, tblField and tblSubfield, there was a need nested queries.

This is an example of the XML result of this method:

```
− <Researchers>
  − <RSR mstid="4" id="3955" status="ACT">
      <firstName>Robert</firstName>
      <lastName>Blinc</lastName>
    − <keyws>
        Fizika faznih prehodov, feroelektriki, tekoči kristali, jedrska magnetna resonanca.
      </keyws>
      <science>Naravoslovno-matematične vede</science>
      <field>Fizika</field>
      <subfield>Fizika kondenzirane materije</subfield>
      <tel>(01) 477 32 81</tel>
      <fax>(01) 426 32 69</fax>
      <email>robert.blinc@ijs.si</email>
    </RSR>
  − <RSR mstid="25" id="3956" status="ACT">
      <firstName>Peter</firstName>
      <lastName>Fajfar</lastName>
      <keyws>Potresno inženirstvo, konstrukcije v gradbeništvu</keyws>
      <science>Tehniške vede</science>
      <field>Gradbeništvo</field>
      <subfield>Potresno inženirstvo</subfield>
      <tel>(01) 476 85 92</tel>
      <fax>(01) 425 06 93</fax>
      <email>pfajfar@ikpir.fgg.uni-lj.si</email>
      <url>http://www.ikpir.fgg.uni-lj.si</url>
    </RSR>
</Researchers>
```

3.1.2 Exporting researchers data by name of researchers

It is possible to obtain researcher data by providing researcher name using web method `RsrByName(fname, lname)`. There are two input parameters of the method; these are first name and last name of the researcher. Method returns data about one or more researchers. Input parameters are processed before execution of the query, in order to handle incomplete inputs and Slovenian accented characters. To support retrieving of the right researcher, it is possible to input just few characters of the first and last name of the researcher. Method will return all results which

match the input. This is done by adding '%' sign to the end of input parameters. Slovenian researcher in Atlas of Slovenian Science may have accented characters in their names. To enable international users (which may not use these characters on their keyboard) easy retrieving of data about researcher, every character which could be accented is treated as potentially accented in the query. This is done in the following way (`fname` and `lname` are the input parameters):

```
fname = fname.Replace("c", "[č|ć|c]");
lname = lname.Replace("c","[č|ć|c]");
fname = fname.Replace("z","[ž|z]");
lname = lname.Replace("z", "[ž|z]");
fname = fname.Replace("s", "[š|s]");
lname = lname.Replace("s", "[š|s]");
```

The query in this method is similar to the one for retrieving all researcher data, whit difference of adding WHERE clause in the end of it:

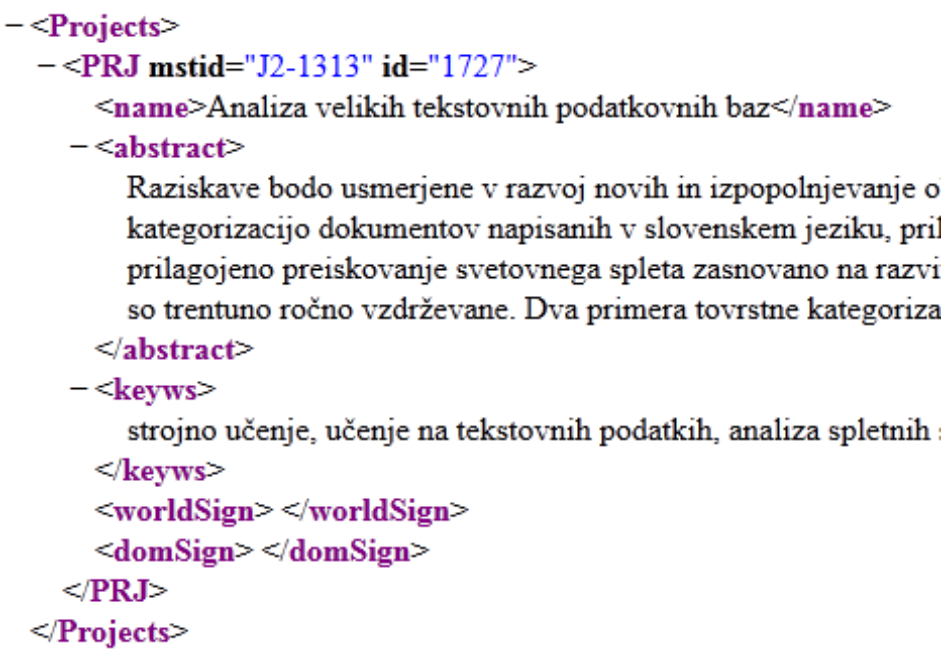
```
SELECT id, mstid, firstName, lastName, status, keyws,

      (select tblScienceCodes.description from tblScienceCodes where
      tblResearchers.science = tblScienceCodes.scienceId) as science,

      (select tblFieldCodes.description from tblFieldCodes where
      tblResearchers.science = tblFieldCodes.scienceId and
      tblFieldCodes.fieldId = field) as field,

      (select tblSubfieldCodes.description from tblSubfieldCodes where
      science = tblSubfieldCodes.scienceId and tblSubfieldCodes.fieldId =
      field and tblSubfieldCodes.subfieldId = subfield) as subfield,

tell, fax, email, url FROM dbo.tblResearchers
WHERE firstName like 'fname' and lastName like 'lname';
```

3.1.3 Exporting project data by project name

Data about projects according to the name of the project, can be exported using web method `PrjByName(name)`. Since query is performed only on one table of the database – tblProjects, it is quite simple:

```
SELECT * FROM dbo.tblProjects WHERE id like 'name';
```

This is the resulting XML, when input of the method is "analiza velikih tekstovnih podatkovnih baz":



We can see from the example that the result carries id, mstid, name, abstract, keywords, world significance and domestic significance of the project.

3.1.4 Exporting all projects of a researcher

All project of a researcher can be exported in XML format using PrjOfRsr(id) method of AtlasWebService. This method takes id of the researcher as an input parameter. After the connection with the Atlas of Slovenian Science database is established, query which retrieves data about all the project of a researcher is defined:

```
select * from tblProjects where id IN( select prjId from tblRsrHasPrj where
rsrId IN( select id from tblResearchers where id = 'id') ) order by name;
```

The data is taken from two main tables: tblProjects and tblResearchers which are connected via table tblRsrHasPrj, therefore, the query needs to be implemented with two nested queries.

The projects of the researchers in the result are given with the same attributes and nodes like in the case of method `PrjByName(name)`, which is shown in previous subchapter.

3.1.5 Exporting collaboration data of a researcher

Using the method `CollaborationOfRsrOnPrj(string id)` of the web service AtlasWebService, data about collaboration of researcher can be obtained. By calling this method, an XML record of all researchers with which selected researcher is/was collaborating is returned. Input parameter of the method is id of researcher whose collaboration is to be retrieved.

This is the query used in the method to retrieve collaboration data from the database:

```
select id, mstid, firstName, lastName, status, keyws, tell, fax, email,
url,
```

```
    (select name from tblOrganizations where id IN(select TOP 1
    tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
    tblRsrIsinOrg.rsrId)) as orgName,
    (select TOP 1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
    tblResearchers.id = tblRsrIsinOrg.rsrId) as orgId,

    (select tblScienceCodes.description from tblScienceCodes where
    tblResearchers.science = tblScienceCodes.scienceId) as science,

    (select tblFieldCodes.description from tblFieldCodes where
    tblResearchers.science = tblFieldCodes.scienceId and
    tblFieldCodes.fieldId = field) as field,

    (select tblSubfieldCodes.description from tblSubfieldCodes where
    science = tblSubfieldCodes.scienceId and tblSubfieldCodes.fieldId =
    field and tblSubfieldCodes.subfieldId = subfield) as subfield,

    N from  tblResearchers,

    (select rsrId, count(*) as N from tblRsrHasPrj where prjId in
          ( SELECT prjId FROM tblRsrHasPrj where rsrId = 'id') group by
    rsrId) as t2

where tblResearchers.id = t2.rsrId order by N desc, CASE rsrId WHEN 'id'
THEN 1 ELSE 100 END, lastName
```

The query is complex because the data is taken from nine tables of the database and counting the number of collaboration is performed. In query are included three main tables (tblResearchers, tblProjects and tblOrganizations), tables for connections (tblRsrHasPrj and tblRsrIsinOrg), and four tables used for classification (tblClassification, tblSubfieldCodes, tblFieldCodes and tblScienceCodes).

Here is shown an example of result of this method:

```
− <Researchers>
  − <RSR mstid="12570" id="7778">
      <firstName>Dunja</firstName>
      <lastName>Mladenić</lastName>
    − <keyws>
        Umetna inteligenca, inteligentni sistemi, strojno učenje, izkopavanje znanja iz podatkov, analiza tekstovnih podatkov
      </keyws>
      <collaborationCount>15</collaborationCount>
      <orgName>Institut 'Jožef Stefan'</orgName>
      <orgId>559</orgId>
      <science>Tehniške vede</science>
      <field>Računalništvo in informatika</field>
      <subfield>Inteligentni sistemi - programska oprema</subfield>
      <tel>(01) 477 32 72</tel>
      <fax>(01) 425 10 38</fax>
      <email>dunja.mladenic@ijs.si</email>
      <url>http://www-ai.ijs.si/DunjaMladenic/home.html</url>
    </RSR>
  − <RSR mstid="17137" id="9594">
      <firstName>Marko</firstName>
      <lastName>Grobelnik</lastName>
      <collaborationCount>12</collaborationCount>
    </RSR>
  − <RSR mstid="8952" id="6592">
      <firstName>Damjan</firstName>
      <lastName>Bojadžiev</lastName>
    − <keyws>
        Računalniška logika, formalno samo-nanašanje, računalniška refleksija, strojno učenje, kognitivna znanost
      </keyws>
      <collaborationCount>5</collaborationCount>
      <orgName>Institut 'Jožef Stefan'</orgName>
      <orgId>559</orgId>
      <science>Tehniške vede</science>
      <field>Računalništvo in informatika</field>
      <subfield>Inteligentni sistemi - programska oprema</subfield>
      <tel>(01) 477 37 68</tel>
      <fax>(01) 425 10 38</fax>
      <email>damjan.bojadziev@ijs.si</email>
      <url>http://nl.ijs.si/~damjan/me.html</url>
    </RSR>
      ⋮
</Researchers>
```

In the above example, input parameter of the method was ‘7778’. This is an identification number of a research whose collaboration is outputted as the result, this researcher can be called selected researcher. Result contains RSR nodes for those researchers which collaborated on project with the selected researchers. First RSR node of each result is selected researcher. Each RSR node contains basic data: first name, last name, keywords, science, field, subfield, telephone number, fax number, email and URL; and data about researcher’s organization: organization name and id. `CollaborationCount` node of each RSR node defines number of projects in common with the selected researcher. For the firs RSR node, collaboration count means number of projects selected researchers was/is on. RSR nodes are sorted in the way that those researchers with greater number of common project are closer to the selected researcher.

3.2 Data indexing module

Index is created for the purpose of fast searching of full text. Index is created using the Lucene.NET search engine library (The Apache Software Foundation). It is intended to be used for purpose of searching researchers according to search terms which can be: name of the researcher, keywords of researcher, titles of project of researcher, keywords of researcher's projects or abstracts of researcher's projects. Method for creating index is integrated into AtlasWebService. It is called with input parameter which defines the name of the index.

To create index, first a query which returns all researchers is performed. For each researcher one document of type Lucene.Net.Documents.Document is created. First name, last name and keywords of each researcher are added as new fields into created documents. Next, with new query, all projects of each researcher are obtained. Titles, keywords, abstracts, domestic and world significance descriptions of all projects are added as five new fields in a document. While adding new fields in a document different options can be applied. These options with description of fields are shown with the following table (Table 1):

Table 1 – Structure of the index

Name of the field	Description of the content	Indexed	Store	Boost
idrsr	Id of a researcher	NO	YES	-
firstName	First name of a researcher	UN_TOKENIZED	YES	8
lastName	Last name of a researcher	UN_TOKENIZED	YES	8
keywsRsr	Keywords associated with the researcher	TOKENIZED	NO	6
projectTitle	Titles of all the projects of the researchers	TOKENIZED	NO	4
abstract	Abstracts of all the projects of the researcher	TOKENIZED	NO	-
keyws	Keywords of all the projects of researcher	TOKENIZED	NO	-
sign_world	Description of world significance of all the projects of the researcher	TOKENIZED	NO	-
sign_dom	Description of domestic significance of all the projects of the researcher	TOKENIZED	NO	-

Field idsrs is not indexed, because id of a researcher will not be used as a search term; but it is stored into the index, because it will be displayed as retrieved result. firstName is indexed, but since it is a single term – it is not tokenized. This field is stored into the index, because it will be displayed as

retrieved result of a search. Also, boost is performed on this field with the value 8, because this field is more important in search than others indexed fields. Field lastName has the same properties as the firstName. Field keywsRsr is indexed and tokenized, because it consists of more terms. It is not stored and it is boosted with value 6, because it is less important than first and last name, but more important than other indexed fields. projectTitle field is indexed with tokenization, not stored in the index for retrieving and boosted with the value 4, because it is less important for the search than firstName, lastName, keywsRsr and projectTitle fields, but more important than other fields. Fields: abstract, keyws, sign_world, sign_dom are all indexed without boosting, tokenized and not stored in the index for retrieving.

4 Conclusion

This document cover two major topics: (1) connecting different data modules and (2) exporting and indexing data. Main focus in the first part is on connecting the data obtained from two different sources: SICIRS and VideoLectures.NET. Step by step description is given of how this data is automatically imported into Atlas of Slovenian Science database. In the second part, implementation, usage and result of developed web service for exporting data is described. Finally, creation of index of the data is described.

5 References

[1] Center for Knowledge Transfer, Jozef Stefan Institute . (brez datuma). *VideoLectures.NET*. Pridobljeno iz http://videolectures.net/

[2] *Classifications*. (brez datuma). Prevzeto 29. 8 2011 iz Slovenian Research Agency: http://www.arrs.gov.si/en/gradivo/sifranti/sif-vpp.asp

[3] IST World Consortium. (brez datuma). *IST-World* . Pridobljeno iz http://www.ist-world.org/

[4] IZUM in ARRS. (brez datuma). Prevzeto 20. 03 2011 iz SICRIS: http://sicris.izum.si/

[5] The Apache Software Foundation. (brez datuma). *Lucene.NET*. Prevzeto 10. 09 2011 iz http://incubator.apache.org/lucene.net/

Atlas slovenske znanosti

RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

R31 – Framework for Data Modeling with Temporal Evolution of Topics

Ljubljana, 20.3.2012

Table of content

1 Introduction

Data on researchers and their activity as captured in publicly available portals, such as SICRIS, videolectures.NET or IST-World, record some textual and/or multimedia description of the researcher's activity at a certain point in time. This enables modeling of data via observing temporal evolution of topics for a specific researcher or a group of researchers, as well as analysis of science development based on the research projects and publications. Connecting different data sources potentially makes the analysis more demanding but has greater potential for more interesting results.

Since the year 2000 we are witnessing intensive development of empirical methods for analysis of different data types including text, social networks and multimedia content. This report is mainly based on data mining, social network analysis, text mining, semantic web, complex data visualization and analysis of complex dynamic systems as addressed in complexity science. Our goal is to include modern data analysis methods form these areas for empirical analysis of science development supporting real-time, multidimensional and sufficiently detailed insights into scientific activities. The science can then be analysed from three aspects:

- social - reflected in connection between individual researchers, organizations and projects,
- content – incorporating topics of projects and publications,
- temporal – analysis of science over time including trends and prediction of future development.

This report consist of five parts addressing modeling of data using social network analysis methods and using text analysis methods, modeling the evolution of data using social network analysis and using text analysis and, a combination of social network and text analysis methods.

2 Modeling the Data with Social Network Analysis

In this chapter social network analysis on the data about researchers and their collaboration on projects are performed. The chapter is divided into three parts. The first part holds the description of the data on which the analysis were made, the second part gives the properties of the network connected to social cohesion and the third part gives the properties of network related to brokerage.

2.1 Description of the Data

Data about researchers and projects was taken from Slovenian Current Information System (SICRIS) database in March 2011, and it includes both active and not active researchers, and projects from 1994 to 2010. Using this data a network was created. Each vertex of the network represents a researcher, while an edge between two vertices represents collaboration on a project between the two researchers. Two researchers can collaborate on more than one project, so the edges can have weights of different values. When visualizing the network we use undirected edges (in opposition to arcs), because the collaboration between two researcher is commutative, i.e. the line from the

researcher "*a*" to the researcher "*b*" is equivalent to the line from the research "*b*" to the researcher "*a*". The network does not contain loops (i.e. edges between researcher "*a*" to "*a*"), because the collaboration of a researcher with himself does not make sense in the proposed context.

The data for analysis using social networks was obtained from the Atlas of Slovenian Science database exported into a suitable format for usage by Pajek (Batagelj & Mrvar, 2012) – software with which the social network analysis is performed. Pajek is a program for analysis and visualizations of large networks. To obtain the network of collaboration between researchers on project, it is needed to get all researchers working on the same projects for each researcher. The query which can construct this is complex, because it has to be performed for each project of a researcher and for every researcher. Since there are 33474 researchers in the last update of the database, the query for constructing the network is computationally demanding. Finally constructed network file consisted of 33474 vertices and 143743 edges. The format of the network file is such, that first all the vertices are listed with id, label and coordinates for x, y and z axis on the 3-dimensional plane. Following the list of vertices is the list of all edges in the vertex pair format together with the vertex weight. This is the example containing few example lines from the researchers network file:

```
*Vertices 33474
...
3820 "12570_Dunja_Mladenic"                    0.6678    0.0424    0.5000
...
5634 "17137_Marko_Grobelnik"                   0.8770    0.2064    0.5000
...
12557 "13325_Mitja_Jermol"                     0.8923    0.5878    0.5000
...
*Edges
...
3820 5634 12
...
3820 12557 4
...
5634 12557 4
...
```

The first line of the example above is indicator of vertices list with the number of vertices. Next are lines are vertices, which contain id, label in the double quotes and coordinates. The id of the vertex cannot be bigger than the number of vertices in the list. This is a constraint which makes addition difficulties with construction of the network file, because MSTID values of the researcher could not be used as ids in the network file. The label of the vertex consists of MSTID of researcher, first and last name of the researcher, connected with dashes. The coordinates are normalized to 1 and are generated randomly during the file construction. The line `*Edges` indicates that following is list of edges. Edge is represented as a pair of vertices ids with the weight of the edge. In the example are 3 edges, i.e. first edge connects first (12570_Dunja_Mladenic) and second (17137_Marko_Grobelnik) vertex from the example with weight 12.

2.2 Cohesion

Cohesion is part of social network analysis investigating who is related to whom and who is not related. Density and degree are the two main measures for cohesion. Density is the number of lines in a simple network, expressed as a proportion of the maximum possible number of lines. More ties between people yield a tighter structure, which is, presumably, more cohesive. The degree of a vertex is the number of lines incident with it. Vertices with high degree are more likely to be found in dense sections of the network. Density is not the best measure for cohesion of a network, because it depends on the size of the network. Density is inversely related to network size: the larger the network, the lower the density, because the number of possible lines increases rapidly with the number of vertices, whereas the number of ties which each person can maintain is limited. Thus In addition to density, one can use average degree of all vertices as a measure of overall cohesion, as it does not depend on the network size (so average degree can be compared to networks of different sizes). Dense groups of actors in a network which interact intensively are called cohesive subgroups. (Wouter de Nooy, 2005)

In the rest of this Section first the density and degree of entire network of researchers and sub networks for each science group is given. Next, the collaboration between science groups is explained. Finally, the cohesive subgroups of researchers in the network are identified.

2.2.1 Density and Degree of Researchers network in Slovenia

In this part are given the results of measuring the cohesion of the networks of researchers in Slovenia, using density and degree measures. First the density and degree of entire network are examined, followed by the density and degree of sub networks for each science group.

2.2.1.1 Density and Degree of the Network

The dataset representing the complete network of researchers in Slovenia, on which the measurements were conducted, consisted of 33474 vertices and 143743 edges. Since the density is the proportion of the number of lines in a simple network and the maximum possible number of lines, it is necessary to calculate the maximum possible number of lines. The maximum possible number of lines (e_{max}) of a simple network (no loops and multiple lines), with the given number of vertices (v), is calculated with the following equation:

$$e_{max} = \frac{v^2 - v}{2}$$

In our case this gives 560.237.601 maximum possible lines, which with 143.743 of existing lines, gives the density of 0.00025658. This means that only 0.025658% of all possible lines are present in the network.

The average degree of all vertices is a better measure of cohesion because it does not depend on the size of the network. The average degree of the network is 8.5883, which means that each researcher in average collaborates with 8 different researchers.

Since edges represent the collaboration on projects, the weights of the edges give information about the number of different projects two researchers have collaborated on. Figure 1 graphically shows the weights of all 143743 edges. The long tail of the graph in Figure 1 indicates the large number of edges with weight 1, which means that in most cases, researchers collaborated on only one project with the same researcher. The highest number of collaborations for two researchers is 22, but there are only a small number of such edges, as indicated by the peak on the left side of Figure 1.

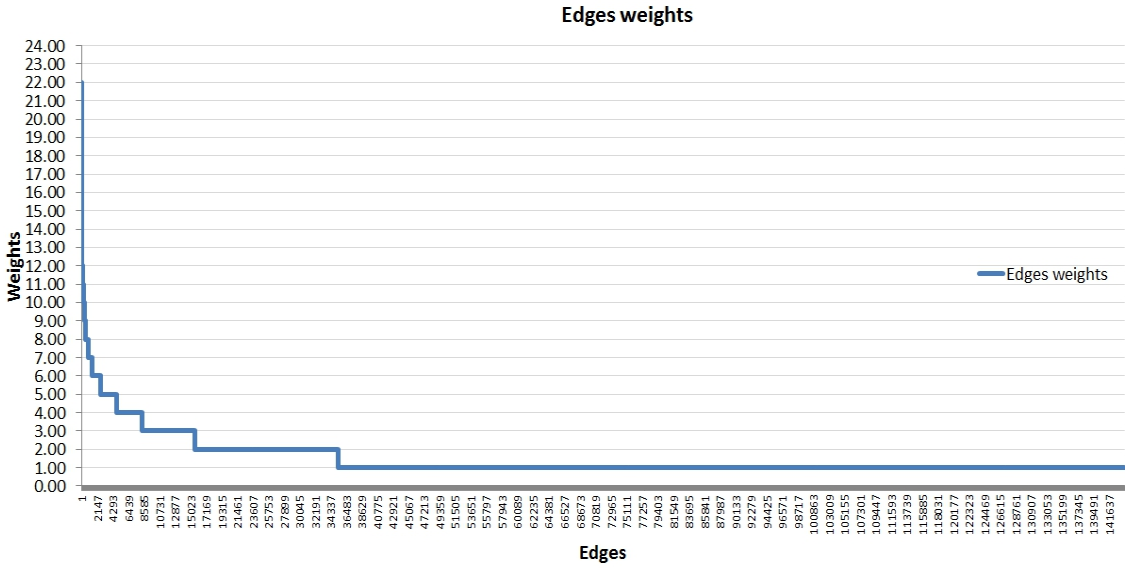


Figure 1 - Weights of edges of the Researchers Network

2.2.1.2 *Density and Degree of the science groups*

In the SICRIS dataset, researchers are classified according to Field of Research Classification. The classification is hierarchical with three hierarchical levels: science, field and subfield. The cohesion measures – density and average vertex degree, were calculated for sciences, which are on the highest hierarchal level of this classification. These are 7 sciences: (1) Natural science and mathematics, (2) Engineering sciences and technologies, (3) Medical sciences, (4) Biotechnical sciences, (5) Social sciences, (6) Humanities and (7) Interdisciplinary studies. Engineering sciences and technologies is the largest science with 8930 researchers, while the smallest one - Interdisciplinary studies has only 83 researchers belonging to it. Even though the sizes of sciences strongly differ, the average vertex degree should be a good measure of cohesion, showing which sciences are the most and which are the least internally collaborative. The average vertex degree, density, number of vertices and number of edges for each science are shown in Table 1. Figure 1 shows graphically average vertex degree of groups of sciences.

Table 1 - Density and Degree of the Sciences

Name of the science group	Number of vertices	Number of edges	Density	Average Vertex Degree
Natural sciences and mathematics	3765	14047	0.0020	7.4619
Engineering sciences and technologies	8930	23877	0.0006	5.3476
Medical sciences	2639	13672	0.0039	10.3615
Biotechnical sciences	1602	13490	0.0105	16.8414
Social sciences	3243	10563	0.0020	6.5143
Humanities	1801	7069	0.0044	7.8501
Interdisciplinary studies	83	4	0.0012	0.0963

We can see that biotechnical sciences have the highest cohesion with the average vertex degree 16.8414. The second most cohesive group of sciences is medical sciences with the average vertex degree 10.3615. The groups of sciences: Natural sciences and mathematics, Engineering sciences and technologies, Social sciences and Humanities, all have similar cohesion with average vertex degree between 5.3476 and 7.8501. Interdisciplinary studies group of sciences has extremely low cohesion compared with the other groups. The size of this group is much smaller as well. There are 83 researchers in this group, with only 4 distinct collaborations and the average degree 0.0963 .

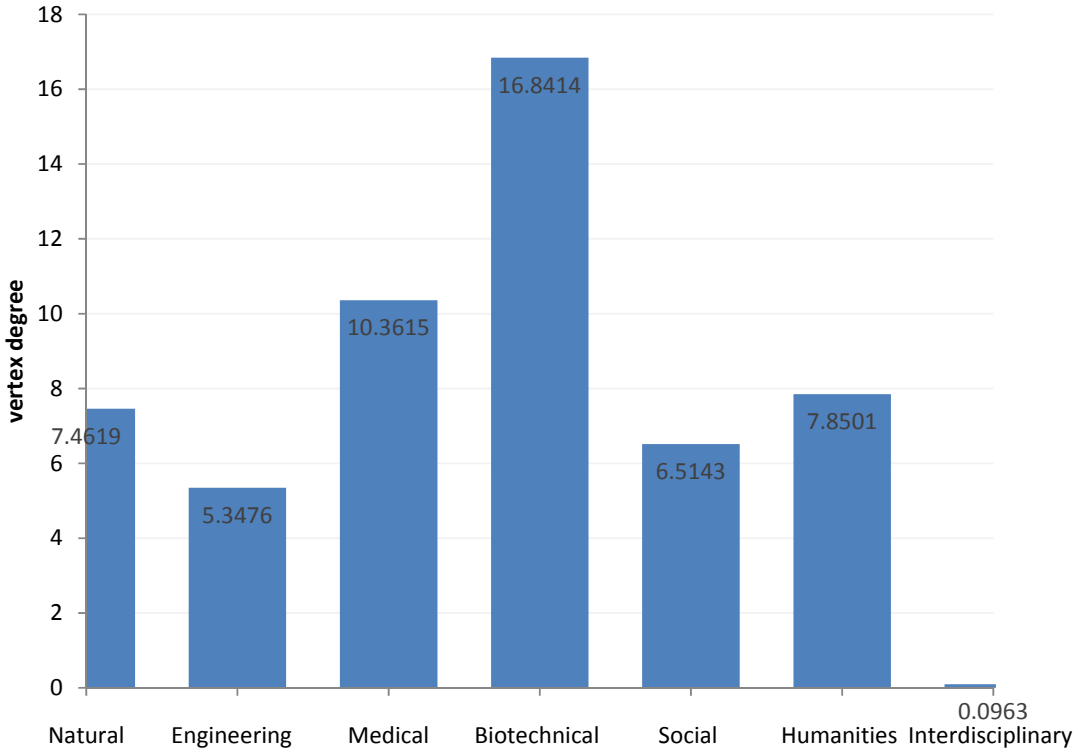


Figure 2 - Average vertex degree of science groups

2.2.2 Collaboration between science groups

Science groups can be observed from the global view by merging all the researchers belonging to that group and representing each group as a single vertex. Figure 2 shows the collaboration between the groups, where the connection between vertices represents collaboration between all researchers from one group to the researchers from other group. To improve the clearness of the graph, all lines with weight less than 1000 were discharged from the graph. The information about the connections between all the vertices is shown in Table 2. We can see that the highest collaboration is between Biotechnical sciences and, Natural sciences and mathematics (9245 collaborations).

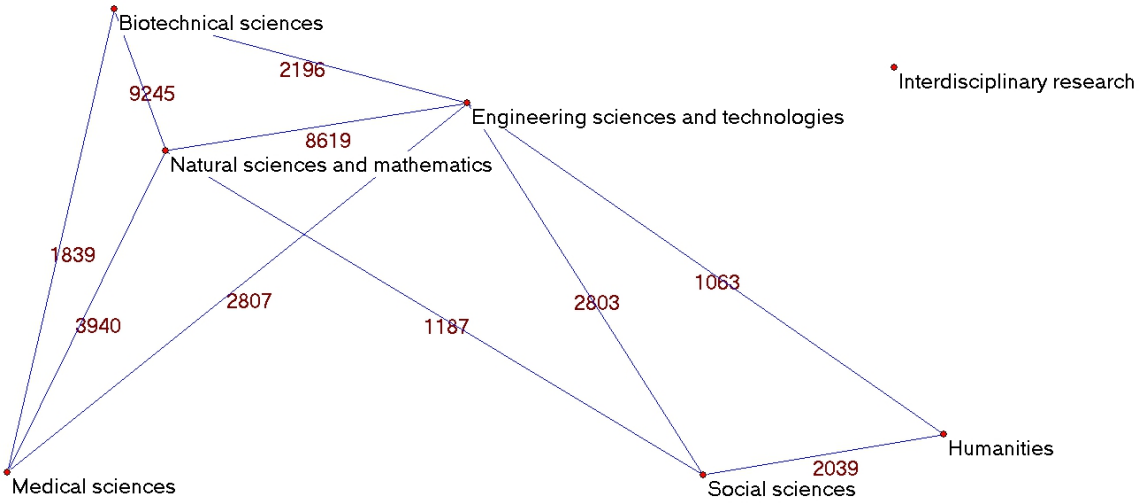


Figure 3 - Collaboration between science groups

In Figure 3, we can see that Engineering sciences and technologies has the highest degree, it connects with all other groups of sciences, with connections stronger than 1000 collaborations. The strongest connection (8619 collaborations) is between Engineering sciences and technologies and the Natural sciences and technologies. Natural sciences and technologies is the group that is the second in the number of connections to other groups of sciences. The group that is the least connected to sciences from other groups is Humanities, which connects only to the Engineering sciences and technologies and to the Social sciences.

Table 2- Collaboration between science groups

No	Science group 1	Science group 2	weight
1	Natural sciences and mathematics	Biotechnical sciences	9245
2	Natural sciences and mathematics	Engineering sciences and technologies	8619
3	Natural sciences and mathematics	Medical sciences	3940
4	Engineering sciences and technologies	Medical sciences	2807
5	Engineering sciences and technologies	Social sciences	2803
6	Engineering sciences and technologies	Biotechnical sciences	2196
7	Humanities	Social sciences	2039
8	Biotechnical sciences	Medical sciences	1839
9	Natural sciences and mathematics	Social sciences	1187
10	Engineering sciences and technologies	Humanities	1063
11	Natural sciences and mathematics	Humanities	963
12	Medical sciences	Social sciences	955

13	Biotechnical sciences	Social sciences	559
14	Engineering sciences and technologies	Interdisciplinary research	191
15	Biotechnical sciences	Humanities	188
16	Natural sciences and mathematics	Interdisciplinary research	133
17	Social sciences	Interdisciplinary research	60
18	Humanities	Medical sciences	56
19	Humanities	Interdisciplinary research	27
20	Biotechnical sciences	Interdisciplinary research	18
21	Medical sciences	Interdisciplinary research	7

2.2.3 Cohesive subgroups of Researchers Network in Slovenia

The complete network consists of 33474 researchers, 23214 (69.35%) of which are isolated, i.e. do not have collaboration on projects with any other researcher. The rest of 10260 researchers which have some collaboration, are mostly part of one big connected component consisting of 10183 researchers. Taking into account just the 10260 vertices which have some connection, the big component includes 99.25% of vertices. Except this component, there are 16 other small connected components with 3 to 10 vertices. In order to identify cohesive subgroups of similar sizes the technique of removing low weighted edges was applied. When all the edges with the weight smaller than 2 were removed, 4036 of previous researchers became isolated. After removing the isolated researchers, remains one dominant connected component with 5674 (91.16%) vertices and 81 small ones neither containing more than 1% of vertices. Removing all lines with weight lower than 3 results in 2161 newly isolated connected component, one large component containing 60% of vertices and 168 small components of which the biggest contains 1.8% of vertices. It can be noticed that each removal of edges with weights under some threshold detaches a big number of vertices from the large connected component and makes them isolated, and creates additional number of new connected components of small size. This indicates that the network does not contain many bridges and cut-vertices, which is shown in more detail in the next section. We can see that the researchers which collaborate on projects are not connected in big cohesive subgroups, but area part of one big group that involves the whole network of researchers in Slovenia.

To obtain the connected components from the core of the network, all the edges with the weight lower than 15 were removed. This is an extreme reduction since the network mostly consists of lower weight edges (covered in last part of chapter 3.1.1 – Density and Degree of the Network). Next, the connected components consisting of at least 3 vertices were identified. The result is 6 connected components showed in Figure 4. These components are located in the center of the network, because the edges between the connected vertices are very strong. The biggest of the components consists of 6 vertices (upper-left part of Figure 4); two researchers from this component belong to Medical sciences, while there was no information about Science classification for the rest of the researchers in the component. The second biggest component (upper-right part of Figure 4) has 4

researchers, all from Biotechnical Sciences. There is one component with three researchers from Engineering sciences and technologies (lower-right part of Figure 4), while all other researchers in components are from Biotechnical sciences (including one from unknown science group). It can be noticed that the majority of researchers of these components connected with very strong edges, comes from biotechnical sciences. This is not surprising taking into account that biotechnical sciences had the strongest cohesion, measured by the average vertex degree (see Section 3.1.2 - Density and Degree of the science groups).

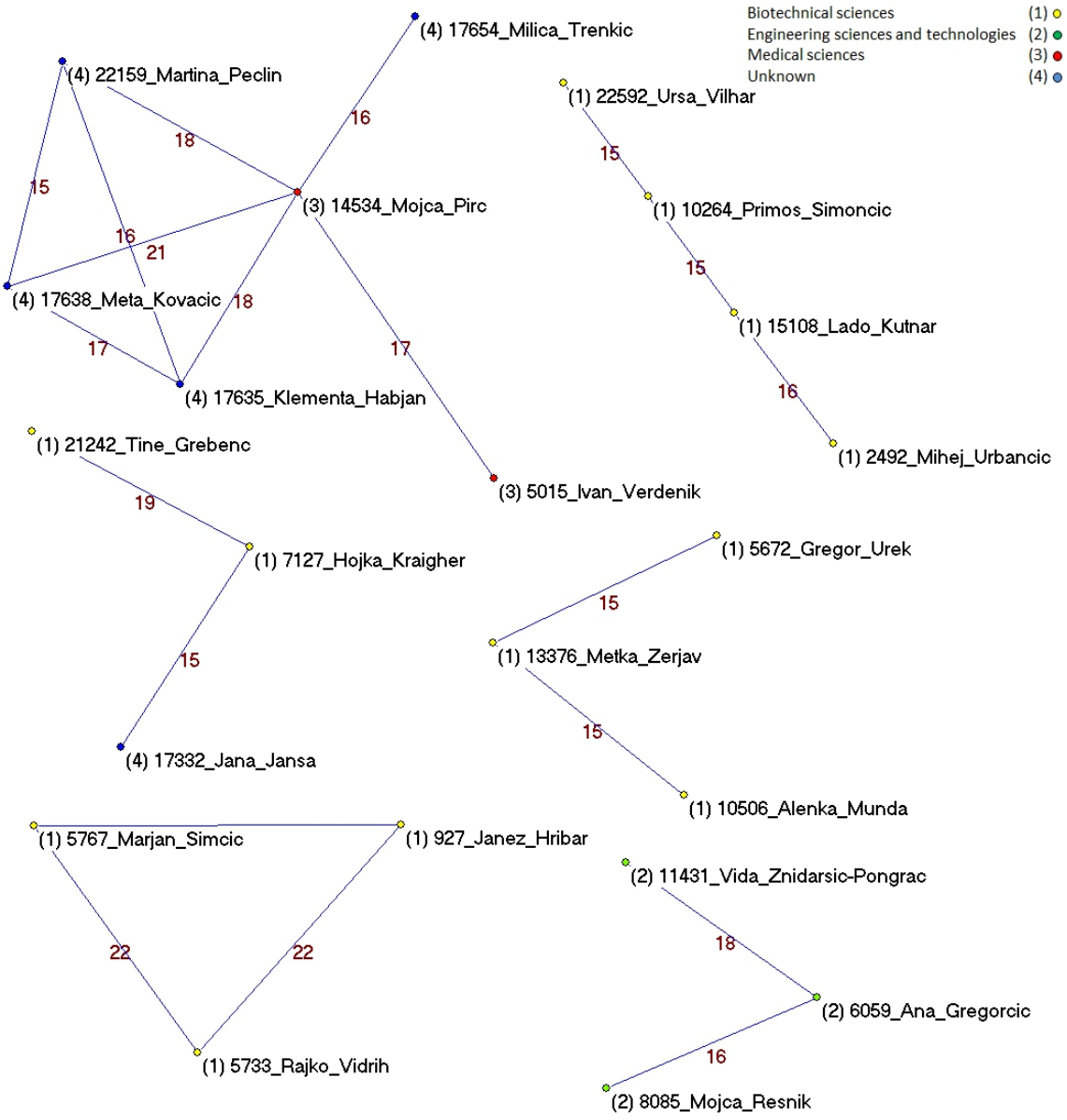


Figure 4 - Connected components of researcher network with at least 3 vertices and edges weight 15 or more

2.3 Brokerage

Brokerage deals with social networks as structures that allow for the exchange of information. Main parts of brokerage are centrality, brokers and bridges, and diffusion. Centrality refers to positions of individual vertices within the network, whereas we use centralization to characterize an entire network. People or organizations that are central have better access to information and better opportunities to spread information. Highly centralized network spreads information easily but the center is indispensable for the transmission of information. Brokers and bridges are respectively: elements of the network that are important for identifying bottlenecks of the network and the individuals which are in the best position to profit from their social ties. Brokerage analyses ties and vertices (called bridges and cut-vertices) in the network that are indispensable for network to stay connected. Diffusion is a special case of brokerage with a time dimension that analysis how structure of ties influences spreading of something (opinions, products, diseases, attitudes) through time. (Wouter de Nooy, 2005)

In the rest of this Section we first examine how central are the researchers (centrality) and how central is the entire network of researchers (centralization), and sub networks of researchers belonging to the science groups. In the second part of this Section bi-components of the network are identified, with the finding that network behaves like a strongly connected component, without bridges that could disconnect the component into other components with some significant size.

2.3.1 Centralization of Researchers network in Slovenia

Central people have quick access to information circulating in the network. Centrality is the term used to express how central is a person in some network, while the centralization applies to the measure expressing how central is the network as a whole. The centrality and centralization can be expressed by three different measures: degree centrality/centralization, closeness centrality/centralization and betweenness centrality/centralization.

Degree centrality and centralization is based on the idea that information may easily reach people who are central in a communication network. In case of network of researchers, the reachability of information can be considered as reachability of expertise from other researchers. The simplest indicator of centrality is the number of its neighbors – vertex degree. The degree centrality of the vertex is its degree. The network is more centralized if the vertices very more with respect to their centrality. Degree centralization of a network is the variation in the degrees of vertices divided by the maximum degree variation which is possible in a network of the same size.

Closeness centrality is based on the total distance between one vertex an all other, where larger distances yield lower centrality scores. It is defined as the number of other vertices divided by the sum of all distances between the vertex and all others. Closeness centralization is the variation in the closeness centrality of vertices divided by the maximum variation in closeness centrality scores possible in a network of the same size.

Betweenness centrality is not based on the reachability of a person within a network, but on the importance of the person in the network for the flow of information. The betweenness centrality of a

vertex is the proportion of all shortest paths between pairs of other vertices that include this vertex. Betweenness centralization is the variation in the betweenness centrality of the vertices divided by the maximum variation in betweenness centrality scores possible in a network of the same size.

2.3.1.1 Centralization of the of the Network

Network Degree Centralization measured on our data is 0.0068. The degree centralization of the largest connected component (10183) is 0.0204. Table 3 shows the top 10 ranked researchers according to the network degree centrality measure. The arithmetic mean of degree centrality is 0.0003 with standard deviation of 0.0006.

Table 3 - Top 10 degree centrality scores

Rank	Degree centrality	Researcher
1	0.0071	2085_Franc_Batic
2	0.0064	10264_Primos_Simoncic
3	0.0064	7127_Hojka_Kraigher
4	0.0063	8800_Gregor_Sersa
5	0.0062	14082_Radojko_Jacimovic
6	0.0061	11595_Tomislav_Levanic
7	0.0061	16283_Borut_Vrscaj
8	0.0060	10807_Sonja_Lojen
9	0.0060	1644_Milan_Pogacnik
10	0.0059	5098_Peter_Dovc

Figure 5 graphically shows the degree centrality scores distribution of the network vertices. It can be seen that most of the researchers have degree centrality score 0, less than 1/3 of the researchers has degree centrality score between 0.0001 and 0.003 and only a small percentage of researcher (less than 1%) has the degree centrality score above 0.003.

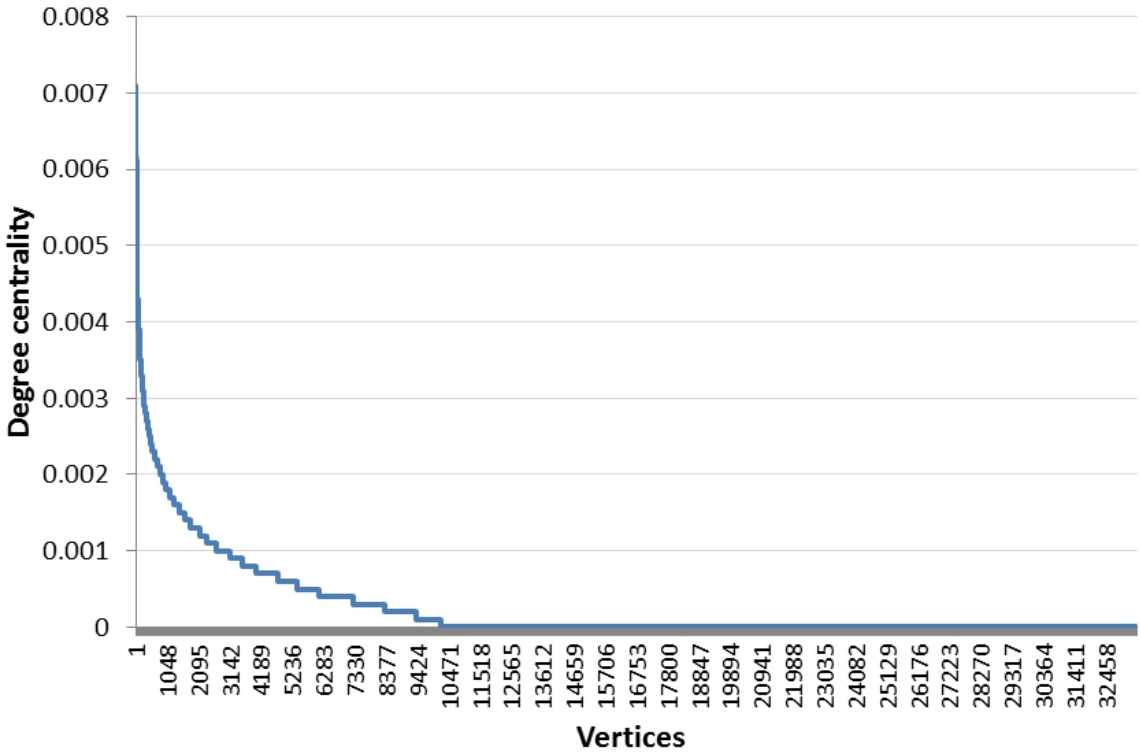


Figure 5- Degree centrality scores

Closeness centrality of the networks largest connected component is 0.1669. Betweenness centrality of the largest connected component is 0.0230.

2.3.1.2 Centralization of the science groups

Table 4 shows the centralization of each group of sciences. The first column shows degree centralization for the complete network. The degree centralization, closeness centralization and betweenness centralization in the second, third and fourth column of the table are calculated for the largest connected component of each science group. The centralization measures are performed on connected component rather than on the whole sub network, because the condition for calculating closeness centralization is interconnection of all vertices.

Table 4 - Centralization of the science groups

Name of the science group	Network Degree Centralization	Component Degree Centralization	Component Closeness Centralization	Component Betweenness Centralization
Natural sciences and mathematics	0.0320	0.0690	0.2068	0.0769
Engineering sciences and tech.	0.0113	0.0343	0.2070	0.0449
Medical sciences	0.0378	0.0645	0.2504	0.0873

Biotechnical sciences	0.0801	0.1315	0.2490	0.0538
Social sciences	0.0279	0.0629	0.1918	0.0637
Humanities	0.0340	0.0651	0.1976	0.1224
Interdisciplinary studies	0	0	0	0

We can see that biotechnical sciences have the biggest degree centralization, measured both on the whole science sub networks and on the largest components of the science sub networks. This is the confirmation of findings described in the chapter 3.2.1 - Density and Degree of the science groups, where it was showed that biotechnical sciences have the biggest average vertex degree. The closeness centralization biotechnical sciences are not the biggest, medical sciences are slightly more centralized according to this measure. The betweenness centralization is the biggest for humanities, meaning that it is the most centralized group of sciences taking into account connections which are indispensable for information flow. The variations in the centralization rank of the science depending on centralization measure indicate that it is important to understand which centralization measure is used and what does it means. Biotechnical sciences are highly centralized according to the centralization measures based on the reachability of the person, while the centralization is not so strong from the information flow point of view. On the other hand Humanities are showing opposite centralization characteristics, with high betweennees centralization and lower reachability based measures – degree centralization and closeness centralization. Engineering sciences and technologies is the group of sciences which does not have big centralization score according to any of the measures.

2.3.1.3 Bridges and Bi-Components of Researchers Network in Slovenia

A bridge is a line whose removal increases the number of components in the network. Cut-vertex or articulation point is a vertex, deletion of which disconnects the network or a component of the network. A bi-component is a component of minimum size 3 that does not contain a cut-vertex.

Slovenian researcher network has 53 bi-components. 23274 researchers (69.53%) do not belong to any bi-component of minimum size 3. One bi-component contains 9988 researchers, what makes 30.37% of all the researchers from the network, or 97.59% of researcher taking into account just the ones belonging to some bi-component. From other bi-components which make the minor part of the network, the most of them are of size 3. Number and sizes of bi-components are shown in Table 5.

Table 5 - Bi-components

Bi-component size	**Number of components**	**Percentage of vertices**
9988	1	97.59%
27	1	0.26%

13	1	0.13%
11	1	0.11%
10	0	0.00%
9	2	0.18%
8	2	0.16%
7	1	0.07%
6	2	0.12%
5	3	0.15%
4	11	0.43%
3	28	0.82%

Bi-components represent components which have very good connectivity since there is no articulation point which could make the network disconnected. The percentage of vertices belonging to a single dominant bi-component indicates that the network of researcher's collaboration has a specific structure not usual for a typical social network.

2.4 Conclusion

We have applied social network analysis methods on the data capturing collaboration on national projects between Slovenian researchers. We have measured cohesion of the researcher network. First the density and degree of the entire network and the sub networks of researchers belonging to science groups are examined. The entire network is sparse with density measure of 0.025658%, while the average vertex degree of entire network is 8.5, meaning that the average researcher collaborates with 8 other researchers on projects. Measuring the cohesion of the sub networks according to the science group, biotechnical sciences turned out to be the most cohesive sub network with density of 0.0105 and average vertex degree of 18.4. Examining the collaboration between science groups, it is found that Engineering sciences and technologies and Natural sciences mathematics and have the most central position, being connected with strong edges (with weight greater than 1000) to 5 and 4 other science groups respectively. In the last part of this chapter, some small cohesive subgroups from the core of the network were identified.

Last part examines social networks as structures that allow for the exchange of information - brokerage. In the first part centrality and centralization measures were applied to entire network and to sub networks according to science groups. The network degree centralization is 0.0068, closeness centrality of the networks largest connected component is 0.1669 and the betweenness centrality of the largest connected component is 0.0230. The list of top 10 researchers with the highest centrality scores was given, together with the distribution of centrality scores across the whole network. The comparative results of all tree centralization scores were given for all six groups of sciences for the researcher network. In the final part of the chapter bi components were identified, with the conclusion that the network has a special structure with one dominant bi-component.

3 Modeling the Data with Text Analysis

This chapter has three main parts, each part covering a different aspect of text analysis and uses different software tool for it. The first part is about discovering text properties using R software. The second part tackles unsupervised learning on the text for creating topic ontology using OntoGen software. The third part covers supervised learning for classifying researchers into categories on the basis of their projects, using the Text Garden software. The data used in all three parts includes researchers and projects from Slovenian Current Information System (SICRIS) database, obtained in March 2011 (see Section 2.1 for details on the dataset).

3.1 Text mining with R

In the first part of this chapter, various operations which show the properties of text using on R statistical software are described. The operations are described in the step-by-step manor, so they could be easily reconstructed. R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity. (Gentleman, Robert; Ihaka, Ross; et al)

First, the loading and preprocessing of the data is described. This includes loading the data in R and text preprocessing operations like: white space removal, numbers removal, converting to lowercase, stop words removal and stemming. The second is focused on analyzing the text in R. This covers discovering the properties of text corpus (creating and exporting of textual corpus, creating document-term matrices with different weightings, identifying most frequent terms and density of the matrices, discovering correlations between terms and bigrams), reductions of document-term matrices, empirical laws (Zipf's and Heap's) on the corpus and creating the word cloud in R.

3.1.1 Loading and preprocessing

Loading and preprocessing is an unavoidable step in every process. This chapter describes the retrieving of data from database into a file used for loading into R and different text preprocessing operations which include: white space removal, numbers removal, converting to lowercase, stop words removal and stemming.

3.1.1.1 Preparation of the data

The data, on which the text mining analysis is performed, is located in the Atlas of Slovenian Science database. The subset of data in the focus of the analysis is the textual descriptions of the research projects. The database is implemented using the Microsoft SQL Server platform and the data is accessed by the SQL Server Management Studio. The data which we want to obtain is located in the tblProjects table. The columns of the table that carry the textual descriptions of the projects are: name_en (names of the projects in English language), keywords_en (keywords of the projects in

english language), abstract_en (abstracts of the projects in English language), abstract_en (abstracts of the projects in English language), sign_dom_en and sign_world_en (description of domestic and world significance of the project in English lanuage). The following query was performed in the SQL Server Management Studio with the option `Query>Save To File` selected:

```
SELECT [name_en] ,[keyws_en], [abstract_en], [sign_dom_en], [sign_world],
[sign_world_en] FROM [AtlasTest1].[dbo].[tblProjects]
```

and the results of the query were saved into file `projects.rpt`. The file is opened with Notepad++ text editor, where the column headings are removed and the file is saved in UTF-8 encoding as projects.txt.

3.1.1.2 Data loading and preprocessing operations in R

Before doing actual work with the data, proper package has to be installed and loaded in the R. The main package used for our data mining tasks is called *tm* (Feinerer, tm: Text Mining Package. R package version 0.5-7.1., 2012). Nice introduction and a detailed description of the text mining infrastructure in R can be found in (Feinerer, Introduction to the tm Package Text Mining in R, 2011) and (Feinerer, Hornik, & Meyer, Text Mining Infrastructure in R, 2008) respectively.

The following command is used to get a list of all available packages for R:

```
utils:::menuInstallPkgs()
```

This gives a list of all available packages, from which the tm should be selected. In case of installing the packages for the first time, the list of the available mirror server from which the download should be performed is also displayed. After the tm package is installed, it should be loaded with the following command:

```
library(tm)
```

Next, the projects descriptions saved in textual file are loaded into an R object. One useful thing to do, is to check and set the working directory of R. Working directory is checked with the command:

```
getwd()
```

The new working directory can be set with the command:

```
setwd(newPathToDirectory)
```

When we know the working directory, we can place the file projects.txt into it and perform the following command to load the projects textual data into an R object:

```
x <- readLines("projects.txt", encoding = "UTF-8")
```

This command creates vector x with each project as one of its element. Typing x into R's command line interface prints all of the project description. First 10 elements of the vector can be displayed in the following way:

```
x[1:10]
```

Next, all the words in the vector x are transformed into lower case words and stored into vector lCx with the command:

```
lCx = tolower(x)
```

It can be noticed that some names of the projects are written in Slovene rather than in English language, even thou all the data was retrieved from database columns which should contain the English text. Nevertheless, many names of the projects were written in Slovene with added note: (Slovene), which warns that this text is in the different language than it was meant to be. Since this text mining tasks are oriented to analysis of English text, the projects with Slovene descriptions must be filtered out, which is done with the command:

```
lCxEn <- grep(pattern="(slovene)",lCx, invert=TRUE, value=TRUE)
```

This creates a vector lCxEn with all documents containing (Slovene) mark discharged. Filtering is performed using the grep function. Pattern is the peace of text which should be identified in the document. Since all the words were transformed into lower case, the pattern is (slovene). lCx is vector on which the filtering is performed. Parameter invert is set to TRUE which means that the documents which do not match the pattern will be written into new vector. Parameter value is also set to TRUE, which means that the whole content of the documents, not just the number of the document will be saved into the new vector lCxEn. The lCx vector contained 5384 projects, while the lCxEn contains only 2054 projects. This means that 62% of the English descriptions of the projects are missing from the dataset, or in other words, only 38% of the projects contain textual description in English language.

From the vector containing textual descriptions of projects in English language, new corpus is created with the following command:

```
prjCorpusEn <- Corpus(VectorSource(lCxEn))
```

The documents of the corpus can be observed with the inspect command. I.e. first 10 document can be view with this command:

```
inspect(prjCorpusEn[1:10])
```

The corpus can be processed with different transformation. The following code shows the commands for removing the punctuation signs, removing the numbers and removing the redundant whitespaces between words respectively:

```
prjCorpusEn <- tm_map(prjCorpusEn, removePunctuation)

prjCorpusEn <- tm_map(prjCorpusEn, removeNumbers)

prjCorpusEn <- tm_map(prjCorpusEn, stripWhitespace)
```

Next, the stop-words can be removed from the corpus. Stop-words are words commonly used to form the structure of the sentence (e.g. “the”,” a”, “of”, “and”, etc.), for which is showed that they do not contribute much to information about content of the text (Joachims, 1998). The predefined list of stop-words from the English language can be removed with the following command:

```
prjCorpusEn<-tm_map(prjCorpusEn,removeWords,stopwords("english"))
```

The additional words can be removed by creating a vector of words which should be removed and applying it in the removeWords function. Word null is appearing as a symbol representing empty fields in the database. It can be removed in the following way:

```
wordsToRemove <- c("null")

prjCorpusEn<-tm_map(prjCorpusEn,removeWords,wordsToRemove)
```

Performing the stemming of the corpus:

```
prjCorpusEnStem <- tm_map(prjCorpusEn,stemDocument)
```

Finally, the corpus can be saved in a directory so it can be easily accessed for later use. The corpus is written in a directory in the way that every document of a corups is written to one textual file with txt extension. This is the command for writing the corpus in the directory prjCorpusEn, which is a child directory or R's working directory.

```
writeCorpus(prjCorpusEnStem,path="./prjCorpusEnStem")
```

3.1.2 Analyzing the textual corpus

This part describes the process of analyzing the textual data using R. Firstly it is shown how the properties of textual corpus can be discovered. These properties are: most frequent terms, document-term matrix density, terms co-occurrence and bigrams co-occurrence. Next, it is described how the document-term matrix can be reduced by removing sparse elements. Following is the description of inspecting two empirical laws – Zipf's and Heap's law, on the text corpus. Finally, the process of building the word cloud is shown.

3.1.2.1 Corpus properties

At first a term-document matrix for the corpus is created. Rows of the term-document matrix correspond to document IDs and the columns correspond to terms. Element of the matrix contains the frequency of a specific term in a document. Term-document matrix is created with the following command:

```
dtmStem <- DocumentTermMatrix(prjCorpusEnStem)
```

The frequencies in the term-document matrix are expressed with term frequency (TF), but other measures can be easily applied, e.g. term frequency-inverse document frequency:

```
weightTfIdf(dtm, normalize = TRUE)
```

Typing the name of the term-document matrix displays some main properties:

```
dtmStem
```

The output of this command is:

```
A document-term matrix (2054 documents, 14192 terms)
Non-/sparse entries: 99658/29050710
Sparsity           : 100%
Maximal term length: 61
Weighting          : term frequency (tf)
```

The first row of the output gives information about how many documents and terms are in the matrix. The number of document is equal to number of rows of the matrix and the number of terms is equal to the number of columns. Since there are 2054 rows and 14192 columns, the matrix has total of 29150368 elements (2054*14192). Next row tells us how many non-spars and spars entries are in the matrix. Spars entry is value 0 in the particular place in the matrix, which means that for the document defined by the row, there is no term defined by the column. The document-term matrix of the research projects consists of 99658 non-sparse entries and 29050710 spars entries (sum of these to values gives the total number of entries). Information about spars and non-sparse entries is used to calculate the sparsity of the matrix, which is given with the next row of the output of previous command. The sparsity of the matrix is the proportion of the sparse entries from the total number of entries in the matrix. Since this proportion is very close to 1 (29050710/29150368 = 0.9965812), R rounds the sparsity of the matrix to 100%. Next row of the output gives information about term with biggest number of characters. Finally, the last row gives the information about weighting used in the matrix, which is term frequency (tf) in this case.

The matrix can be displayed with the inspect function, but since it is very big and the output on the command line display would not be readably, a small portion of the matrix can be displayed. The following command displays the matrix from the row 175 to 180 and columns from the column 9005 to 9010:

```
inspect(dtmStem[175:180, 9005:9010])
```

With the following output:

```
Docs  oustand outbreak outburst outcom outcrop outdat

  175       0        0        0      0       0      0
  176       0        0        0      0       0      0
  177       0        1        0      0       0      0
  178       0        0        0      0       0      0
  179       0        0        0      0       0      0
  180       0        0        0      0       0      0
```

Displayed portion of the matrix has only one non-sparse entry for the document number 177 which contains one term *outbrake*.

It is possible to output the terms which have frequency above some define value. With the following command the terms with the frequency 700 and more are displayed:

```
findFreqTerms(dtmStem,700)
```

Next command plots the correlation of the terms which have frequency above 780; the edge between words is draw if the correlation is above 0.4 (ⁱRgraphviz libray needs to be installed for plotting):

```
plot(dtmStem, terms = findFreqTerms(dtmStem,780), corThreshold = 0.4)
```

The Figure 5 shows the correlation of most frequent words from the corpus.

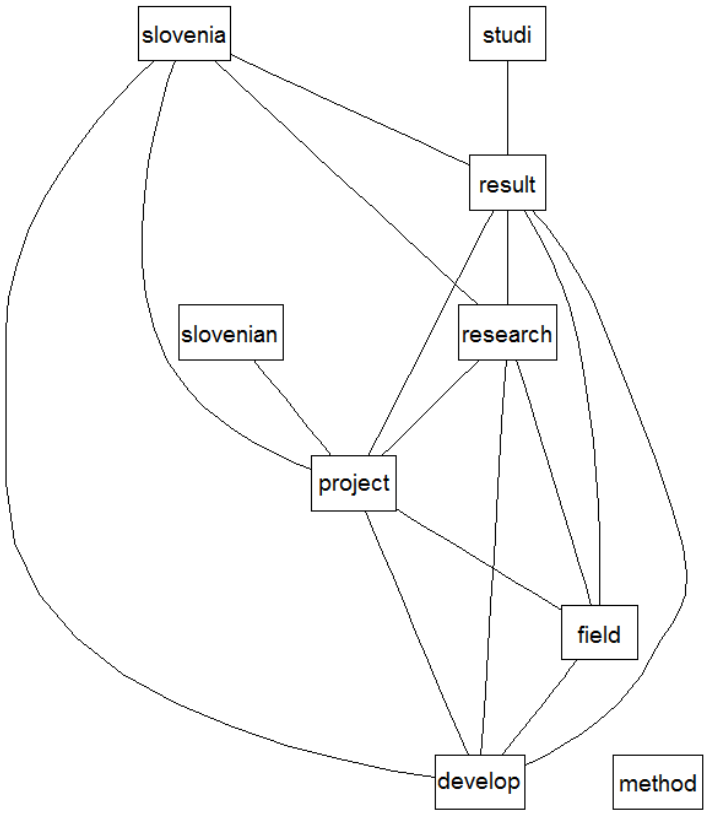


Figure 6 - Correlation between most frequent words from the corpus

The most frequent terms from the corpus and associations between them, give a nice summary of the whole corpus. The dominant terms: *project*, *research*, *slovenia*, *develop*, *result*, *study*, *field*; and connections between them, are good high level indicator that these corpus is about research in Slovenia.

R has the functionalities to work with document-term matrices of n-grams. For this the RWeka package must be installed. Next command illustrate creating document-term matrix of bigrams and creating a plot of most frequent bigrams of the corpus (Figure 7):

```
library("RWeka")
```

```
BigramTokenizer <- function(x) NGramTokenizer(x, Weka_control(min = 2, max
= 2))

dtm2Gram <- DocumentTermMatrix(prjCorpusEnStem, control = list(tokenize =
BigramTokenizer))

plot(dtm2Gram, terms = findFreqTerms(dtm2Gram,80), corThreshold = 0.15)
```

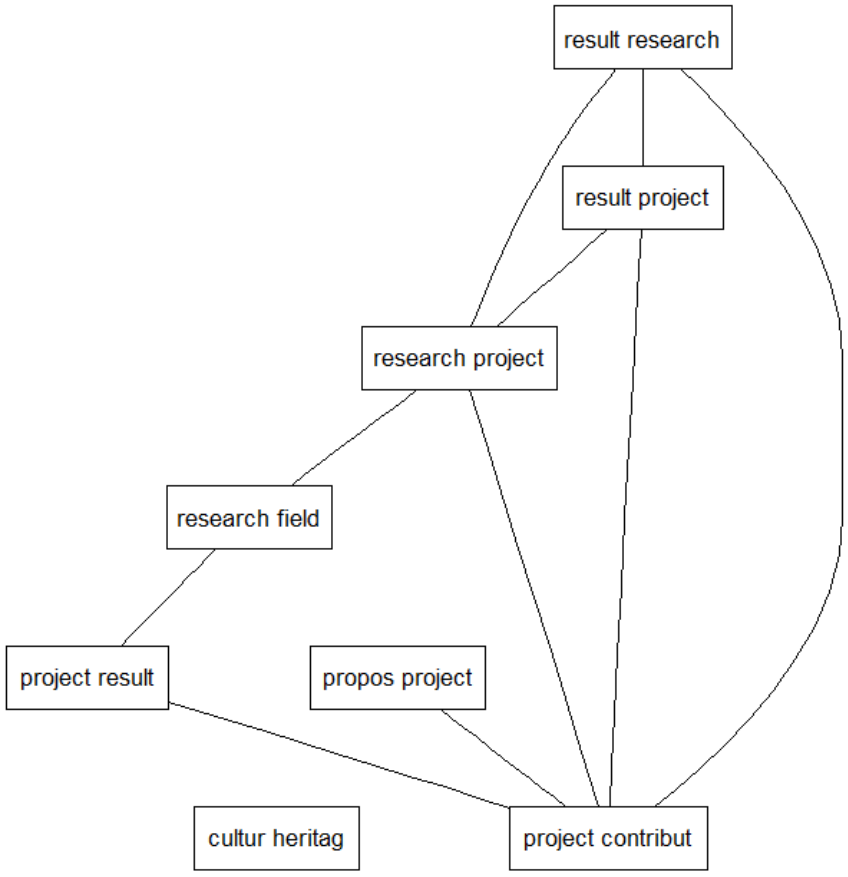


Figure 7 - Correlation of most frequent 2-grams

The correlation between most frequent bigrams nicely shows Slovenian research projects domain of the corpus.

3.1.2.2 Term-document matrix reduction

Since the term-document matrices tend to be very big, there are methods to remove sparse terms (terms occurring in very few documents), which can dramatically reduce the size of the matrix. Sparse terms can be removed using the removeSparseTerms function. This function takes term-document matrix and value for the maximal allowed sparsity as arguments. Performing this function on a term-document matrix, all terms which have sparsity larger than the specified value for allowed

sparsity will be removed from the term-document matrix. Every term occurs at least in one of the documents from the term-document matrix; otherwise the term would not be in the matrix. This means that every term has sparsity at least divided by the number of documents in the term-document matrix decreased for 1, divided by the number of documents in the term-document matrix. Previously constructed matrix with documents representing research projects contains 2054 documents. If we apply the `removeSparseTerms` function on this matrix and set the maximal allowed sparsity to 1-(1/2054), the matrix should remain the same. This can be tested with the following command:

```
removeSparseTerms(dtmStem,1-(1/2054))
```

The actual removal of sparse terms can be accomplished if all the terms which occur in less than two documents are removed. The maximum allowed sparsity for these terms in the projects term-document matrix is 1-(2/2054):

```
dtmStem2min <- removeSparseTerms(dtmStem,1-(2/2054))
dtmStem2min
```

The sparsity of the newly created matrix dtmStem2min is decreased to 99% which is still very sparse matrix, but the number of terms is 5762, which is decrease for almost 60% compared to the dtmStem matrix:

```
wsHC50 <- hclust(dist(tdm50), method="ward")
```

```
plot(wsHC50, hang = -1, labels=FALSE, ann = TRUE, sub=NULL, xlab=NULL,
ylab=NULL)
```

3.1.2.3 Empirical laws

Using the *tm* package, R enables validation of some empirical laws from mathematical statistic, namely: Zipf's and Heap's law. Zipf's law, named after the Harvard linguistic professor George Kingsley Zipf (1902-1950), states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. We can conveniently explore the degree to which the law holds for our corpus by plotting the logarithm of the frequency against the logarithm of the rank, and inspecting the goodness of fit of a linear model. The plot can be made with the following command (Feinerer, R Documentation):

```
Zipf_plot(dtmStem
```

The result of the command is the plot in Figure 8.

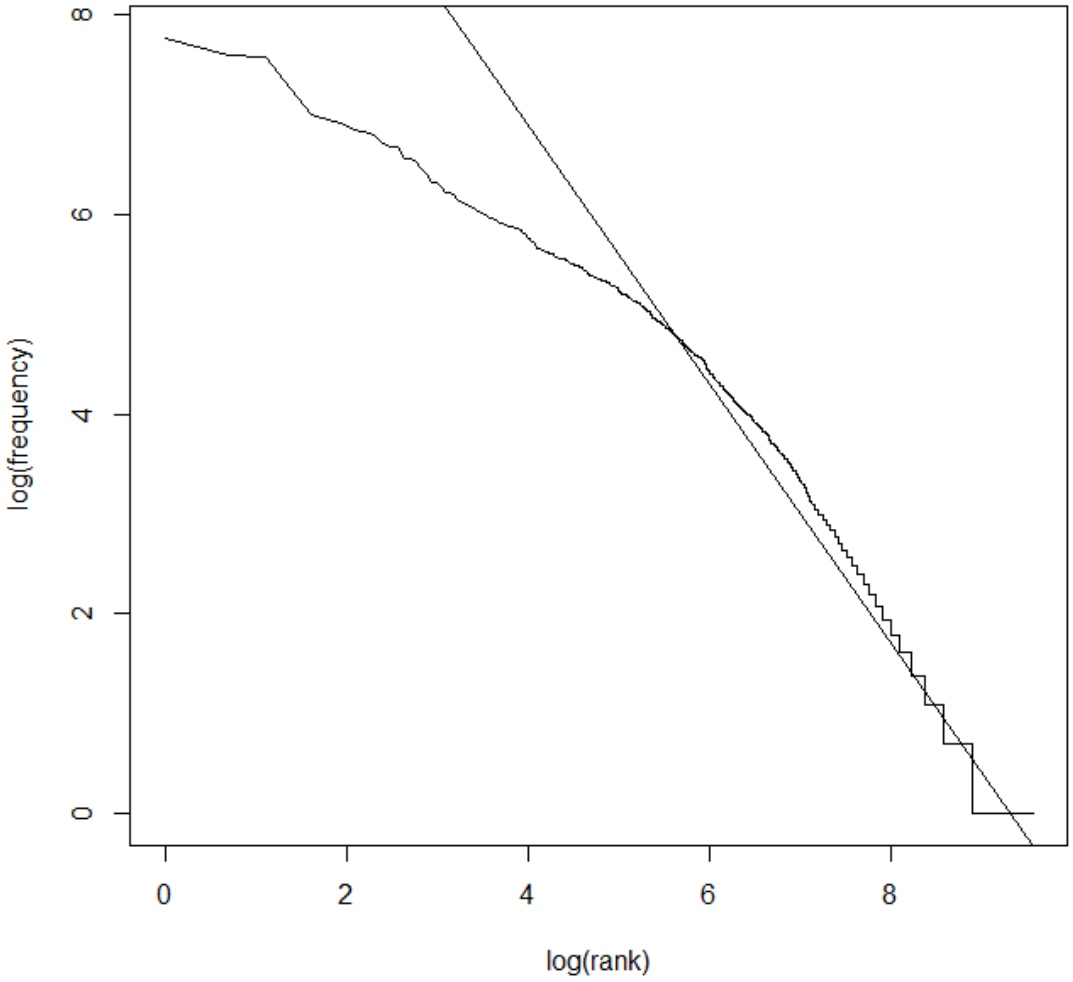


Figure 8 - Logarithm of the frequency against the logarithm of the rank of the projects corpus, with linear line representing the Zipf's law

It can be noticed that for the lower ranked documents, the frequency does not follow the Zipf's law function, i.e. the frequencies of the lower ranked words are lower than predicted with Zipf's law.

Heaps' law states that the vocabulary size *V* (i.e., the number of different terms employed) grows polynomial with the text size *T* (the total number of terms in the texts). The graph which plots logarithm of vocabulary size against the logarithm of text size can be plotted with the following command (Feinerer, R Documentation):

```
Heaps_plot(dtmStem)
```

The result of the command is the plot in Figure 5.

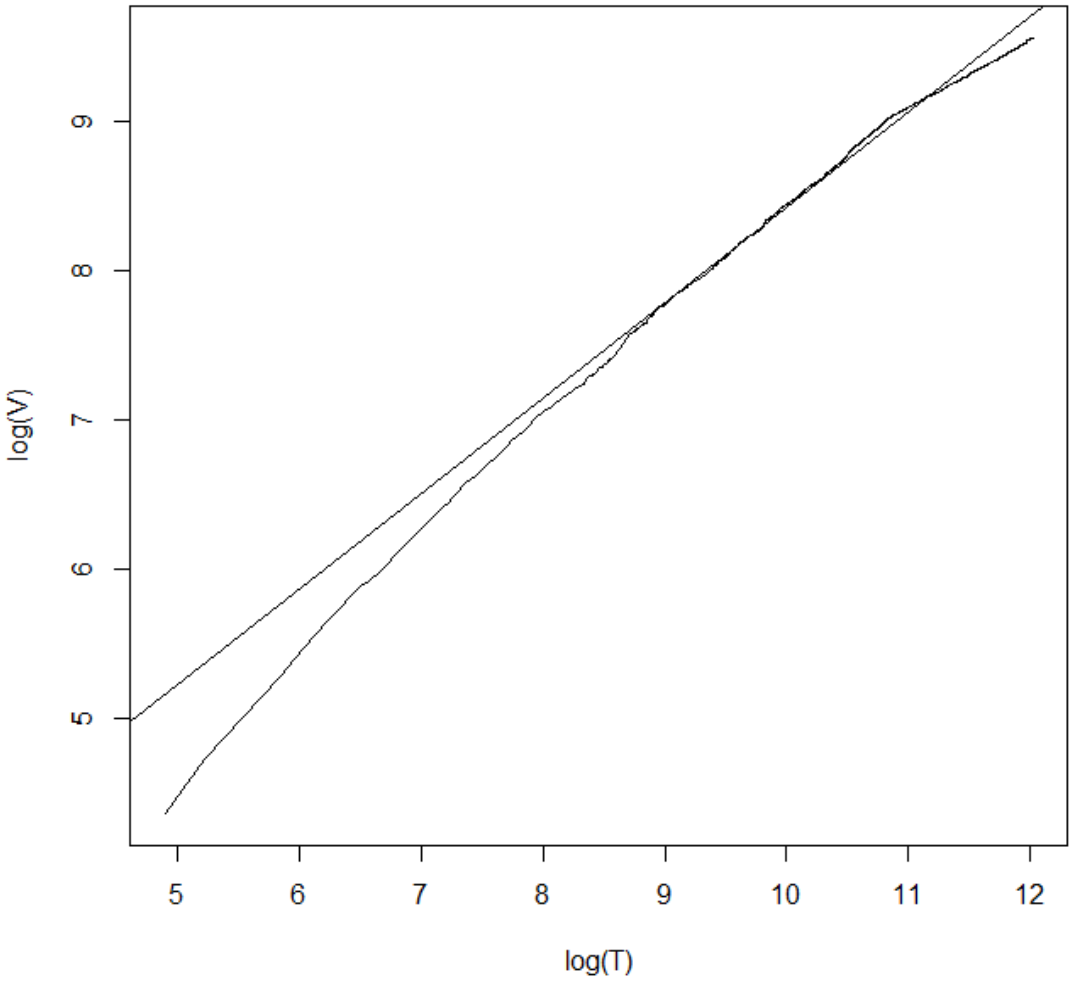


Figure 9 - Logarithm of the vocabulary size agains the logarithm of the text size of the projects corpus, with linear line representing the Heap's law

The plot from figure 9 shows that for the smaller number of total terms employed the vocabulary grows slower than predicted with the Heap's law, but for the bigger number of total terms the vocabulary grows in accordance to Heap's plot function.

3.1.2.4 Creating word cloud with R

Word cloud is visualization useful for highlighting the most commonly used words from some corpus. Based on examples from (Sonego, 2011) and usage of wordcloud (Fellows, 2012) package, next commands show creation of the word cloud of projects corpus:

```
library(wordcloud)
```

```
tdm <- TermDocumentMatrix(prjCorpusEnStem)

m <- as.matrix(tdm)

v <- sort(rowSums(m),decreasing=TRUE)

d <- data.frame(word = names(v),freq=v)

pal2 <- brewer.pal(8,"Dark2")

png("wordcloud_packages.png", width=1280,height=800)

cloud(d$word,d$freq, scale=c(8,.3),min.freq=2,max.words=300,
random.order=T, rot.per=.15, colors=pal2)

dev.off()
```

Figure 10 - Word cloud for the projects corpus

The Figure 10 shows the word cloud of the textual corpus. The biggest and most visible words in the word cloud are the most frequent ones. These words are: “research”, “develop”, “project”, “slovenia”, “system”, “result”, etc., and one could guess that the domain of the corpus described by this word cloud are the Slovenian research projects.

3.2 Topic Ontology in Ontogen

To get a deeper insight into the corpus, Ontogen (Fortuna, Grobelnik, & Mladenic, OntoGen: Semi-automatic Ontology Editor., 2007) utility was used to visually inspect the data and create topic ontology in the semi-automatic fashion. Ontogen is an ontology editor as implemented which integrates machine learning and text mining algorithms into an efficient user interface. The main features of the systems include unsupervised and supervised methods for concept suggestion and concept naming, as well as ontology and concept visualization.

First the visual inspection of the data is made, following by the process of creating topic ontology from the projects corpus data.

3.2.1 Visual inspection of the data

New ontology was created using the Ontogen system by `File>New ontology>Folder` command. The selected folder was the folder containing the corpus of projects documents created with R (and described in chapter 2.1.2 - Data loading and preprocessing operations in R). The preprocessed textual corpus consists of 2054 documents written in a folder with one textual file for each document. Loaded corpus was visualized using the Concept Visualization command from the Ontology details menu.

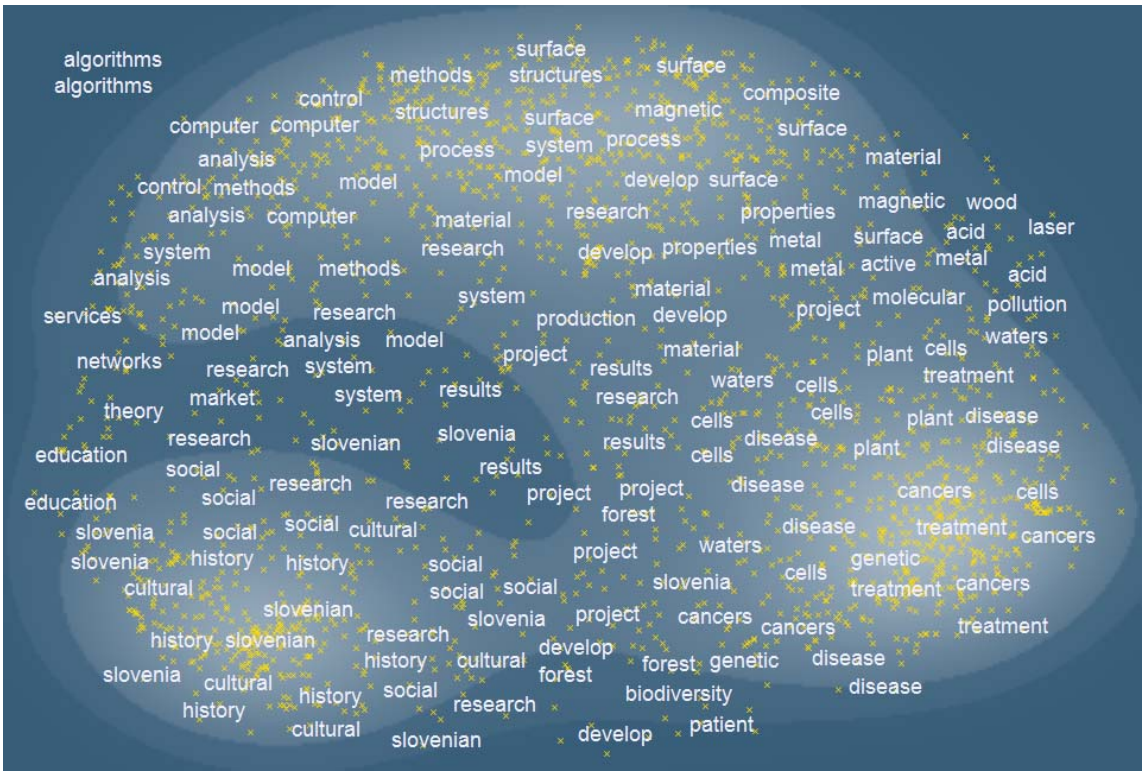


Figure 11 - Visual representation of the research projects corpus using Ontogen

The Figure 11 shows the documents of the projects corpus visualized in the 2-dimensional plane. Each yellow cross represents a document from the corpus. Words around the crosses are the keywords which describe the content of the documents in the surrounding. The denser areas containing more documents have brighter background. The visualization indicates there are three major clusters of documents. The first dense group of documents is on the top of Figure 11. Keywords like: computer, surface, system, model, material, develop, control, etc., indicate that the documents in this group are mostly from engineering and technical sciences. The group in the lower left corner of the Figure 11 contains documents from social sciences and humanistic, as indicated by keywords like: social, cultural, history, education, etc. Finally, the third group is located in the lower right corner of the Figure 11 and the keywords which describe it are: diseases, genetic, treatment, cells, cancers, etc., which can be related to medicine.

3.2.2 Semi-automatic construction of research projects topic ontology

The topic ontology was created using the Ontogen functions: suggestions – for automatic suggestions of desired number of subtopics, query – for building a subtopic using the active learning method, concept visualization – for inspection of the documents using in the 2D view and concept documents – for detail inspection of the documents content. Different functions were used in different stages of topic ontology construction.

The first step in generating topic ontology was building a topic with documents from medical sciences, by performing command `Query` and labeling around 15 documents as belonging to medicine topic or not. The topic named *medicine* was created, containing 612 documents. Next, from 1442 unused document from root, 4 suggestions for topics was made, from each one containing 378 documents was created. The topic contains these keywords: "*cultural", "social", "research", "slovenia", "history", "Slovenian", "project", "education", "politics", "develop''*; and was named *Social sciences and humanities*. From next 4 suggestions of 1064 unused documents in the root new topic was created, containing 234 documents and described with the following keywords: "*waters", "forest", "soil", "plant", "karst", "pollution", "biodiversity", "natural", "species", "carbon"*. The topic was named Biotechnologies. Using the query "*natural science chemistry biology mathematics"*, Natural sciences topic with 657 documents was created. The lastly created topic was Engineering and it takes all 170 unused documents from root. Keywords describing Engineering topic are: *"structures", "production", "system", "model", "computer", "project", "develop", "control", "machine", "services".* All unused documents can be used to create new topic by simply performing a query with keywords that describe all the documents. The ontology build till this phase is shown in Figure 12.

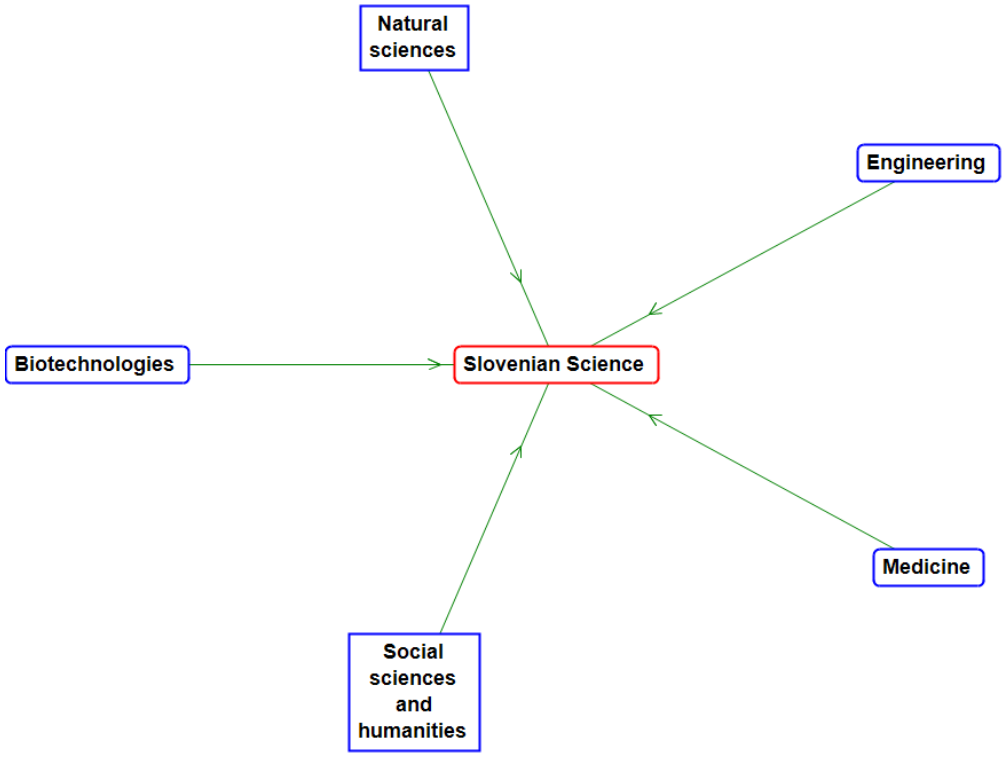


Figure 12 - First level of research project topic ontology

The topics on the next level were mostly created using the suggestions made with the Ontogen utility. The reason for this is lack of detail domain knowledge for every science group. The Medicine was further divided in the following topics: radiology, hart diseases, cancer treatment, sensors and equipment and genetics. Subtopics for social sciences and humanities are: sociology, anthropology, art; law, politics; and economy. With the same technic of generating suggestions, biotechnology gets subtopics: forestry; green biotechnology and blue biotechnology. Natural sciences are divided in the subtopics: Nanotechnologies, chemistry, biology and mathematics. Finally engineering is subdivided to: industrial engineering, mechanical engineering, networks and computer science. The complete topic ontology is shown in Figure 13.

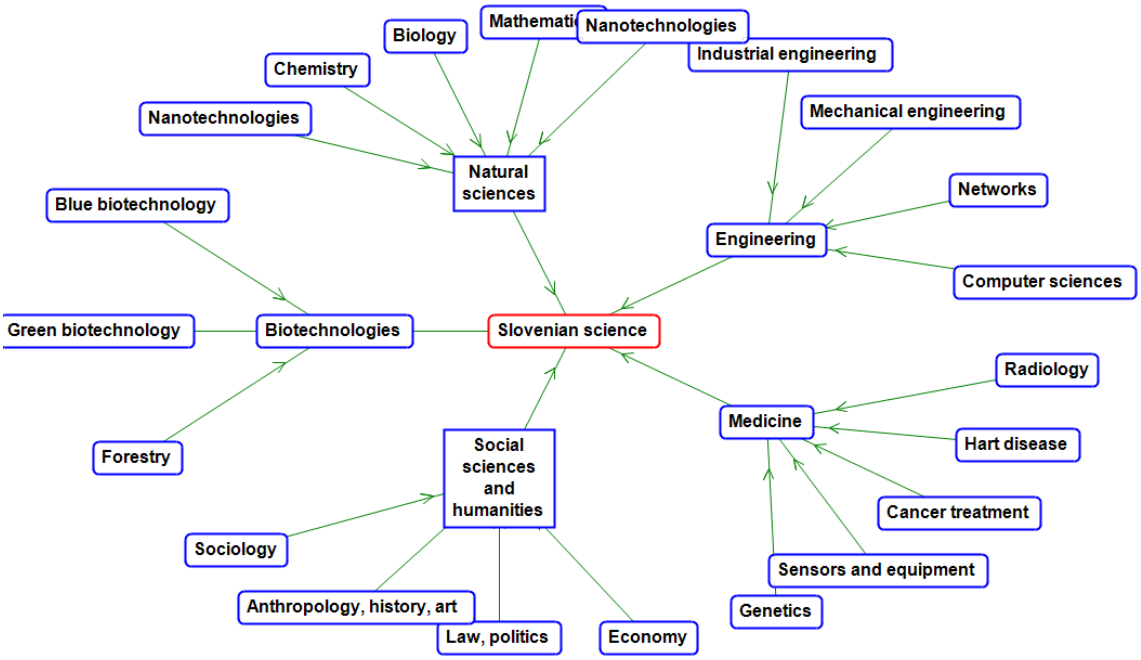


Figure 13 - Topic ontology of research projects corpus

3.3 Text Classification using Text-Garden

In this part the classification of researchers based on their projects is performed. The classification is performed using the Text Garden (Artificial Intelligence Laboratory - Institute Jozef Stefan). Text-Garden is a set of Text-Mining Software Tools that enable easy handling of text documents for the purpose of data analysis including automatic model generation and document classification, document clustering, document visualization, dealing with Web documents, crawling the Web and other.

First, the process of preparation of data and construction files suitable for inputting into Text Garden is described. Next, the approach for classification process is explained. This includes the preprocessing steps on the text, constructing the feature vector and the classifier. Finally, the results of the testing on the prepared datasets are given.

3.3.1 Data preparation

The classification was performed on researchers. Researchers are categorized into science, field and subfield according to the ARRS classification. The ARRS categorize of researchers are used as the labels for the classification problem. The descriptions associated with the projects of each researcher are used as the attributes according to which classification is performed. Two datasets were prepared, first one where labels for the researchers are the scientific fields to which the researchers are categorized and the second one where labels are science groups to which the researchers are categorized in. The first dataset represents much harder classification problem, because the

researcher must be classified in a correct filed among 74 possibilities, whereas the second dataset does not represent so hard problem as there are only 6 different science groups.

3.3.1.1 Building the named line documents

The data used for text classification was taken from the Atlas of Slovenian Science database and transformed into Named Line-Document format used by Text-Garden. Other applications that use this file format are: Document Atlas (Fortuna, Grobelnik, & Mladenic, Visualization of text document corpus, 2005) and Ontogen. Each line of the Named Line-Document format contains one document. The document in a line has three parts: name of a document, categories of a document and the content of a document.

Since the documents represent researchers, the names of the documents were constructed by combining the MSTID of a researcher, first and last name of a researcher. The three parts of the name were combined with the dash, as in the example:

```
12570_Dunja_Mladenić
```

Since there are some researchers which have more than one last name, the whitespaces of the last names were replaced with dashes in the process of constructing the names.

The categories in the named line document format are represented with preceding exclamation mark. Named line document allows multiple categories for a document, but since the most of the researchers have only one category, single categories were used. Instead of codes, the actual names of science fields and science groups were used, so that the interpretation of the document would be simpler. Since the categories can be made of multiple words, the whitespaces were replaced with dashes similar like in constructing the names of the documents.

For the content of the documents, all the data about the projects of the researchers were taken. Only those researchers which participated in at least one project were taken into account. The attributes of the project which were used to construct the content of the document were: name of the project, abstract of the project, keywords of the project, domestic and world significance of the project. The projects of the researchers in the Atlas of Slovenian science have descriptions in both Slovene and English language. Since the classification was based on text in English language, the attributes for the descriptions of project in English were obtained. But even thou the database support English and Slovene data, many projects contain Slovene names in the fields intended for English. These projects were not included in the constructed dataset and they could be identified by mark “(Slovene)” in the ending of the project name. To avoid project with names written in Slovene language in the fields intended for English, following condition was used in the SQL query used to obtain the data:

```
where name_en not like '%(Slovene)'
```

Finally constructed document consist of 6000 lines where each line represents a researcher. These are two example lines from the finally constructed named line document:

```
12570_Dunja_Mladenić !Computer_science_and_informatics Analysis of large…
13325_Mitja_Jermol !Systems_and_cybernetics Intelligent materials, polya…
```

3.3.2 Approach

The created named line-document files were used for creating the bag-of-words (BOW) objects. A BOW is a structurally simple representation of text produced without linguistic or domain knowledge (David, 1998). Text is represented as a vector which elements are frequencies of the words of represented text. BOW object was created with the process of excluding stop-words from the text, using the predefined list for English language named – EN 523. Stop-words are words commonly used to form the structure of the sentence (e.g. “the”,” a”, “of”, “and”, etc.) (Joachims, 1998). The BOW was created with performing stemming. Stemming is a word-level transformation, with the task of grouping the words derived from a common stem (e.g. grouping “fish”, “fishes”, and “fishing”) (Croft, Metzler, & Trevor, 2010). In the BOW representation, the structure of the sentence is lost and these words used on their own do not contribute to information about content of the text.

Next, TF-IDFs (term frequency–inverse document frequency) for the BOW were computed. TF-IDF is a statistical measure used in text -mining, which outperforms a simple count of words, by increasing the importance of less common words. (Jones, 1972)

Finally, based on the BOW and the computed TFIDF values, Support Vector Machine (SVM) classification model was built. SVM is a machine learning method that non-linearly maps the input vectors to a very high dimension vector space in which the decision surface is constructed (Cortes & Vapnik, 1995). In (Joachims, 1998) it is shown that the SVM is an appropriate method for text classification. The main reasons include the ability to handle high dimensional input space and suitability for problems with dense concepts and sparse instances.

3.3.3 Results

The classification model was tested using 5 folds cross validation technique. The datasets were divided into 5 subsets of equal size. Five experiments were made, each time different subset was a testing sample and the other four were the training samples. The average of five experiments gives the final result.

The results of testing the dataset with 74 science fields give the classification accuracy of 0.68. This means that 68% of the researchers were classified in one of 74 scientific fields to which they are actually categorized. The classification accuracy for the dataset with 6 different science groups is 0.84.

3.4 Conclusion

We have described tree aspects of dealing with textual corpus: preprocessing and exploring the properties of text, clustering and developing data driven topic ontologies and, classification of textual documents.

Various possibilities of discovering properties of text using R software are described in the first part of this Chapter. Covered are various preprocessing operations like: white space removal, numbers removal, converting to lowercase, stop words removal and stemming, but also construction of visualizations like: word cloud and term co-occurrence graph. It was shown that the corpus can be validated against some empirical laws, like Zipf's and Heap's law.

In the second part topic ontology based on the textual corpus of research project was build. The ontology consists of 5 subtopics at the 1st level from the root and 21 subtopics on the 2nd level. It is build using the methods for semi-automatic ontology construction like: active learning by performing queries, K-Means classification of text by performing suggestions and visual inspection of document visualizations on 2-dimensional plain constructed using latent semantic indexing and dimensionality reduction.

In the last part of the Chapter classification of textual documents is described. Two problems were constructed: classifying researchers to one of 74 science fields and into one of 6 science groups. Both classifications were made on the basis of textual descriptions of projects researchers were conducting. Classification was performed with Text-Garden utility using the SVM classification model. The classification accuracy of 0.68 and 0.84 for first and second problem respectively, can be considered as good results, taking into account that researcher can be involved in projects that are different from their official categorization.

4 Modeling the evolution of data with social network analysis

This chapter gives the results of analyzing the dynamical aspects of the network of researcher collaboration. The chapter has two parts, first one in which the data preparation process is described and the second one with actual results of the performed analysis.

The evolution of the network in time was analyzed using the Stanford Network Analysis Package (SNAP) library. SNAP is a general purpose network analysis and graph mining library that easily scales to massive networks, is efficient and easily extendible. It is written in C++ and easily scales to massive networks with hundreds of millions of nodes, and billions of edges. It efficiently manipulates large graphs, calculates structural properties, generates regular and random graphs, and supports attributes on nodes and edges. SNAP was built on top of a general purpose STL (Standard Template Library)-like library GLib that was developed at Jožef Stefan Institute. (Leskovec)

4.1 Data preparation

Network data was obtained from the Atlas of Slovenian Science database and written into textual files containing the edges of the network in the vertex pair format. Network was created using the data about projects and researchers taken from Slovenian Current Information System (SICRIS) database in March 2011 and it includes both active and not active, and projects from 1994 to 2010.

Each vertex represents a researcher and is symbolized with the MSTID of the researcher, and the edge represents collaboration of researchers on projects. Each edge has weight assigned to it, defining the number of different projects on which two researchers have collaborated. Atlas of Slovenian science database contains data for the projects dating from year 1994 to 2010. For the purpose of social network evolution analysis, 17 files were created, each containing the snapshot of collaboration for one year.

Created files were used as in input to applications that implement SNAP library. The outputs of the applications are plots and tabular data for various analyses. Tabular data is used in R application, where some additional analysis is performed and the figures are plotted.

4.2 Researcher Network Evolution Analysis

Four different aspects of researcher network evolution were analyzed and reported. The network was analyzed from the aspect of network size, density of the network, network diameter and connected component of the network, taking into the temporal dimension into focus of the studies.

4.2.1 Network Growth

In this section it is examined how the researcher network collaboration grows in respect to number of nodes and edges thru period of 17 years (from 1994 to 2010). The Figure 14 shows the number of nodes in the network in each year. The node becomes the part of the network when it creates first edge with another node. In this context arrival of new node is event when a new researcher (first time mentioned in the project from the dataset) participates on some project. The number of nodes grows linearly in time. The growth can be approximated with the linear function $f(x) = 600.5x + 815.8$, with squared R error (R^2) of 0.9636. Number of nodes in year 1994 was 159 and the final number of nodes in year 2010 is 10341. These nodes represent researchers involved in at least one research project.

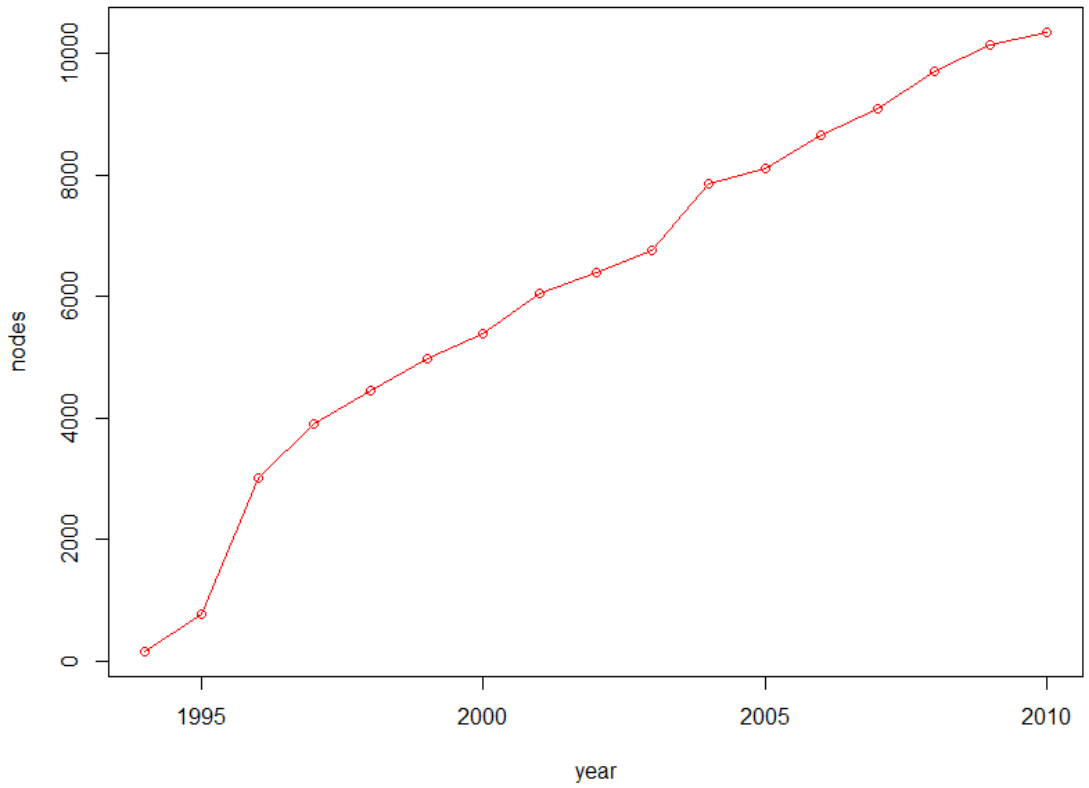


Figure 14 - Number of nodes of the network in time

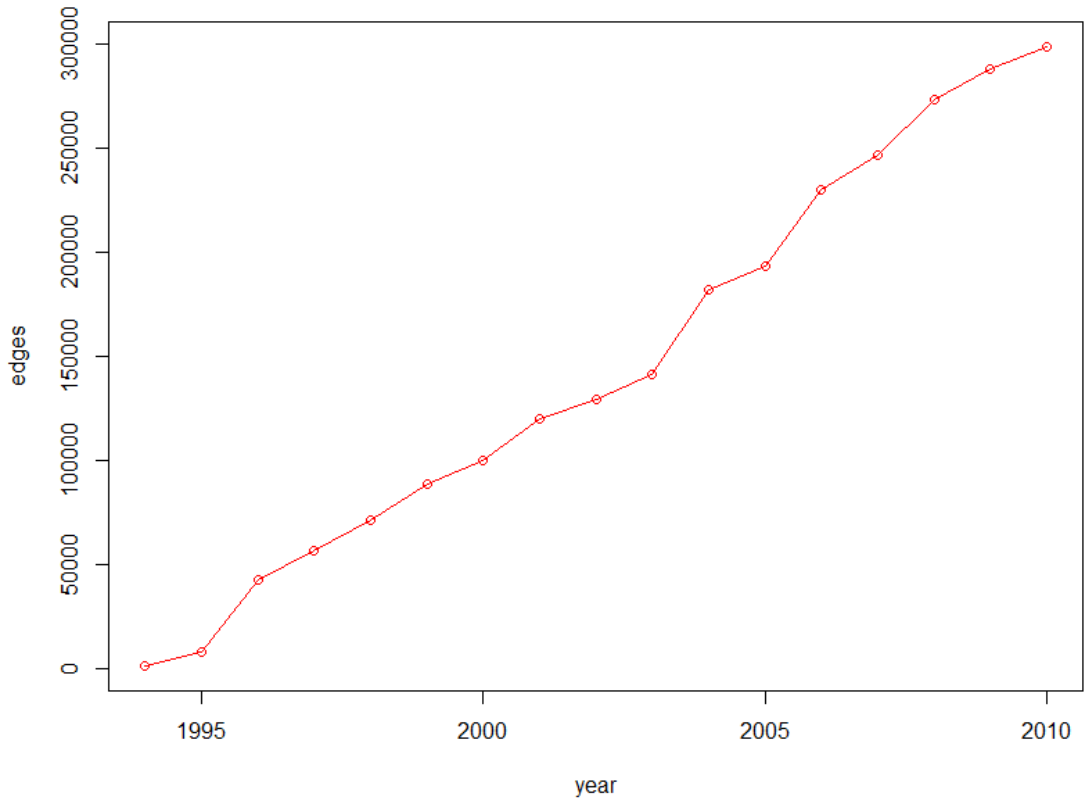


Figure 15 - Number of edges of the network in time

The Figure 15 shows the growth of the number of edges in the network thru time. The numbers of edges increase linearly, the liner function: $f(x) = 19111x - 26655$ approximates the growth with$R^2 = 0.9893$. The initial number of edges in the year 1994 is 1749 and the final number of edges in year 2010 is 298453.

4.2.2 Density of the Network

In this section the density of the network through time is observed. The empirical observations show that the network is becoming denser over time, with number of edges growing superlinearly in the number of nodes. It is found that the network of researcher is growing according to the densification power law (Leskovec, Kleinberg, & Faloutsos, Graph evolution: Densification and shrinking diameters, 2007):

$$e(t) \propto n(t)^a,$$

Where $e(t)$ and $n(t)$ denote the number of edges and number of nodes of the graph at time t, and a is an exponent that generally lies strictly between 1 and 2. Exponent $a = 1$ corresponds to constant average degree over time, while $a = 2$ corresponds to an extremely dense graph where each node has, on average, edges to a constant fraction of all nodes.

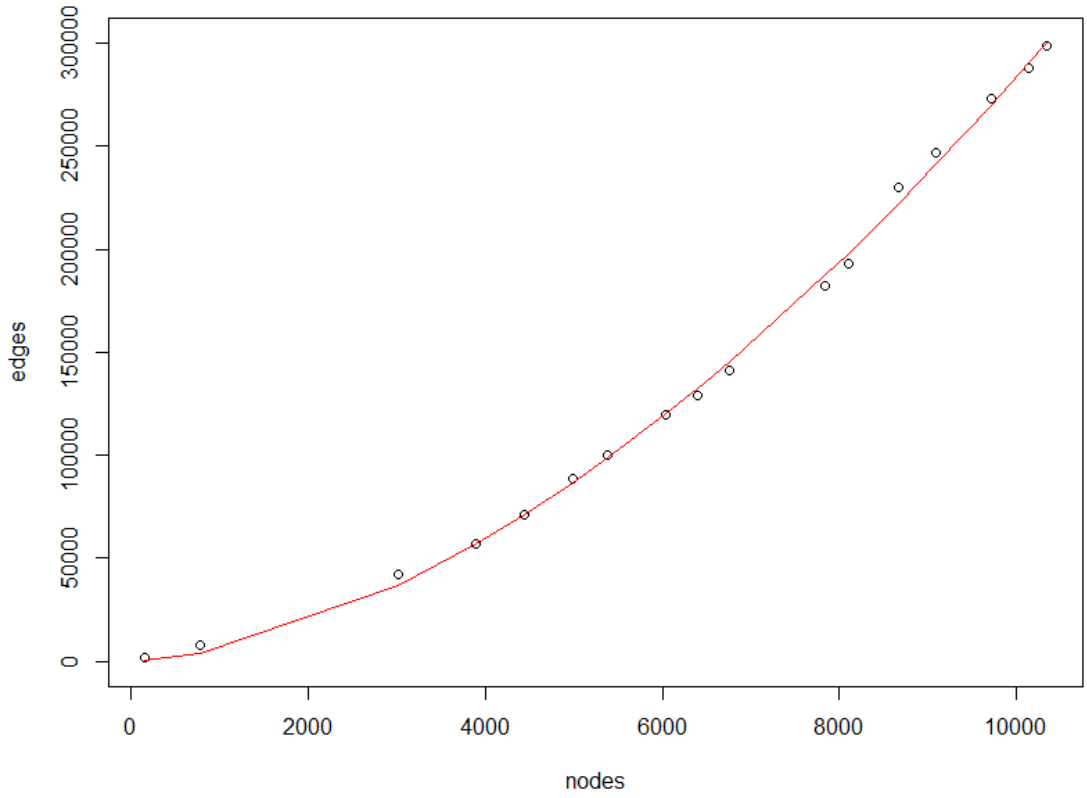


Figure 16 - Number edges against number of nodes in the network

The exponent a in the densification power law for the examined network is 1.70. For the examined network a is significantly higher than 1, indicating a large deviation from linear growth. This means that the average degree is increasing over time and the network is becoming denser.

Figure 16 shows the number of edges against the number of nodes. It can be noticed that the edges grow exponentially to the number of nodes. The red line on the figure represents the exponential function: $f(x) = 0.04417x^{1.70191}$,that approximates the dynamics of number of edges versus number of nodes in the network. Initially in with 159 nodes there were 1749 edges and finally with 10341 nodes, the number of edges was 298453. Figure 17 shows the same data as Figure 16, only with x and y axis in logarithmic scale. The plot on Figure 17 is the log-log plot of number of edges $e(t)$versus number of node $n(t)$and it is called the Densification Power Law Plot (DPL). The slope of the DPL plot corresponds to the exponent in the densification power law, which is $a = 1.70$ in this case.

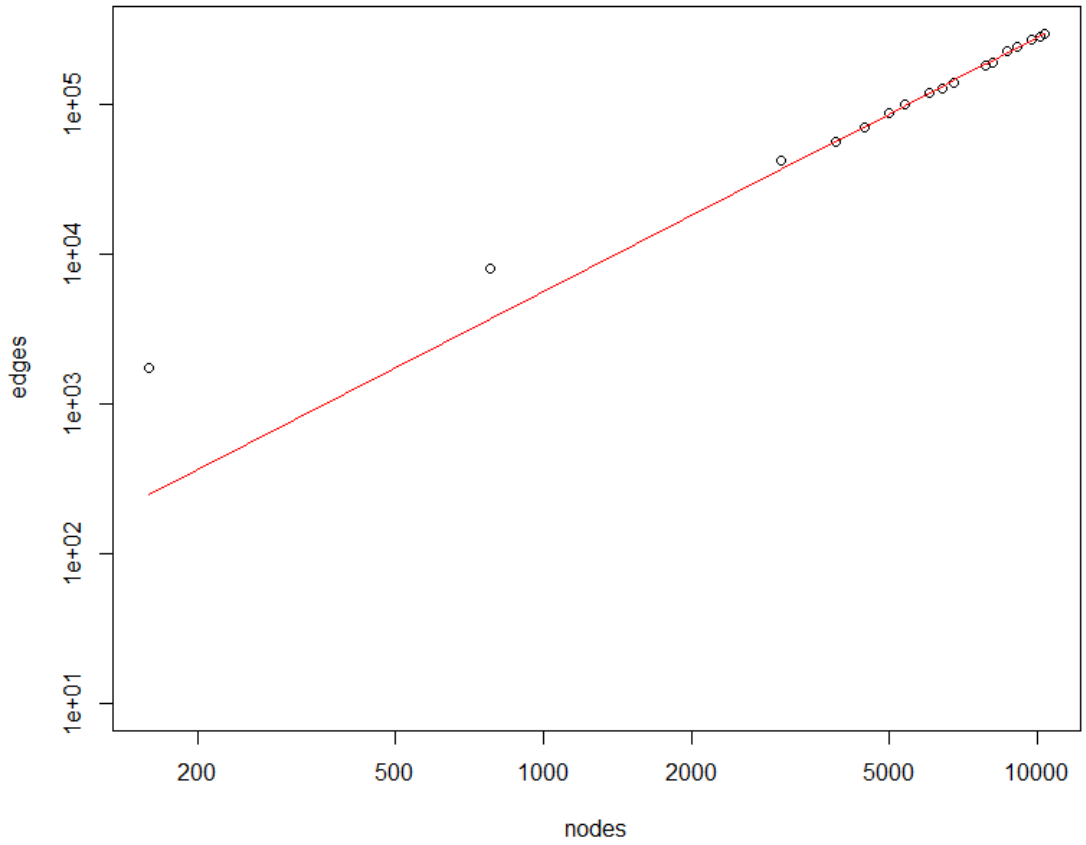


Figure 17 - Densification Power Law Plot (DPL) - number of edges versus number of nodes in a log scales for both x and y axis

4.2.3 Diameter of the Network

In this section, the behavior of effective diameter over time is inspected. Two nodes in a graph are connected if there is an undirected path between them. For each natural number d, let g(d) denote the fraction of connected node pairs whose shortest connection path has length at most d. Effective diameter of the network is the value of d at which the function $g(d)$ achieves the value 0.9. In other words, the effective diameter is the smallest number of hops at which at least 90% of all connected pairs of nodes can be reached. Definition of network diameter differs from the definition of effective diameter. The diameter of the network d is the maximum length of undirected shortest path over all connected pairs of nodes. The effective diameter is a more robust quantity than diameter, since the diameter is prone to the effects of degenerate structures in the graph. (Leskovec, Kleinberg, & Faloutsos, Graph evolution: Densification and shrinking diameters, 2007)

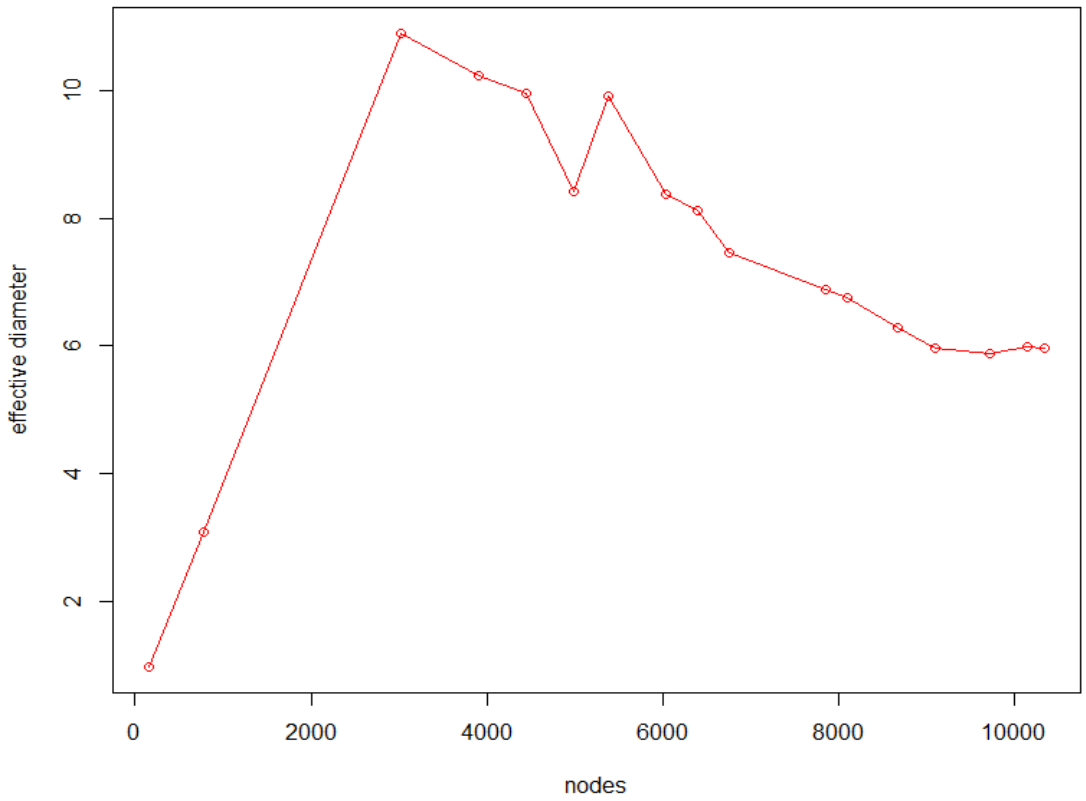


Figure 18 - Effective diameter of network versus number of nodes in the network

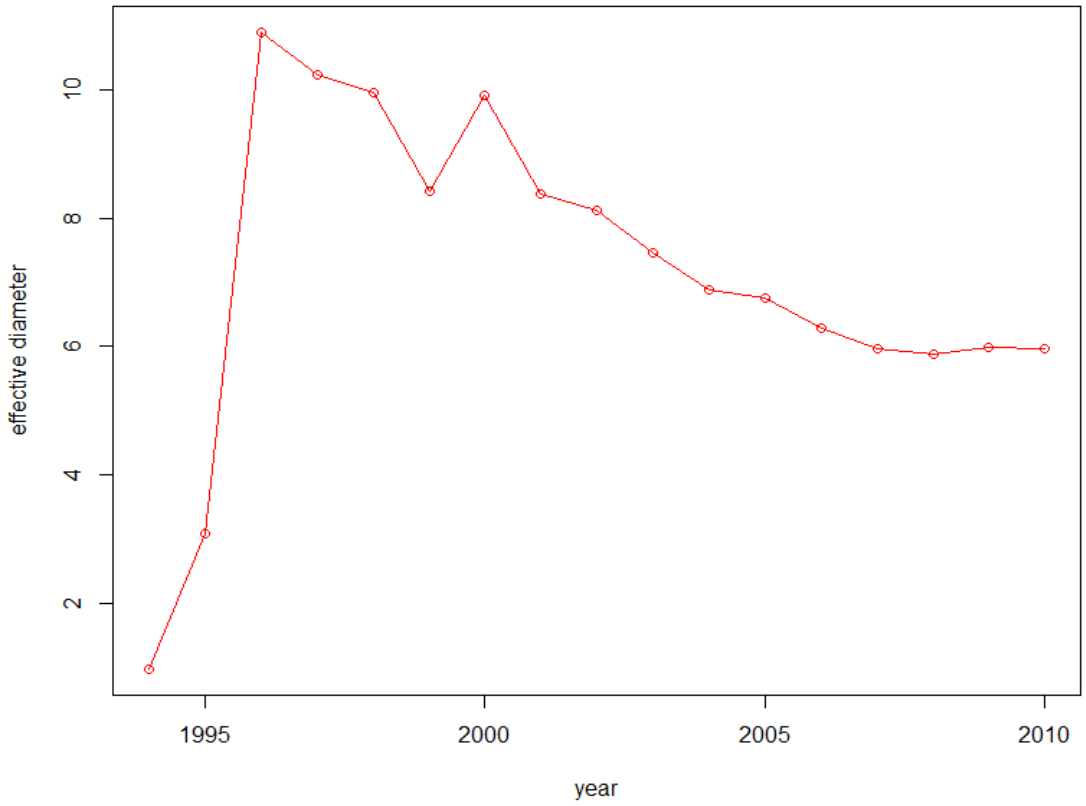


Figure 19 - Effective diameter of the network in time

Figure 18 and 19 show the effective diameter of the network versus the number of nodes in the network and through the years. It can be noticed that in the first few years or until the number of nodes grows to around 3000, the effective diameter of the network grows, but after that period it is constantly shrinking. To confirm the diameter behavior of the network in time, Figure 20 shows the effective diameter of the largest connected component in the network.

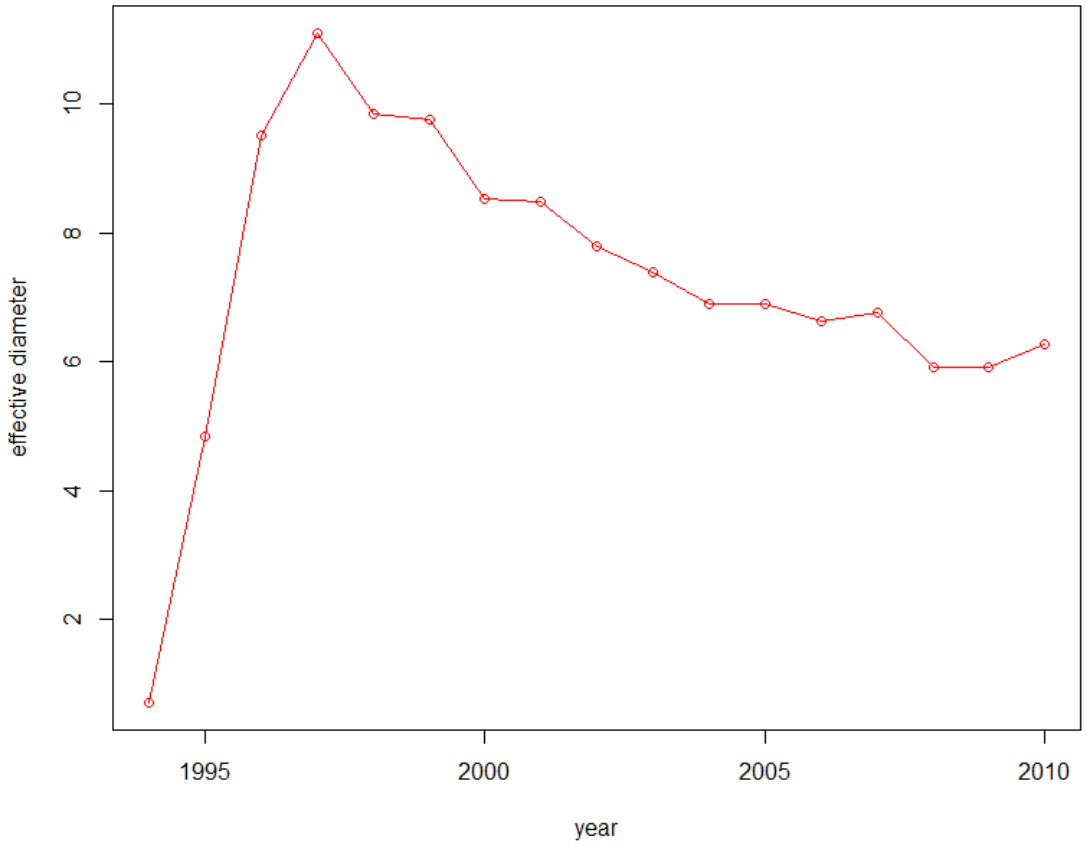


Figure 20 - Effective diameter of largest connected component in the network

The effect of shrinking of effective diameter with time is the same for the complete network and for the largest connected component. The network had biggest effective diameter of 10.91 in year 1996, while size of effective diameter in last network snapshot from 2010 is 5.96. This means that at least 90% of all connected pairs of nodes can be reached within 6 hops.

4.2.4 Connected Component of the Network

In this section we analyze size of the largest connected component of the network with respect to time. In undirected networks, where nodes are connected with edges rather that with arcs, weekly connected component is identical to strongly connected component. Since examined network of researcher's collaboration is undirected network, the term connected components can be applied, without need to specify if it is weekly or strongly connected. The examined network contains one giant connected component. Analyzed are the dynamics of proportion of nodes belonging to the largest connected component (Figure 21), number of nodes (Figure 22) and number of edges (Figure 23) from 1996 to 2010.

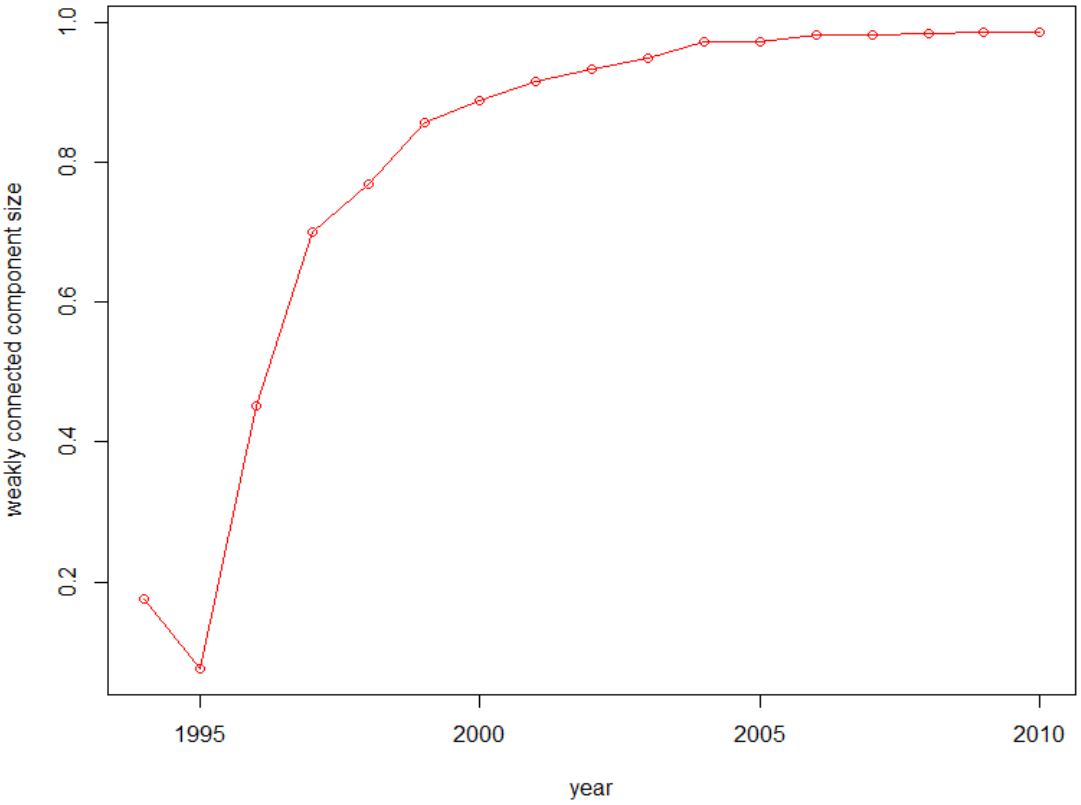


Figure 21 -Proportion of nodes belonging to the largest connected component in time

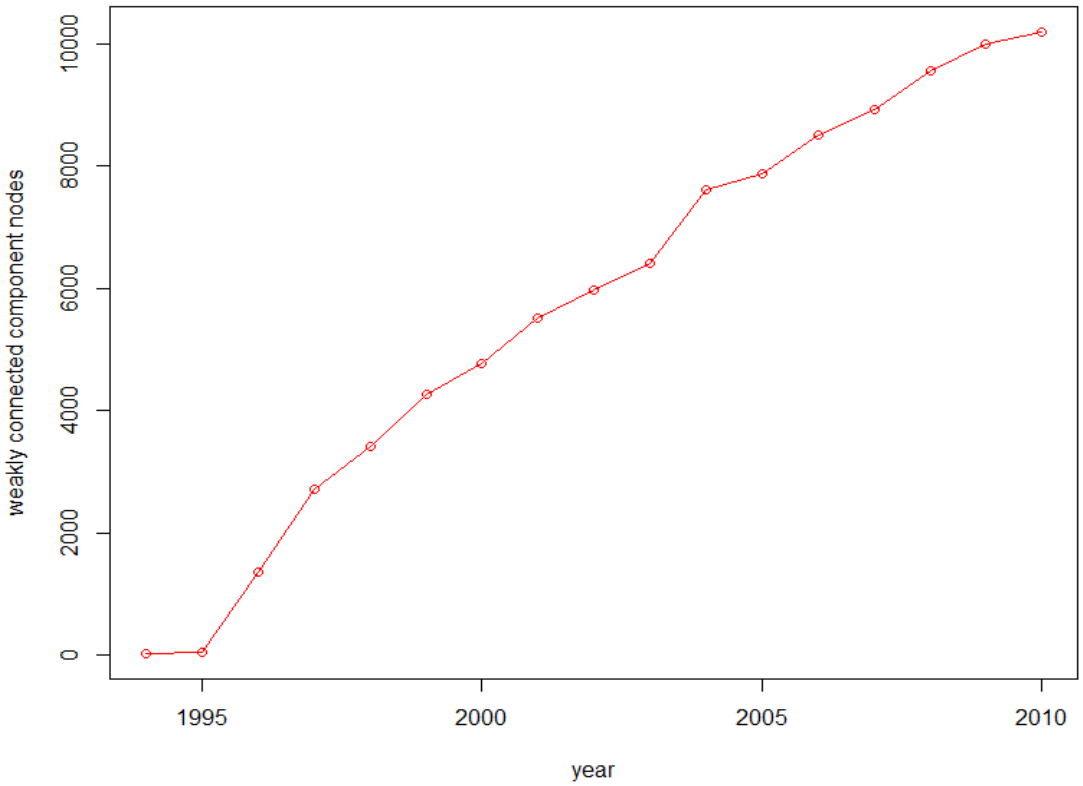


Figure 22 - Number of nodes in the largest connected component in time

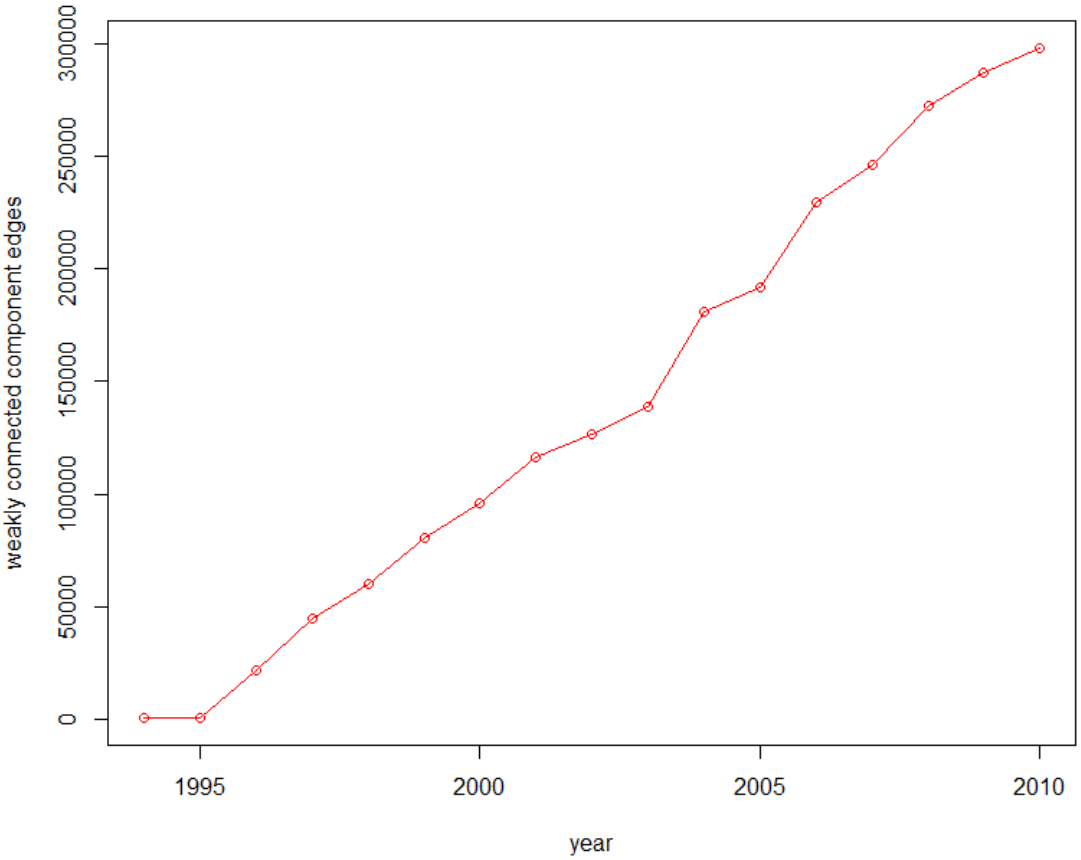


Figure 23 - Number of edges in the largest connected component in time

As shown in Figure 21, the proportion of nodes belonging to largest connected component in respect to all nodes of the network grows logarithmically. This means that the connected component obtains a big proportion of all nodes very fast and then the proportion keeps growing slower. In the first six years (till tear 1999), the size of the connected component reaches 85% of all nodes, while in the next 11 years (till year 2010) the proportion reaches final 98%. The only phase where the proportion does not grow is the "knee" on the figure 21 where the proportion falls from 17% to 7% from initial year 1994 to year 1995.

The growth of the connected component in respect to number of nodes and number of edges is linear, as it can be notices in Figures 22 and Figure 23. The size of the connected component is very large, it takes most of the nodes and edges of the network and it grows linearly. The number of nodes of the network which are the part of the largest connected component in the last snapshot of the network in the year 2010 is 10200 from 10341 of all nodes (0.9863%) and the number of such edges is 297994 from 298453 (0.9985%) of all edges.

4.3 Conclusion

The network was analyzed from the aspect of network size, density of the network, network diameter and connected component of the network. The size of the network can be approximated with linear function. The structure of the network is such that the most of the nodes (98.6%) are

connected in a largest connected component. By examining the density of the network and effective diameter of the network in time, it was discovered that the network is becoming denser and that the diameter is shrinking. The findings are in line with the studies of large networks properties in (Leskovec, Kleinberg, & Faloutsos, Graph evolution: Densification and shrinking diameters, 2007).

5 Modeling the evolution of data with text analysis

In this chapter, the evolution of the textual data in time is examined. The dynamical aspect of the textual data is analyzed in two different approaches. First approach is analyzing the content of the complete dataset and breaking it down into different timeframes. This approach can be called top-down approach. Second approach which can be called bottom-up approach, is examining the content of different snapshots in time separately and then building the global understanding by combining the results of analyzes from different timeframes.

5.1 Data preparation

The data used for this analysis is obtained from the Atlas of Slovenian science database. The Atlas of Slovenian science database contains the data about researchers, research projects, and research organizations, which were collected from the Slovenian Current Research Information System (SICRIS) in March 2011, using the SICRIS web service. Atlas of Slovenian Science contains 33519 researchers and 5384 research projects from the year 1994 to 2010. For the purpose of this analyzes the textual data about research projects written in English language was used. The SICRIS database contains the data in both Slovenian and English language, but some fields intended for English language were empty, or filled with text written in Slovenian language. From 5384 projects, 2250 projects which contained at least the title of the project in English language were extracted for the analysis. Along with the title, other data used as the descriptions for the projects were: keywords, abstract, domestic and world significance.

Data from the database was written into Named-Line document which was used for Ontogen to build topic ontology and making top-down analysis of the topic evolution in time. This document contains all 2250 projects with starting and ending year labels. Also, 17 textual documents were created with the snapshots for each year from 1994 to 2010, containing descriptions of active projects in that year. These documents were used for the bottom-up analyzes of projects dynamics.

5.2 Text Dynamics Analysis

In this part of the chapter the analysis of text in time are performed. First the topic evolution is analyzed, following by the analysis of snapshots content.

5.2.1 Evolution of Topics

To analyze topic ontology, first topic ontology is build is the Ontogen tool. Then the topic ontology evolution was analyzed by examining the temporal attributes of the documents that from the topics of the ontology.

5.2.1.1 Building the Topic Ontology

Using the Ontogen tool the set of 2250 projects from year 1994 to 2010 was analyzed and topic ontology was created. The ontology was created semi-automatically, based on the content of the projects and the domain knowledge. Based on the content of the projects, i.e. cosine similarity between project descriptions, Ontogen can suggest an arbitrary number of clusters by applying K-Means clustering algorithm. For each cluster, Ontogen provides set of keywords that describe the cluster and which can be used for naming the clusters. Rather than using the method of suggesting some number of clusters, for building the topic ontology the active learning method provided by Ontogen was used. Reason for using this method was the sufficient background knowledge about the domain and the dataset, which was acquired by prior analysis of the dataset. Active learning method is in Ontogen implemented by making a query and then accepting or declining some number of different projects. In this way the system 'learns' which projects should be included into a cluster. The created topic ontology consists of six subtopic, these subtopics were created due to the science categories from the ARRS classification. The created subtopics are: medical sciences, natural sciences, engineering, humanistic, social sciences, and biotechnology. The topic ontology is showed in Figure 24.

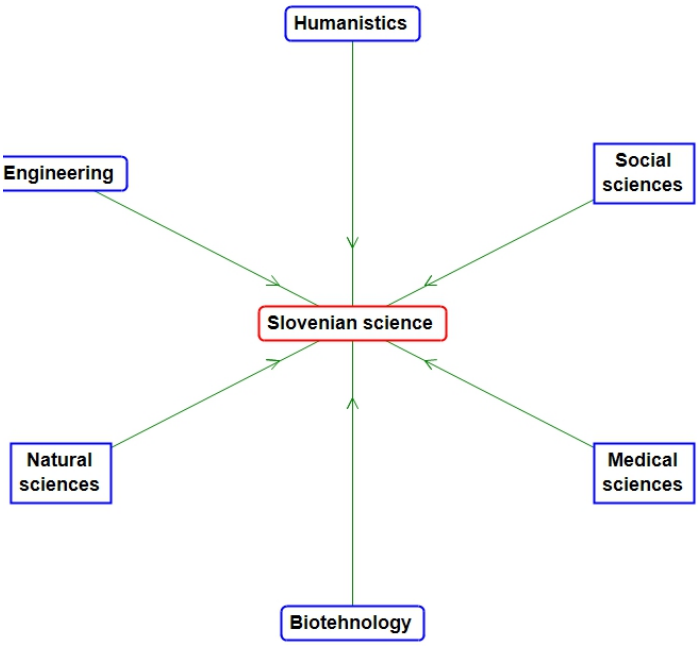


Figure 24 - Topic ontology build on 2250 projects with description in English language

5.2.1.2 Topic Ontology Evolution

The topic ontology was created on the complete set of 2250 projects, which include projects from year 1994 to year 2010 written in English language. The subtopics contain the following number of projects: medical sciences – 542 projects, natural sciences – 832 projects, engineering – 544 projects, humanistic - 308, social sciences - 553 and biotechnology – 181 projects. The sum of numbers of projects in each year is bigger than 2250, because some projects are belonging to more subtopics. To

analyze the ontology evolution, the number of projects in each subtopic, for the every year in the time span from 1994 to 2010, has to be examined.

Figure 2 shows number of new projects initiated in each year from 1994 to 2010, for every subtopic created with Ontogen (the figures were generated with Many Eyes tool (IBM Research and the IBM Cognos software group)). The graph in Figure 25 has six layers in different color, each layer representing one subtopic. Going from the upper layer direction down, the red layer represents medical sciences, purple represents biotechnologies, blue layer represents natural sciences, aquamarine layer represents engineering, green layer represents humanistic and the lowest yellow layer represents social sciences. Graph in Figure 25 evidently has four peaks and five pits. The first pit is in the beginning of the time interval, from year 1994 to 1995 in that period database contained small number of projects with English description. The number of initiated projects in each year grows, with the peak in year 1997 (406 new projects), after when the number of new projects peer year keeps falling with pit in year 2000, in which only two new projects are initiated. After pit in 2000, in year 2001 is a next peak with 485 projects following by a new pit in 2002 with only 10 projects. The number of new projects peer year start again growing and in year 2004 appears new peak with 492 projects. Then the number of projects peer year falls until year 2006 in which only 14 new projects were initiated. Year 2007 is another peak with 377 projects after which the numbers of new projects peer year falls with a pit in 2009 with no projects at all. Finally, in year 2010 there were 138 new projects initiated.

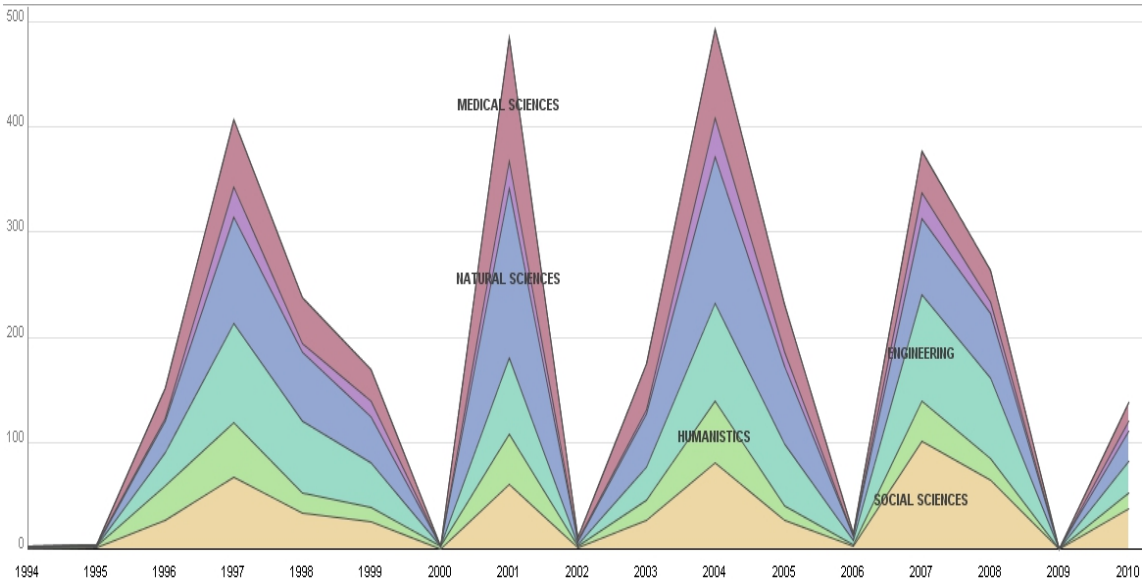


Figure 25 - Number of projects per subtopic initiated in each year from 1994 to 2010

From illustrated dynamics of new projects emergence, a pattern can be clearly recognized. The pattern reveals that new projects emerge in 2 to 5 years intervals between which there is a year with very small number of projects, or no projects at all. Each interval has its peak with the largest number of new projects somewhere in the middle of the interval, i.e. in the first or second year of the interval. The number of new projects peer year always grows from the beginning of the interval to

the peak, and it always falls from the peak of the interval to the end. The value of the peak has average value 440.

Examining the number of new project according to subfields to which they belong, it can be noticed that the proportion of the project for a subfield is always similar. This proportion can is clearer with Figure 26 which shows the proportion of projects peer each subfield. In the years 1994, 1995, 2000 and 2009 the layers become distorted because the number of new projects is very low, so a subfield takes large proportion for that year even thou it has very small number of projects. If these years containing "pits" are ignored, the proportion for each subfield seems consistent, with similar proportion of new projects thru time.

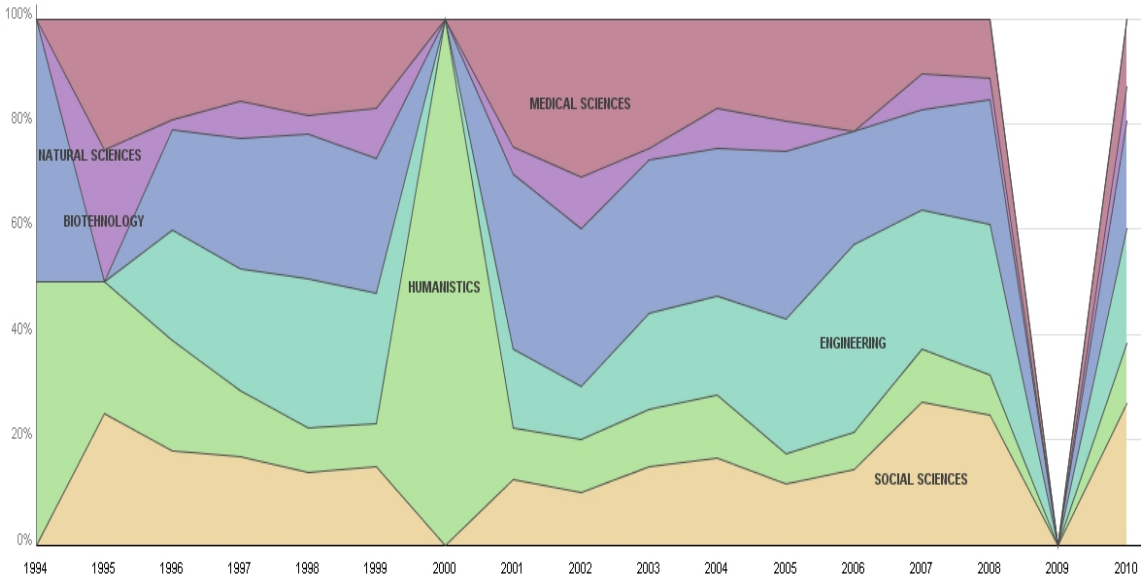


Figure 26 - Proportion of projects per subtopic initiated in each year from 1994 to 2010

Figure 27 shows the cumulative number of new projects peer year. Previously identified pits are on this figure areas of the graph where there is no growth. Since in years 1995, 2000, 2002, 2006 and 2009 there were no or very little new projects, the part of the line which leads to these years from one previous year is flat or almost flat. The areas on graph which have the steepest slope are the years with the biggest number of new projects initiated.

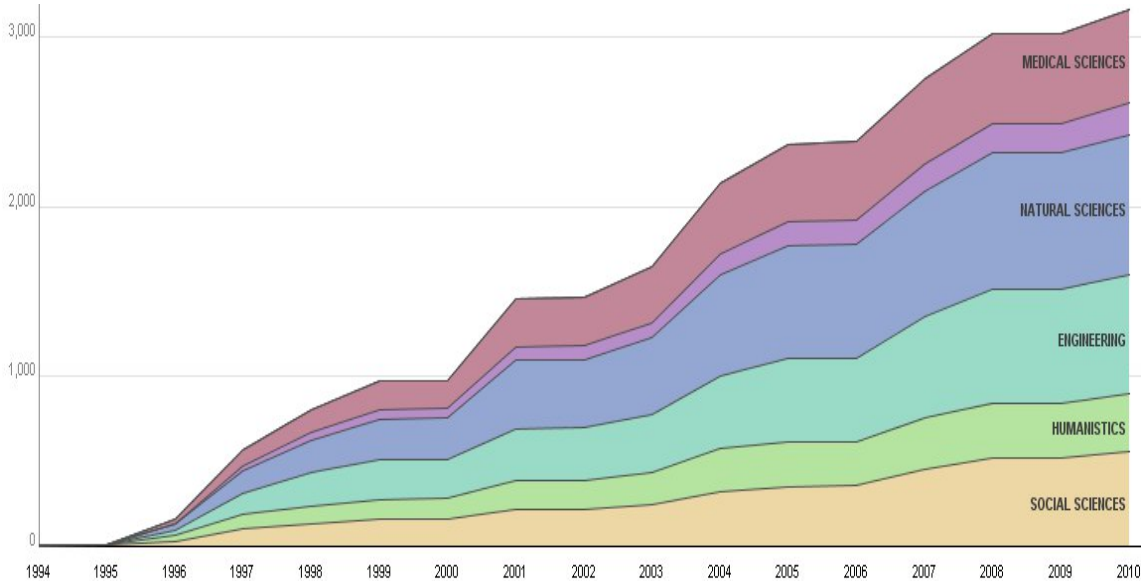


Figure 27 - Cumulative number of projects per subtopic initiated in each year from 1994 to 2010

The cumulative growth of projects for each subfield thru years is proportional; this can be observed on Figure 28, which shows the proportion of cumulative number of projects in each subfield. The area of graph on Figure 5 before year 1996 is not proportional, but the reason for this is the small cumulative number of projects before that year. The Figure 5 reveals that the natural sciences contained the biggest proportion of projects thru time, following by the engineering, medical sciences, social sciences and finally humanistic and biotechnical sciences with small least proportion of projects through the years.

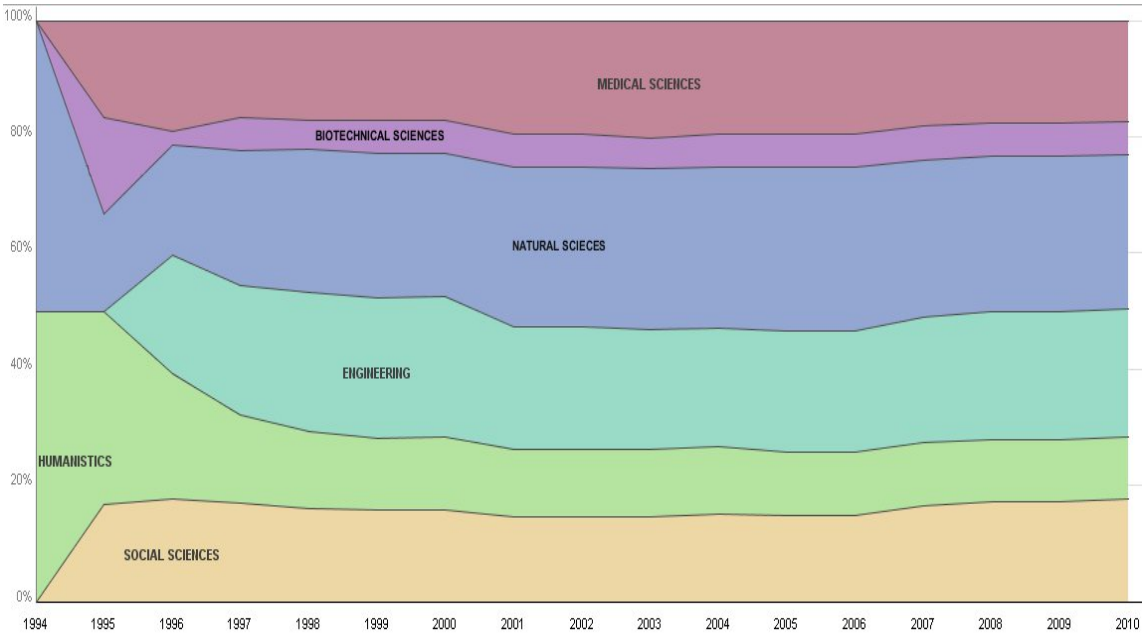


Figure 28 - Proportion of cumulative number of projects per subtopic initiated in each year from 1994 to 2010

5.2.2 Analyzing the Content of Project Snapshots

In this section the dataset containing 2250 projects with English descriptions is divided into 17 files, each for one year in time span from year 1994 to 2010. Each file contains the projects which were active in that year, meaning that the starting year of the project is smaller or equal of the year represented by the file and the ending year of the project is greater or equal to the year represented with the file. Figure 29 shows the number of active projects thru the years. It can be noticed that the highest numbers of active projects with English descriptions was in years 2004 (771 projects) and 2007 (749 projects). Figure 29 gives clear information about changes of number of active projects in time, but the information about content is missing. In order to discover dynamics of topics, the content of each snapshot was analyzed.

5.2.2.1 Dynamics of Content

Each snapshot was treated as a separate document containing descriptions the descriptions of all projects active in that year. The usual text preprocessing steps were made on all documents: removing punctuations, removing numbers, converting to lower case, removing white space, English stop words removal and stemming. Next, the most frequent word were identified for each snapshot these are the keywords that characterize active projects from each year. The most frequent words together with their frequencies are given in Table 6. The set of 28 most frequent words common for all documents were removed from the documents. These words are: active, analysi, applic, base contribut, culture, data, develop, effect, field, intern, knowledge, materi, method, model, process product, project, research, result, scientif, slovenia, slovenian, social, structur, studi, system,

technolog. Notice that stemming was applied on the text, thus some of the listed words differ from the surface form of the corresponding word.

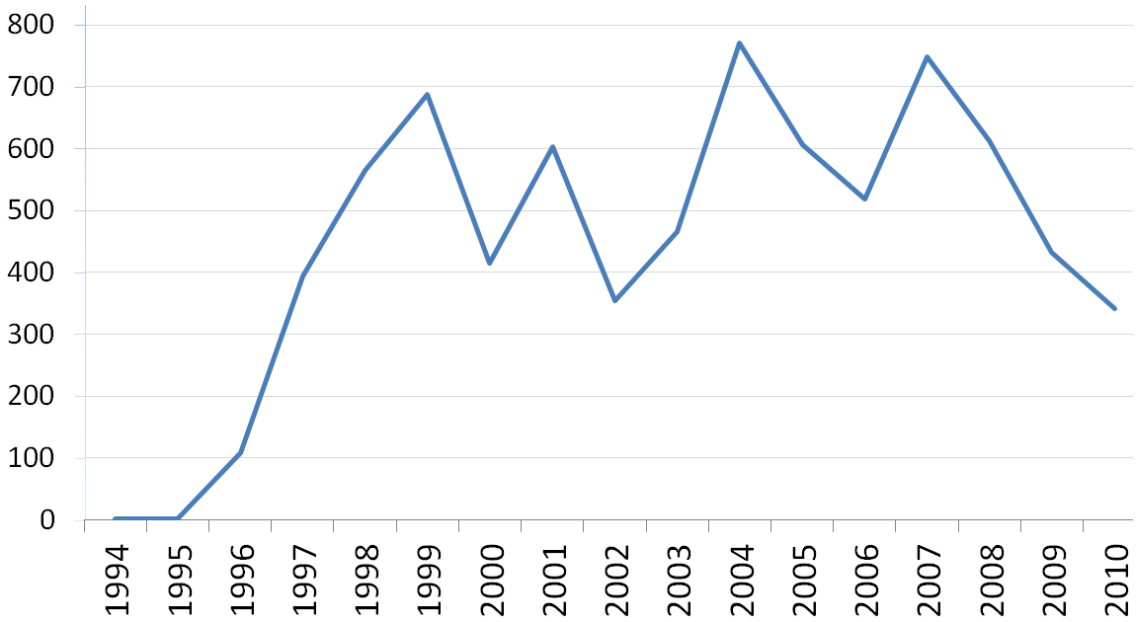


Figure 29 - Number of active projects with English descriptions thru years (1994-2010)

The content of fields in the Table 6 represents the keywords which describe the projects for every year. For every year few most frequent words were selected. Under the year label in is the minimum frequency of words in the field, e.g. >=90 in the field for the year 1999 means that the words describing that year have frequency 90 or more.

Table 6 - Keywords for each snapshot of projects for particular year. Under the year label is the number that shows the minimum frequency of the words in the fields

1994	**1995**	**1996**	**1997**	**1998**
>=3	*>=3*	*>=15*	*>=80*	*>=90*
alp, alpin, insect	alp, alpin, insect	corros, dynam, histori, karst, region, sloven	condit, determin, inform, investing, measure, relat, sloven	control, inform, investing, measure, relat, sloven
1999	**2000**	**2001**	**2002**	**2003**
>=90	*>=31*	*>=31*	*>=20*	*>=25*
condit, control, influenc, inform, investing, measure, relat, sloven	comput, dynam function, histori, measure, mechan, optim sloven, theori	dynam, magnet, metal, molecular, simul, treatment, water	diseas, factor, metal, plant, sloven, treatment, water	cancer, diseas genet, surface, treatment, water
2004	**2005**	**2006**	**2007**	**2008**
>=43	*>=35*	*>=32*	*>=210*	*>=340*
histori, property, sloven, surface, treatment, water	cancer, histori, measure, property, surfac water	comput, control, histori, measure, mechan, properti	enable,improve, increase, qualiti, sloven	enable, improve, increase, scienc, understand
2009	**2010**			
>=300	*>=240*			
enable, improv increase, scienc time, understand	enable, european, repres, scienc, time, understand			

To inspect the dynamics of topics in time in more detail, the keywords from Table 6 are added to graph with number of projects like in Figure 29. The combination of number of projects and their dynamic is shown on Figure 30.

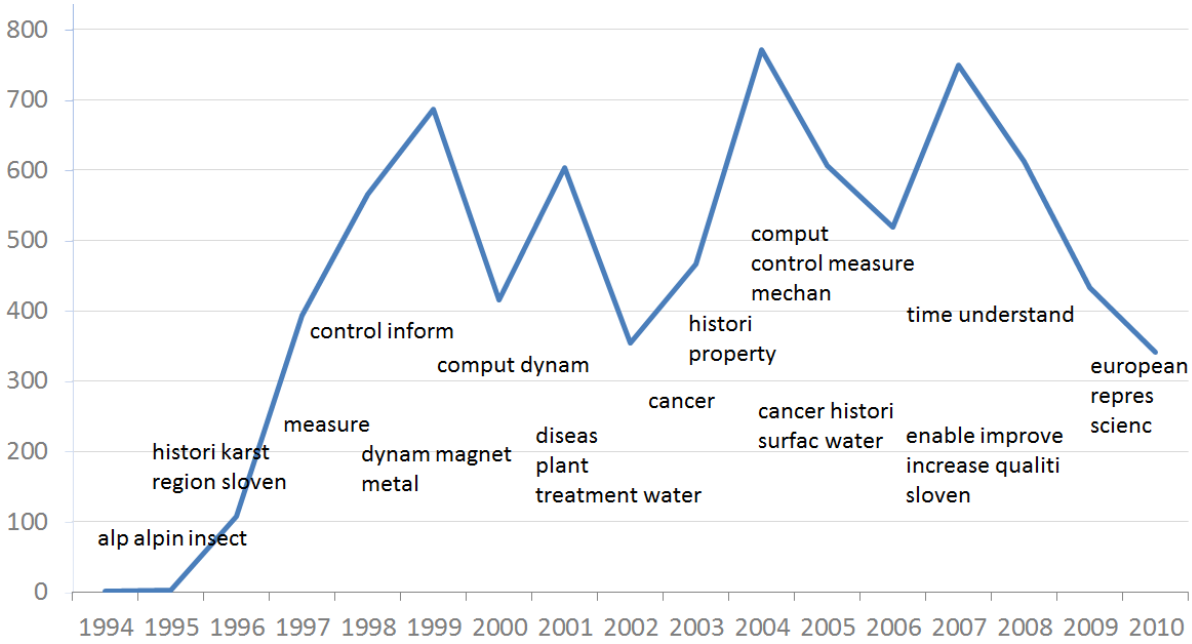


Figure 30 - Graph with the combination of project number and content dynamics

5.3 Conclusion

The dynamical aspect of the textual data is analyzed in two different approaches. First approach is analyzing the content of the complete dataset and breaking it down into different timeframes. Topic ontology was created and the temporal attributes of the documents that belong to topics were analyzed to get the insight into the evolution of topics in time. Second approach was separately analyzing different snapshots in time separately and connecting the results into a time flow to obtain the understanding of text evolution.

6 Combined modeling (temporal, text and social network analysis)

In this chapter a standard text analysis – classification, and a social network analysis task – vertex centrality measurement, were performed by combining the three types of analysis: temporal, text and social network analysis.

6.1 Classification

6.1.1 Data

The data used for this analysis was obtained from the Atlas of Slovenian science database. The Atlas of Slovenian science database contains the data about researchers, research projects, and research organizations, which were collected from the Slovenian Current Research Information System (SICRIS) in March 2011, using the SICRIS web service. Atlas of Slovenian Science contains 33519 researchers and 5384 research projects from the year 1994 to 2010.

For the purpose of this analyzes the data about researchers which had some keywords in English language was used. The number of such researchers was 8585. In order to create network data, research projects were used to connect the researchers with common projects. The constructed graph consisted of 8588 vertices representing the researchers and 58243 edges representing the common projects between researchers.

6.1.2 Classification with including attributes of neighboring nodes

The classification was performed using only the textual data assigned to the researchers – keywords written in English language, similarly as in chapter 3 - Modeling the Data with Text Analysis, the researchers were classified using the Text Garden utility. Before the classification, the standard text-mining procedures were applied on the dataset – stop words removal and stemming. The classification was performed using the SVM algorithm available with the Text Garden utility. The testing was done with the 5 fold cross validation method and the classification accuracy was 59.17%.

In order to use the network data to improve the classification, an approach of obtaining the attribute values of neighboring nodes was used. First, in the naïve way, every node was enriched with the keywords of all of its neighbors. Classification accuracy was slightly improved (for 1.76%) to 60.93%. Examination of results in more detail revealed that the 1164 researchers which were classified wrongly with the first approach, classified correct with the second approach. But at same time, the 1013 researchers which were classified correctly with the first approach were wrongly classified with the second approach. Since the number of samples with the improved classification was bigger than number of those degraded (for 151 samples), the classification accuracy was better, but the improvement was not sufficient.

Further improvement of classification was tackled by selecting the neighboring nodes from which the attribute values were obtained, based on the cosine similarity to them (table 1). First, each node was enriched with the keywords from neighboring nodes which had cosine similarity greater than 0. This improved the classification accuracy to 63.8905%. Next, each node obtained keywords of all neighboring nodes which have cosine similarity greater than 0.1. The classification accuracy with this approach was 63.7772%. The best results were obtained with obtaining the keywords from nodes with cosine similarity greater than 0.2, the classification accuracy was 66.888% in that case. Testing was performed also with threshold set to 0.4 (classification accuracy 62.749%), which did not give the best results, but still better from those when none or all neighbors of a node were used to obtain the additional keywords.

Table 7 - Classification accuracy with different criteria for acquiring attributes from neighboring nodes

Method	**Classification accuracy**
a) Original keywords	0.59173
b) Keywords from all neighbors	0.60939

c) Keywords from neighbors (cosine similarity > 0)	0.63890
d) Keywords from neighbors (cosine similarity > 0.1)	0.63777
e) Keywords from neighbors (cosine similarity > 0.2)	**0.66888**
f) Keywords from neighbors (cosine similarity > 0.4)	0.62749

There were 30814 edges (53% of all the edges in the graph) connecting researchers with different categories. The average similarity of these edges was 0.072123. The number of edges connecting researchers belonging to the same category was 27429 (47% off al edges) and the average similarity of these edges between there edges was 0.2023624. The reason why the results of classification were the best when acquiring attributes from neighbors with cosine similarity greater than 0.2, was because 0.2 is the closest value to the average similarity between nodes belonging to the same category. The field related to classification, network and temporal analysis that takes this kind of information to improve the classification is called relational learning. In the next chapter, some of the relational learning algorithms will be applied on our dataset.

6.1.3 Relational learning

Relational learning, also known as graph labeling, collective inference and collective classification, is an automatic classification of data items by exploiting the information about relationship between data items. It is useful to divide such systems into three components (Table 8). One component, the relational classifier, addresses the question: given a node and the node's neighborhood, how should a classification or a class-probability estimate be produced? For example, the relational classifier might combine local features and the labels of neighbors using a naive Bayes model (Chakrabarti, Dom, & Indyk, 1998) or a logistic regression (Lu & Getoor, 2003). A second component addresses the problem of collective inference: what should we do when a classification depends on a neighbor's classification, and vice versa? Finally, most such methods require initial ("prior") estimates of the values. A common estimation method is to employ a non-relational learner, using available "local" attributes. (Macskassy & Provost, 2007)

Table 8 - Graph labeling learning framework

1. Non-relational (local) model
2. Relational model
3. Collective inference

Viewing network classification approaches through this decomposition is useful for two main reasons. First, it provides a way of describing certain approaches that highlights the similarities and differences among them. Secondly, it expands the small set of existing methods to a design space of methods, since components can be mixed and matched in new ways. In fact, some novel combination may well perform better than those previously proposed; there has been little

systematic experimentation along these lines. Local and relational classifiers can be drawn from the vast space of classifiers introduced over the decades in machine learning, statistics, pattern recognition, etc., and treated in great detail elsewhere. In (Macskassy & Provost, 2007) authors present a large-scale, systematic experimental study of machine learning methods for within-network classification and offer a network learning toolkit (NetKit-SRL) that enables in-depth, component-wise studies of techniques for statistical relational learning classification with network data. Network Learning Toolkit is designed to accommodate the interchange of components and the introduction of new components. For our purpose two standard algorithms are applied with the toolkit and the results are presented in the table 9. Both algorithms significantly improve the performance of the classification compared to the naïve approach of including attributes of neighbors under different conditions (as shown in table 8).

Table 9 - Results of relational learning based classification methods

Network classification method	Classification accuracy
Network-only Link-Based classifier (nLB) (Lu & Getoor, 2003)	**0.785547348**
Weighted-vote relational neighbor (wvRN) procedure (Macskassy & Provost, A simple relational classifier, 2003)	0.762896

6.2 Centrality measure

The traditional degree centrality measure states that the actors who have more ties to other actors may be in advantaged position. Because they have many ties, they may have alternative ways to satisfy needs, and hence are less dependent on other individuals. Because they have many ties, they may have access to, and be able to call on more of the resources of the network as a whole. (Hanneman & Riddle, 2005).

6.2.1 Description of the dataset

The proposed centrality measure was tested on two networks: a network of projects and a network of researchers.

Network of researchers consists of 8585 researchers from the database of Slovenian National Research Agency from 1994 to 2010. The researchers which contained descriptive keywords in English language were chosen from the database. The network contained 58 243 edges, connecting researchers which worked on the same projects in the period from 1994 to 2010 as recorded in the database of Slovenian National Research Agency.

The nodes of the projects network are national research projects recorded in the database of Slovenian National Research Agency from 1994 to 2010. Two projects of the network are connected with an edge, if there is a same researcher working on both projects. The following attributes were assigned to the nodes: name of the project, starting year of the project, ending year of the project and description of the project. The description of the project included the title, abstract, keywords,

description of domestic and world significance of the project, written in English language. The network was constructed with the projects that contained at least titles of the projects in English language. The number of such projects, and the number of nodes in the constructed network was 2 250, and the number of edges connecting the nodes was 10 535.

In order to calculate the centrality with the proposed approach, first the similarity property had to be assigned to each edge. This property gives the information about how much are two nodes similar. In the case of projects network, the cosine similarity between the textual descriptions of two projects was calculated; and in the case of researchers network, the cosine similarity between keywords of researchers was calculated. When the two projects connected with an edge had identical descriptions, the value of similarity was 1. On the other hand, when the descriptions of the two projects had no words in common, the value of similarity was 0. For the two projects that had some common words, the similarity was between 0 and 1.The cosine similarity was calculated after performing the standard text-mining preprocessing operations - stemming and stop-words removal. After each edge had a similarity property assigned to it, the centrality of each node was calculated using the proposed equation.

6.2.2 Related work

Freeman (Freeman L. , 1979) gives a review of centrality measures and the conceptual clarification of them. According to him, Shaw (Shawn, 1954) introduced the idea of using vertex degree as an index of point centrality. Shawn and other authors (Mackenzie, Czepiel, Niemien, Rogers) equaled vertex degree with centrality. The conception that centrality is some function of the degree of a point is simplest and perhaps the most intuitively obvious, according to Freeman. His conceptual explanation with respect to social network is that a person who is in a position that permits direct contact with many others should begin to see himself and be seen by those others as a major channel of information. Our approach includes the vertex degree as a main part of the centrality measure and ads the consideration of content if the connected nodes into the centrality function. This results with conceptual explanation which extends the one with vertex degree by noting the importance of a person in a position that permits direct contact with many others that have different information.

In (Opsahl, Agneessens, & Skvoretz, 2010) authors propose generalization that combines both the number of ties (or vertex degree which is the central component of the original measures) and the tie weights, for the three common node centrality measures: degree, closeness, and betweenness. The proposed degree centrality measure is the product of the number of nodes that a focal node is connected to, and the average weight to these nodes adjusted by the tuning parameter, which can make the high degree of a vertex favorable (value of the parameter between 0 and 1), or the low degree of a vertex favorable (value of the parameter above 1). Our work is in the lines of the research of (Opsahl, Agneessens, & Skvoretz, 2010) in the sense of incorporating additional value, which is easily tunable, so it can capture the real world settings in which it is important for the measure of centrality. Since there are similarities in the research methodology, the comparison and the combination of effects of our proposed measure and the measure from (Opsahl, Agneessens, & Skvoretz, 2010) will be shown on the Freeman's EIES dataset (Freeman & Freeman, 1979).

6.2.3 Approach

We propose a network centrality measure that takes into account the content of the nodes in the network. Our approach is based on the reasoning that not only the number of ties to other actors is important for the centrality, but also the properties of the actors which are connected with the ties. The hypothesis is that the ties to the diverse actors contribute more to the centrality than the ties to the similar actors. The explanation would be that the ties to the similar actors do not give as much opportunities to get new resources as the ties to the diverse actors. If the actors are described with text, the similarity of two actors connected with a tie can be expressed with one of the measures for comparing the similarity of textual documents, like for example cosine similarity. We propose the following equation for measuring the centrality of an actor, by taking into account the vertex degree of the actor and the cosine similarity with connected actors:

$$CM(a) = \begin{cases} VD(a)^{1+x-\left(\frac{\sum_{i=1}^{VD} similarity(a,b_i)}{VD}\right)\cdot x}, VD > 1 \\ VD(a) + \frac{\sum_{i=1}^{VD}(1 - similarity(a,b_i))}{VD} \cdot x, VD = 1 \end{cases}$$

Where VD is the vertex degree of the node a. $similarity(a, b_i)$ is the similarity between the textual content of node a and the node b_i which is connected with an edge to the node a. x is the parameter which enables users to set the relative importance between the number of edges and the similarity between connected nodes. Parameter x can be set to any positive real number. If x is set to 0, the proposed measure is identical to the vertex degree, because no importance is given to the similarity of the neighboring nodes. If the parameter x is set to 1, the value of proposed centrality measure can be in the range from vertex degree to the vertex degree squared, depending on the similarity. The greater values for parameter x will further expand the range of the possible values of the measure. The influence of different value of parameter x is illustrated on Figure 2. The value of proposed measure is plotted against 5 different values of vertex degree (from 1 to 5). For each vertex degree, 11 different similarity values are plotted (0, 0.1, 0.2, …, 1), to show the range of possible values of the similarity measure. For each degree the point with similarity 1 has value equal to vertex degree, while the point with similarity 0 has maximal possible value. On the Figure 2.a the parameter x has value 0.2 and in that case the influence of similarity between nodes is not so strong, because the values cannot vary much from the vertex degree. For example, the node with vertex degree 5 can obtain value of centrality measure at most round 6.9 because of low similarity with the neighboring nodes. On the Figure 2.b the parameter x is set to 0.5, making the influence of similarity between nodes more important. On Figure 2.c, parameter x is 1, and every node can obtain value of centrality measure between value of vertex degree and vertex degree squared. On Figure 2.d x is set to 2, making the influence of similarity between nodes very strong. For example a node with vertex degree 5 can obtain value of centrality measure between 5 and 125 depending on the similarity with its neighboring nodes.

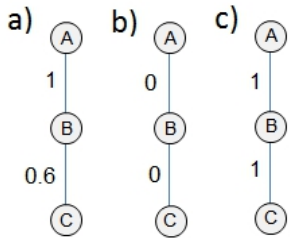


Figure 31 – Networks a, b and c with different similarities between nodes.

For example let's suppose there is a node connected with two other nodes (node B on Figures 1.a, 1.b and 1.c), with the edges with similarity property 1 and 0.6 (Figure 1.a). The vertex degree of this node is 2, but considering that one of the nodes connected to it is somewhat different, the proposed centrality measure with x=1 will be: $2^{2-(1.6/2)} = 2^{1.2} \approx 2.297$. If the two nodes would be completely different (similarity equal to 0 - Figure 1.b), the proposed measure would increase quadratically versus the vertex degree: $2^{2-(0/2)} = 2^2 = 4$. In the case the two nodes would be identical (similarity equal to 1 - Figure 1.c), the measure would stay equal to the vertex degree: $2^{2-(2/2)} = 2$.

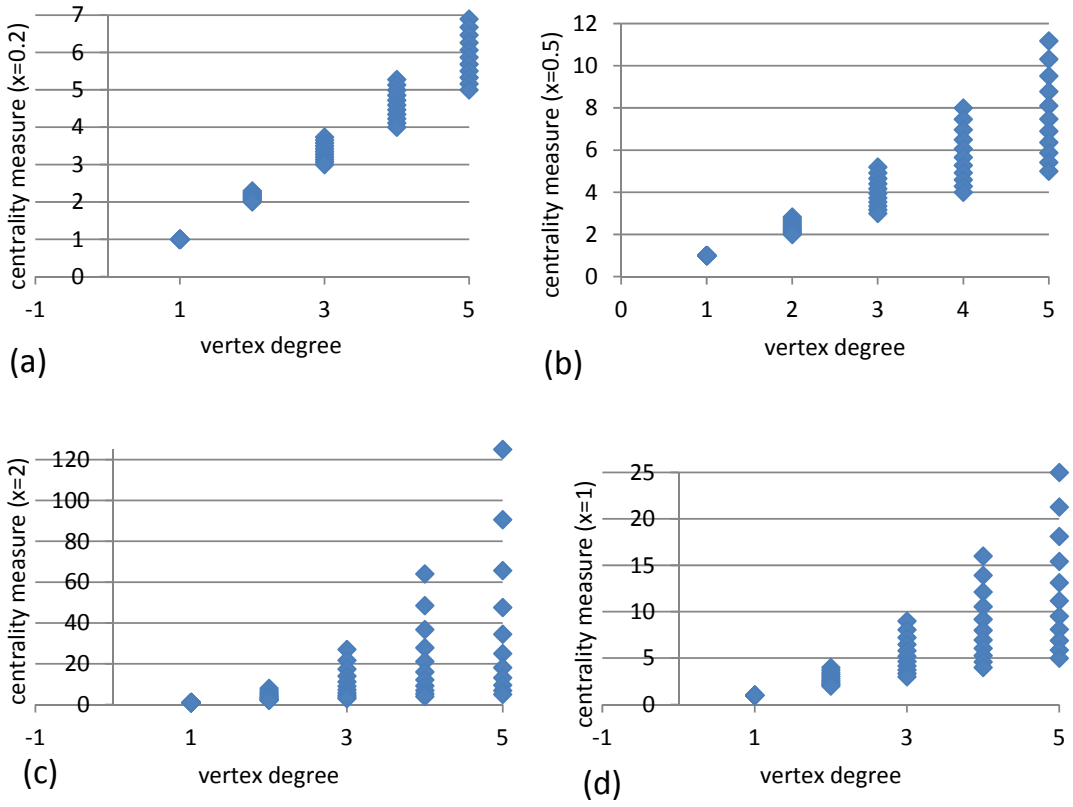


Figure 32 – Range of possible centrality measure values for vertex degrees from 1 to 5, with four different values of parameter x: 0.2 (a), 0.5 (b), 1 (c) and 2 (d).

6.2.4 Illustrative example

To illustrate the benefit of considering the content of the connected nodes when measuring centrality, a simple network with 10 nodes and 9 edges is created and drawn on Figure 2. Let's suppose the nodes of the network represent the researchers and the edges represent the collaboration of researchers on projects. The nodes of the network are described with keywords, which represent the expertise of the researchers. The nodes A and E are described with "linguistics", nodes B, C, D and H with machine learning; nodes F and I with "graph theory"; and nodes G and J with "social science". The four mentioned keywords (linguistics, machine learning, graph theory and social sciences) come from four different fields of science: humanistic, technical sciences, natural sciences and social sciences, but combining these fields can have synergy effect. Since it is clear that it is beneficial for researcher to have not only many connections to other researchers, but also connections to researchers with different competences, the proposed centrality measure is used to rank the researchers of the network.

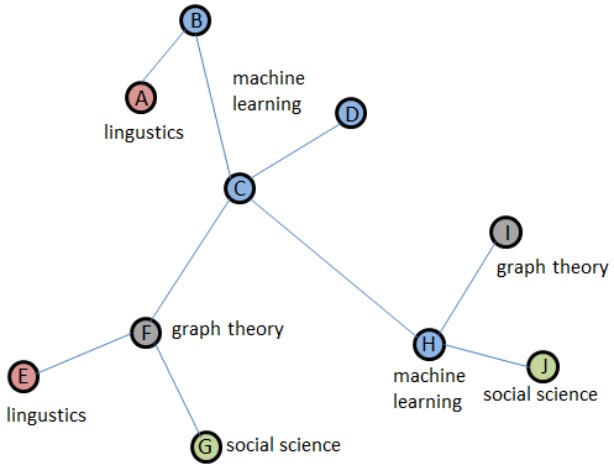


Figure 33. Illustrative network of researchers described with different keywords

Table 1 shows the rankings of the nodes from the illustrative network. In first column of the table the nodes are ranked according to the vertex degree, in columns 2 to 5 the nodes are ranked using the proposed centrality measure with different values of x (greater value of parameter x gives bigger importance to diversity of collaboration).

Table 10.Rankings of the nodes from the illustrative network of researchers

Vertex degree	x=0.2	x=0.5	x=1	x=1.5
C (4)	C (4.29)	F (5.20)	F (9.00)	F (15.59)
F (3)	F (3.74)	**C (4.76)**	H (6.24)	H (9.00)
H (3)	H (3.47)	H (4.33)	**C (5.66)**	C (6.73)
B (2)	B (2.14)	B (2.38)	B (2.83)	B (3.36)
A (1)	A (1.20)	A (1.50)	A (2.00)	A (2.50)
D (1)	E (1.20)	E (1.50)	E (2.00)	E (2.50)

E (1)	G (1.20)	G (1.50)	G (2.00)	G (2.50)
G (1)	I (1.20)	I (1.50)	I (2.00)	I (2.50)
I (1)	J (1.20)	J (1.50)	J (2.00)	J (2.50)
J (1)	**D (1.00)**	D (1.00)	D (1.00)	D (1.00)

As shown with the Table 8, using only the vertex degree the node C is ranked first because it has 4 connections with other nodes. Using the proposed measure and very low value of parameter x (0.2), node D which is connected to only one researchers with the same expertise as his (machine learning) is ranked lower than nodes A, E, G, I and J. Setting the value of parameter x to 0.5 rank the node F higher than node C, because even thou node F has less connections than node C, node F is connected to nodes with more diverse expertise. With value of parameter x equal to 1, node H which has 2 connections less than node C, is ranked higher than node C, because of diverse collaborators. With the further increase of parameter x the ranks stay the same.

6.2.5 Experimental testing

The proposed centrality measure was tested on two networks: a network of projects and a network of researchers.

For the network of researcher, the traditional degree centrality is telling us, with how many other researchers each researcher collaborated on projects. The ones with higher centrality can in greater extend gain and/or give knowledge to others, through the activity of collaborating on project. With our approach, the competences of each researcher are taken into account. The proposed measure ranks higher those researchers which, not only collaborated with many different researchers, but which collaborated with researchers with competences different to his. Sharing of knowledge and skills has bigger potential between those who have diverse competences. The competences of the researchers were captured with the keywords whit which each researcher was described, while cosine similarity between bag-of-words of researcher's descriptions was used to define how much competences between two researchers differ. It is clear to us that the keywords do not perfectly describe the skills and knowledge of researchers, but at least it is some descriptions which are sufficient for testing our centrality measure which takes the dimension of node content into the account. In order to improve the descriptions with keywords, they were extended with synonym terms using the Wordnet.

For the network of projects, in addition to ranking higher those projects which are connected with many others, the proposed centrality measure enables us to increase the rank when the connected projects have diverse competences. The connection between projects exists if they have a common researcher working on it. The competences of projects were captured with the attributes: title, abstract, keywords, domestic and word significance of the projects. The size of these descriptions vary for different projects; some contain few paragraphs of abstract or world and domestic significance with keywords, while others contain only the titles. The central project is the one connected with many other projects from which many different parts cold be inherited and combined. The more central the project, the bigger is the potential impact of it.

6.2.6 Results

Tables 10 and 11 show the results of applying proposed centrality measure on the network of researchers and the network of projects respectively.

Table 9 shows the 20 researchers from the network of researchers. The selection was made according to the vertex degree, choosing the top 20 with the highest value of vertex degree. The first column of the table contains the names of the researchers; second column contains the values of vertex degree; and the third column the rank according to the vertex degree. The fourth column "distinct fields" indicates to how many distinct fields of science are classified the connected researchers - since this information was not used for the proposed centrality measure, it can used for the evaluation. The columns: 5, 6, 7 and 8 show the values of proposed centrality measure with parameter x set to: 0.2, 0.5, 1 and 2. The tables shows how the ranks of the first 20 researchers determined with vertex degree, change with applying the proposed measure with the different values of parameter x. "Gregor" from the first place with vertex degree 141 stays on the first position with all four variations of the proposed measure. This indicated that his collaboration is diverse, what can be confirmed with the high number of distinct fields of science of his collaborators (24). "Milan", "Franc" and "Peter" were all ranked second because they have the same vertex degree (140). The proposed measure with x set to 0.2, ranks "Franc" 2nd, "Milan" 3rd and "Peter" 4th, this ranking fits the number of distinct fields of these researchers (24, 22 and 15). With the higher value of x, meaning that the diversity of collaboration has bigger influence on ranking, the relative ranking of those researchers stays the same, but "Milan" and "Peter" are ranked much lover relative to others, while Franc stays on 2nd position, what means that his collaboration is very divers. The number of distinct science fields of his collaborators points in the right direction, but it does not capture the difference in collaboration diversity between Franc and Milan as accurately as measuring the cosine similarity between their textual descriptions.

Table 11. First 20 researchers from the projects network, ranked by the vertex degree, and the ranks with proposed measure

Name of the Researcher	Vertex degree	Vertex degree rank	Rank with similarities x=0.2	Rank with similarities x=0.5	Rank with similarities x=1	Rank with similarities x=2	Distinct fields
Gregor	141	1	1	1	1	1	24
Milan	140	2	3	4	6	7	22
Franc	140	2	2	2	2	2	24
Peter	140	2	4	5	8	8	15
Sonja	133	3	6	6	5	5	22
Radojko	133	3	5	3	3	3	29
Primož	132	4	7	8	7	6	18
Milena	130	5	9	10	12	13	31
Polona	125	6	10	9	9	10	15
Hojka	123	7	11	12	14	18	22
Tomislav	123	7	8	7	4	4	23
Dragomir	120	8	12	16	20	26	15
Vekoslava	115	9	14	15	15	17	23
Branko	112	10	13	11	10	11	15
Ana	110	11	17	17	16	15	19
Marjeta	109	12	18	18	18	21	13
Miran	108	13	16	14	13	12	25
Polonca	108	13	15	13	11	9	28
Antonija	104	14	21	25	29	34	15
Cvetka	104	14	19	20	19	20	17

Looking at the Table 9 in general, it can be noticed that the changes in the ranking occur in both directions and that the change is stronger in consistent direction, as the value of x increases. Also, the conflicts in the rank because of the duplicate values are resolved. Since the table 1 shows only the small portion of the data from researchers network, the influence of the measure on the complete dataset will be discussed in the last part of these chapter, after analysis of table with data from projects network.

Table 12. First 20 projects from the projects network, ranked by the vertex degree, and the ranks with proposed measure

Name of the Project	Vertex degree	Vertex degree rank	Rank with similarities x=0.2	Rank with similarities x=0.5	Rank with similarities x=1	Rank with similarities x=2	Distinct fields
Stress response	71	1	1	2	2	2	13
Joint effects	69	2	2	1	1	1	12
Lipid peroxidation	57	3	3	3	3	4	12
The use of new	53	4	4	4	4	3	11
Nitrate migration	52	5	6	6	7	7	6
CO2 fixation	52	5	5	5	5	6	10
Clinical and genetic	49	6	7	8	8	8	8
Optimization of MDP	49	6	13	15	18	21	5
Boron Neutron	48	7	8	7	6	5	13
Prion diseases	48	7	9	10	10	11	12
Development of tools	47	8	11	11	12	12	14
Genetic causes of	47	8	10	9	9	9	1
Biological methods	47	8	12	13	13	13	13
Geochemical comp	47	8	15	16	17	18	11
Elaboration and	45	9	16	14	14	14	13
Pathways of carbon	45	9	14	12	11	10	10
Developments of basic	43	10	18	19	20	20	9
Research of degragatio	42	11	17	17	15	15	11
Biosurgery of chronical	42	11	20	21	22	23	1
Prenatal screening	41	12	22	22	24	25	9

Table 11 has the same structure as Table 10, except that it shows 20 top ranked projects with the highest vertex degree. The table points to similar conclusions as the previous table – i.e.: “Stress response” is ranked 1st with vertex degree only, stays first with the proposed measure which takes the content of collaboration into account when the value of x is low (0.2), but drops to second place when bigger importance is given to collaboration diversity.

Even thou the tables 10 and 11 show the influence of the proposed measure only on a small part of the top rated researchers and projects, the similar behavior is noticed on the rest of the data.

6.2.7 Experimental testing on the Freeman’s EIES network

Freeman’s EIES dataset arose from an early experiment on computer mediated communication. Fifty academics interested in interdisciplinary research were allowed to contact each other via an Electronic Information Exchange System (EIES). The data collected consisted of all messages sent plus

acquaintance relationships at two time periods (collected via a questionnaire). The data includes the 32 actors who completed the study (Williams, 2011). The dataset was used in many different studies ((Freeman & Freeman, 1979); (Wasserman & Faust, 1994); (Opsahl, Agneessens, & Skvoretz, 2010); (Opsahl & Panzarasa, Clustering in weighted networks, 2009)). In (Opsahl, Agneessens, & Skvoretz, Node centrality in weighted networks: Generalizing degree and shortest paths, 2010) authors used the dataset to demonstrate the influence of their proposed degree centrality measure, which incorporates vertex degree and weights of the edges, on ranking of different scientists. Here, we will show how our proposed measure influences the ranking on the same dataset and how can our approach of incorporating the content of the nodes, be combined with the generalized degree centrality measure. Since no textual descriptions of the scientists useful for our approach was available in the dataset, the affiliation to one of the four disciplines was used: 1 = Sociology, 2 = Anthropology, 3 = Mathematics/Statistics and 4 = other. This resulted with simplified approach without the need for calculating cosine similarity. Like in the illustrative example from chapter 3, if the scientists belong to the same discipline - the similarity is 1, whereas if they do not – the similarity is 0.

Table 13 - Compariosn and combination of our approach with the generalized centrality measure on the standard Freeman's EIES dataset

Degree $(\alpha = 0; x = 0)$	Degree and weight $(\alpha = 0.5; x = 0)$	Degree and similarity $(\alpha = 0; x = 0.5)$	Degree, weight and similarity $(\alpha = 0.5; x = 0.5)$
Lin Freeman (31)	Lin Freeman (314)	Lin Freeman (75)	Lin Freeman (1382)
Sue Freeman (31)	Barry Wellman (249)	Nick Mullins (75)	Barry Wellman (987)
Nick Mullins (31)	Russ Bernard (200)	Barry Wellman (64)	Nick Mullins (511)
Phipps Arabie (28)	Sue Freeman (180)	Ron Burt (49)	Pat Doreian (342)
Barry Wellman (28)	Doug White (177)	Sue Freeman (41)	Ron Burt (322)
Doug White (28)	Nick Mullins (142)	Phipps Arabie (38)	Russ Bernard (305)
Russ Bernard (25)	Lee Sailer (134)	Pat Doreian (36)	Sue Freeman (273)
Ron Burt (20)	Pat Doreian (129)	Richard Alba (34)	Lee Sailer (239)
Pat Doreian (20)	Ron Burt (85)	Doug White (33)	Doug White (234)
Richard Alba (18)	Richard Alba (77)	Russ Bernard (32)	Richard Alba (203)
Jack Hunter (18)	Steve Seidman (70)	Lee Sailer (24)	Davor Jedlicka (129)
Lee Sailer (17)	Phipps Arabie (66)	Jack Hunter (23)	Phipps Arabie (96)
Steve Seidman (16)	Jack Hunter (65)	Davor Jedlicka (21)	Jack Hunter (92)
Carol Barner-Barry (15)	Al Wolfe (63)	Carol Barner-Barry (18)	Maureen Hallinan (91)
Al Wolfe (14)	Carol Barner-Barry (60)	Al Wolfe (17)	Don Ploch (90)
Paul Holland (12)	Paul Holland (49)	Steve Seidman (16)	Al Wolfe (85)
John Boyd (11)	John Boyd (45)	John Boyd (15)	Carol Barner-Barry (79)
Davor Jedlicka (11)	Davor Jedlicka (45)	Don Ploch (14)	John Boyd (77)
Charles Kadushin (7)	Maureen Hallinan (37)	Paul Holland (13)	Steve Seidman (70)
Nan Lin (7)	Don Ploch (28)	Charles Kadushin (12)	Paul Holland (57)

Don Ploch (7)	Claude Fischer (22)	Nan Lin (11)	Charles Kadushin (51)
Claude Fischer (6)	Mark Granovetter (22)	Claude Fischer (9)	Claude Fischer (48)
Mark Granovetter (6)	Charles Kadushin (21)	Maureen Hallinan (9)	Nick Poushinsky (44)
Maureen Hallinan (6)	Nan Lin (20)	Nick Poushinsky (8)	Nan Lin (38)
Nick Poushinsky (5)	Nick Poushinsky (18)	Mark Granovetter (7)	Mark Granovetter (29)
Sam Leinhardt (4)	Joel Levine (17)	Sam Leinhardt (6)	Joel Levine (24)
Joel Levine (4)	Sam Leinhardt (10)	Joel Levine (5)	Sam Leinhardt (17)
John Sonquist (4)	John Sonquist (9)	John Sonquist (4)	John Sonquist (9)
Brian Foster (3)	Brian Foster (7)	Gary Coombs (4)	Brian Foster (9)
Ev Rogers (2)	Gary Coombs (6)	Brian Foster (4)	Gary Coombs (9)
Gary Coombs (3)	Ed Laumann (6)	Ev Rogers (2)	Ed Laumann (6)
Ed Laumann (2)	Ev Rogers (4)	Ed Laumann (2)	Ev Rogers (6)

The first column of the Table 12, is the ranking of the researchers according to vertex degree only. The second column of the table shows replicated results from (Opsahl, Agneessens, & Skvoretz, Node centrality in weighted networks: Generalizing degree and shortest paths, 2010), where tuning parameter was $\alpha = 0.5$, meaning that the number of different connections increases the value of the measure. The third column of the table shows ranking when applying only our approach with the value of tuning parameter set to 1, while the fourth column shows the ranking when applying the same measure on the values of the second column, i.e. on generalized node centrality measure.

6.3 Conclusion

In this final chapter of the report, the combinations of the analysis methods were combined in order to model the data. In the first part, the classification of nodes into science groups and science fields was performed. First, the textual information of each node was under different criteria combined with attributes of neighboring nodes. Next, two different relational learning algorithms were applied for classifying the nodes. In the second part, textual and network structure information were combined for a centrality measure that takes into account the content of the neighboring nodes. The proposed centrality measure was applied on the network of researchers and the network of projects and the top ranked researchers and projects were listed. The proposed measure was also compared and combined with a generalized degree centrality measure on the standard Freeman's EIES dataset.

7 Conclusion

This report consist of five parts addressing modeling of data using social network analysis methods and using text analysis methods, modeling the evolution of data using social network analysis and using text analysis and, a combination of social network and text analysis methods.

Using the network analysis methods we have measured cohesion and brokerage of the researcher network. The entire network is sparse with density measure of 0.025658%, while the average vertex degree of entire network is 8.5, meaning that the average researcher collaborates with 8 other researchers on projects. Biotechnical sciences turned out to be the most cohesive sub network with density of 0.0105 and average vertex degree of 18.4. Examining the collaboration between science

groups, it is found that Engineering sciences and technologies and Natural sciences mathematics and have the most central position, being connected with strong edges to 5 and 4 other science groups respectively. In the last part of this chapter, some small cohesive subgroups from the core of the network were identified. To measure brokerage centrality and centralization measures were applied to entire network and to sub networks according to science groups.

We have described tree aspects of dealing with textual corpus: preprocessing and exploring the properties of text, clustering and developing data driven topic ontologies and, classification of textual documents. It was shown that the corpus can be validated against some empirical laws, like Zipf's and Heap's law. Topic ontology based on the textual corpus of research project was build and in the last part, classification of textual documents is described.

Modeling the evolution of data using social network analysis and using text analysis included the analyses from the aspect of network size, density of the network, network diameter and connected component of the network. Also, the dynamical aspect of the textual data was analyzed in two different approaches: analyzing the content of the complete dataset and breaking it down into different timeframes. and separately analyzing different snapshots in time separately and connecting the results into a time flow to obtain the understanding of text evolution.

In the final chapter, the combinations of the analysis methods were combined in order to model the data. The classification of nodes into science groups and science fields was performed by including the textual information of neighboring nodes and by applying standard relational learning algorithms. Finally a centrality measure that takes into account the structural properties of the network as well as the content of the network nodes, was proposed and applied on the researchers and projects network, and tested on the standard dataset from the social network analysis research community.

Bibliography

Artificial Intelligence Laboratory - Institute Jozef Stefan. (n.d.). *Artificial Intelligence Laboratory*. Retrieved 11 15, 2011, from Artificial Intelligence Laboratory: http://ailab.ijs.si/tools/text-garden/

Batagelj, V., & Mrvar, A. (2012, 9 26). *Pajek - Program for Large Network Analysis*. Retrieved 03 06, 2012, from http://pajek.imfm.si/doku.php?id=pajek

Chakrabarti, S., Dom, B., & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, (pp. 307–319).

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*, 273-297.

Croft, B. W., Metzler, D., & Trevor, S. (2010). *Search Engines Information Retrieval in Practice.* Boston: Pearson Education.

David, D. L. (1998). Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. *Proceedings of ECML-98, 10th European Conference on Machine Learning.* Chemnitz.

Feinerer, I. (2011). *Introduction to the tm Package Text Mining in R.*

Feinerer, I. (2012). *tm: Text Mining Package. R package version 0.5-7.1.*

Feinerer, I. (n.d.). *R Documentation*. Retrieved 03 06, 2012, from Explore Corpus Term Frequency Characteristics: http://127.0.0.1:16389/library/tm/html/Zipf_plot.html

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software, 25*(5).

Fellows, I. (2012, 2 15). *Package 'wordcloud'.* Retrieved 2 28, 2012, from http://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf

Fortuna, B., Grobelnik, M., & Mladenic, D. (2005). Visualization of text document corpus. *Informatica, 29*(4), 497-502.

Fortuna, B., Grobelnik, M., & Mladenic, D. (2007). OntoGen: Semi-automatic Ontology Editor. *12th International Conference on Human-Computer Interaction* (pp. 309-318). Beijing: Springer-Verlag.

Freeman, L. (1979). Centrality in Social Networks: Conceptual Clarification. *Social Networks*, 215-239.

Freeman, S. C., & Freeman, L. C. (1979). *The networkers network: A study of the impact of a new communications medium on sociometric structure. Social Science Research Reports No 46. .* Irvine CA,: University of California.

Gentleman, Robert; Ihaka, Ross; et al. (n.d.). *R-project*. Retrieved 2 27, 2012, from http://www.r-project.org/

Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods.* Riverside, CA: University of California, Riverside (published in digital form at http://faculty.ucr.edu/~hanneman/).

IBM Research and the IBM Cognos software group. (n.d.). *Many Eyes*. Retrieved 03 10, 2012, from http://www-958.ibm.com/software/data/cognos/manyeyes/

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of ECML-98, 10th European Conference on Machine Learning.* Chemnitz.

Jones, K. S. (1972). A statistical interpretation of term specificity. *Journal of Documentation, 28*(1), 11-21.

Leskovec, J. (n.d.). *SNAP*. Retrieved 3 7, 2012, from SNAP network analysis library: http://snap.stanford.edu/index.html

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*.

Lu, Q., & Getoor, L. (2003). Link-based classification. *In Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, (pp. 496–503).

MacKenzie, K. D. (1964). *A Mathematical Theory of Organizational Structure.* University of California: Berkeley.

Macskassy, S. A., & Provost, F. (2003). A simple relational classifier. *Multi-Relational Data MiningWorkshop (MRDM) at the Ninth ACM SIGKDD International Conferenceon Knowledge Discovery and Data Mining.*

Macskassy, S. A., & Provost, F. (2007). Classification in Networked Data: A Toolkit and a Univariate Case Study. *Journal of Machine Learning*, 935-983.

Opsahl, T., & Panzarasa, P. (2009). Clustering in weighted networks. *Social Networks*, 155-163.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 245-251.

Shawn, M. E. (1954). Group structure and the behavior of individuals in small groups. *The Journal of Psychology: Interdisciplinary and Applied*, 139-149.

Sonego, P. (2011, 7 27). *Word Cloud in R*. Retrieved 2 28, 2012, from One R Tip A Day: http://onertipaday.blogspot.com/2011/07/word-cloud-in-r.html

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press.

Williams, R. (2011, 11 29). *UCINET Software*. Retrieved 09 03, 2012, from Freeman's EIES Data: https://sites.google.com/site/ucinetsoftware/datasets/freemanseiesdata

Wouter de Nooy, A. M. (2005). *Exploratory Network.* Cambridge: Cambridge University Press.

[i] The package can be installed with the commands: `source("http://bioconductor.org/biocLite.R")` and `biocLite("Rgraphviz")`

Atlas slovenske znanosti

RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

R32 Evaluation of data modeling results

Ljubljana, 15.9.2012

Table of content

1 Introduction

This report gives the summary of the results of applying different data modeling methods. Covered are: text analysis methods, modeling the evolution of data using social network analysis and using text analysis and, a combination of social network and text analysis methods.

2 Results of modeling the data with social network analysis

In the first part of the report the data is modeled with social network analysis methods. Two most important covered concepts were: cohesion – investigating who is related to whom and who is not related in the network; and brokerage - dealing with social networks as structures that allow for the exchange of information.

2.1 Cohesion of the network

The average vertex degree of the network was 8.59. This means that in average each researcher is collaborating with 8 other researchers. Observing each of the 7 science groups separately, biotechnical sciences had the highest average vertex degree of 16.84, followed by medical sciences – 10.36, humanities – 7.85, natural sciences and mathematics – 7.46, social sciences – 6.51, engineering sciences and technologies – 5.35, and finally interdisciplinary studies – 0.10. Besides the vertex degree of each node in the network, the connections between science groups were examined by counting the collaborations on projects between researchers belonging to one group with another. The table 1 shows all the connections between science groups.

Table 1. Collaboration between science groups

No	Science group 1	Science group 2	weight
1	Natural sciences and mathematics	Biotechnical sciences	9245
2	Natural sciences and mathematics	Engineering sciences and technologies	8619
3	Natural sciences and mathematics	Medical sciences	3940
4	Engineering sciences and technologies	Medical sciences	2807
5	Engineering sciences and technologies	Social sciences	2803
6	Engineering sciences and technologies	Biotechnical sciences	2196
7	Humanities	Social sciences	2039
8	Biotechnical sciences	Medical sciences	1839
9	Natural sciences and mathematics	Social sciences	1187
10	Engineering sciences and technologies	Humanities	1063
11	Natural sciences and mathematics	Humanities	963
12	Medical sciences	Social sciences	955
13	Biotechnical sciences	Social sciences	559
14	Engineering sciences and technologies	Interdisciplinary research	191
15	Biotechnical sciences	Humanities	188
16	Natural sciences and mathematics	Interdisciplinary research	133

17	Social sciences	Interdisciplinary research	60
18	Humanities	Medical sciences	56
19	Humanities	Interdisciplinary research	27
20	Biotechnical sciences	Interdisciplinary research	18
21	Medical sciences	Interdisciplinary research	7

The researchers from Natural sciences and mathematics, Engineering sciences and technologies and Biotechnical sciences, most intensively collaborate with researchers from other groups.

2.2 Brokerage

Brokerage deals with social networks as structures that allow for the exchange of information. Examined were centralizations of the science groups. Centralized structures are highly efficient for the exchange of information, but they are in the same time to depended on particular parts of the network. Table 2 shows the centrality measure by the three different centrality measures – degree, closeness and betweenness centrality.

Table 2. Collaboration between science groups

Name of the science group	Network Degree Centralizatio	Component Degree Centralization	Component Closeness Centralization	Component Betweenness Centralization
Natural sciences and math.	0.0320	0.0690	0.2068	0.0769
Engineering sciences and tech.	0.0113	0.0343	0.2070	0.0449
Medical sciences	0.0378	0.0645	0.2504	0.0873
Biotechnical sciences	0.0801	0.1315	0.2490	0.0538
Social sciences	0.0279	0.0629	0.1918	0.0637
Humanities	0.0340	0.0651	0.1976	0.1224
Interdisciplinary studies	0	0	0	0

The biotechnical sciences is the most centralized group of sciences according to the degree centralization, what confirms the highest average vertex degree results measured in the previous section. Measured by closeness centralization the most central is medical science group and measured by the betweenness centralization, the most central is humanities group.

3 Results of modeling the data with text analysis

The main part of text modeling was classification of researchers into science and field categories. The classification was performed using the SVM machine classification model. SVM is a machine learning method that non-linearly maps the input vectors to a very high dimension vector space in which the decision surface is constructed (Cortes & Vapnik, 1995). In (Joachims, 1998) it is shown that the SVM is an appropriate method for text classification. The main reasons include the ability to handle high

dimensional input space and suitability for problems with dense concepts and sparse instances. The classification model was tested using the testing dataset, with the 5 folds cross validation technique.

The results of testing the dataset with 74 science fields give the classification accuracy of 0.68. This means that 68% of the researchers were classified in one of 74 scientific fields to which they are actually categorized. The classification accuracy for the dataset with 6 different science groups is 0.84.

4 Results of modeling the evolution of data with social network analysis

The dynamical properties of the network of researcher collaboration were analyzed from the four following aspects:

- Network Growth
- Density of the Network
- Diameter of the Network
- Connected Component of the Network

4.1 Network Growth

The growth of the number of edges in the network in time grows linearly, the liner function: $f(x) = 19111x - 26655$ approximates the growth with $R^2 = 0.9893$. The initial number of edges in the year 1994 is 1749 and the final number of edges in year 2010 is 298453.

4.2 Density of the Network

The exponent a in the densification power law (Leskovec, Kleinberg, & Faloutsos, Graph evolution: Densification and shrinking diameters, 2007) for the examined network is 1.70. This means that the network is becoming denser over time, with number of edges growing superlinearly in the number of nodes. Exponent $a = 1$ corresponds to constant average degree over time, while $a = 2$ corresponds to an extremely dense graph where each node has, on average, edges to a constant fraction of all nodes.

Table 3. Comparison of the SICRIS database densification with the densification of other networks from (Leskovec, Kleinberg, & Faloutsos, 2007)

Dataset	Nodes	Edges	Time	DPL exponent
SICRIS	10,341	298,453	16 y	1.7
Arxiv HEP–PH	30,501	347,268	124 m	1.56
Arxiv HEP–TH	29,555	352,807	124 m	1.68
Patents	3,923,922	16,522,438	37 y	1.66
AS (communication n.)	6,474	26,467	785 d	1.18
Affiliation ASTRO–PH	57,381	133,179	10 y	1.15
Affiliation COND–MAT	62,085	108,182	10 y	1.1
Affiliation GR-QC	19,309	26,169	10 y	1.08

Affiliation HEP–PH	51,037	89,163	10 y	1.08
Affiliation HEP–TH	45,280	68,695	10 y	1.08
Email	35,756	123,254	18 m	1.12
IMDB	1,230,276	3,790,667	114 y	1.11
Recommendations	3,943,084	15,656,121	710 d	1.26

4.3 Diameter of the Network

The effect of shrinking of effective diameter with time is the same for the complete network and for the largest connected component. The network had biggest effective diameter of 10.91 in year 1996, while size of effective diameter in last network snapshot from 2010 is 5.96. This means that at least 90% of all connected pairs of nodes can be reached within 6 hops.

4.4 Connected Component of the Network

The proportion of nodes belonging to largest connected component in respect to all nodes of the network grows logarithmically. This means that the connected component obtains a big proportion of all nodes very fast and then the proportion keeps growing slower. In the first six years (till tear 1999), the size of the connected component reaches 85% of all nodes, while in the next 11 years (till year 2010) the proportion reaches final 98%. The only phase where the proportion does not grow is the "knee" on the figure 21 where the proportion falls from 17% to 7% from initial year 1994 to year 1995. The size of the connected component is very large, it takes most of the nodes and edges of the network and it grows linearly. The number of nodes of the network which are the part of the largest connected component in the last snapshot of the network in the year 2010 is 10200 from 10341 of all nodes (0.9863%) and the number of such edges is 297994 from 298453 (0.9985%) of all edges.

5 Results of modeling the evolution of data with text analysis

The analysis of evaluation of text resulted with several visualizations showing how the content of the projects changed over the years. The graph on Figure 1 shows the number of initiated projects clustered into 6 different science groups according to the description in English language. Different color of layers represent different science groups. Projects classified into natural sciences and engineering hold the biggest portion of the projects thru out the years. Examining the number of new project according to subfields to which they belong, it can be noticed that the proportion of the project for a subfield is always similar.

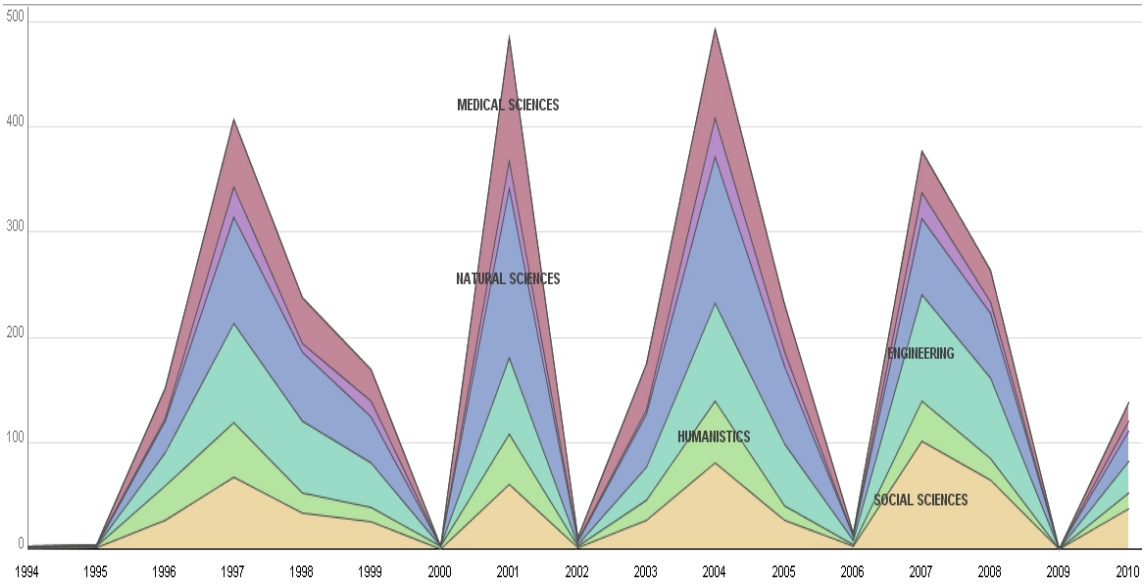


Figure 1. Number of projects in years, clustered into 6 science groups

6 Results of combined modeling (temporal, text and social network analysis)

Finally the results of combining different analysis techniques in classification, and vertex centrality measurement are presented. The results of classification clearly show how even naïve approach of including information from node connections improve the results, while some of the standard relational learning algorithms perform even better. The centrality measure presented in the second part takes into account the textual information with the standard network information.

6.1 Classification

Table 4 shows the result of different classification algorithms applied on problem of classification 8585 researchers into one of the 74 science field. The evaluation was performed using the 5-fold cross validation technique and the results show the classification accuracy. The approach of including attribute of the neighboring nodes improves the classification results (methods from *a* to *f*), with best performance when only attributes from the nodes with the similarity greater than 0.2 are included (method *e*).

Table 4 - Classification accuracy

Method	**Classification accuracy**
a) Original keywords	0.59
b) Keywords from all neighbors	0.61
c) Keywords from neighbors (cosine similarity > 0)	0.64
d) Keywords from neighbors (cosine similarity > 0.1)	0.64
e) Keywords from neighbors (cosine similarity > 0.2)	**0.67**

f) Keywords from neighbors (cosine similarity > 0.4)	0.63
g) Network-only Link-Based classifier (nLB) (Lu & Getoor, 2003)	**0.79**
h) Weighted-vote relational neighbor (wvRN) procedure (Macskassy & Provost, A simple relational	0.76

Applying some of the standard relational learning algorithms improves the classification further more (the best performing is the algorithm *g* from the table 4).

6.2 Centrality measure

Proposed centrality measure was applied on researchers and projects dataset, but also on the Freemans's EIES dataset standard in the social network analysis. The first column of the table 5 shows the rankings of the researchers from the dataset using only the vertex degree. The second column of the table shows the renking with the generalized degree centrality measure that takes the number of connections with the higher average weight of the connections as a positive thing for ranking. The third column of the table shows the rankings with our proposed centrality measure applied. The last column of the table shows the combined usage of vertex degree, connection weights and the similarities between vertices. The rankings in the last column are different from the second and third, but it noticeable that it presents the combination of the two.

Table 5. Aplying te proposed centrality measutre on the Freemans's EIES dataset to compare and integrate the measure with the generalized centrality measure proposed by (Opsahl, Agneessens, & Skvoretz, 2010)

Degree $(\alpha = 0; x = 0)$	Degree and weight $(\alpha = 0.5; x = 0)$	Degree and similarity $(\alpha = 0; x = 0.5)$	Degree, weight and similarity $(\alpha = 0.5; x = 0.5)$
Lin Freeman (31)	Lin Freeman (314)	Lin Freeman (75)	Lin Freeman (1382)
Sue Freeman (31)	Barry Wellman (249)	Nick Mullins (75)	Barry Wellman (987)
Nick Mullins (31)	Russ Bernard (200)	Barry Wellman (64)	Nick Mullins (511)
Phipps Arabie (28)	Sue Freeman (180)	Ron Burt (49)	Pat Doreian (342)
Barry Wellman (28)	Doug White (177)	Sue Freeman (41)	Ron Burt (322)
Doug White (28)	Nick Mullins (142)	Phipps Arabie (38)	Russ Bernard (305)
Russ Bernard (25)	Lee Sailer (134)	Pat Doreian (36)	Sue Freeman (273)
Ron Burt (20)	Pat Doreian (129)	Richard Alba (34)	Lee Sailer (239)
Pat Doreian (20)	Ron Burt (85)	Doug White (33)	Doug White (234)
Richard Alba (18)	Richard Alba (77)	Russ Bernard (32)	Richard Alba (203)
Jack Hunter (18)	Steve Seidman (70)	Lee Sailer (24)	Davor Jedlicka (129)
Lee Sailer (17)	Phipps Arabie (66)	Jack Hunter (23)	Phipps Arabie (96)
Steve Seidman (16)	Jack Hunter (65)	Davor Jedlicka (21)	Jack Hunter (92)
Carol Barner-Barry (15)	Al Wolfe (63)	Carol Barner-Barry (18)	Maureen Hallinan (91)
Al Wolfe (14)	Carol Barner-Barry (60)	Al Wolfe (17)	Don Ploch (90)
Paul Holland (12)	Paul Holland (49)	Steve Seidman (16)	Al Wolfe (85)

John Boyd (11)	John Boyd (45)	John Boyd (15)	Carol Barner-Barry (79)
Davor Jedlicka (11)	Davor Jedlicka (45)	Don Ploch (14)	John Boyd (77)
Charles Kadushin (7)	Maureen Hallinan (37)	Paul Holland (13)	Steve Seidman (70)
Nan Lin (7)	Don Ploch (28)	Charles Kadushin (12)	Paul Holland (57)
Don Ploch (7)	Claude Fischer (22)	Nan Lin (11)	Charles Kadushin (51)
Claude Fischer (6)	Mark Granovetter (22)	Claude Fischer (9)	Claude Fischer (48)
Mark Granovetter (6)	Charles Kadushin (21)	Maureen Hallinan (9)	Nick Poushinsky (44)
Maureen Hallinan (6)	Nan Lin (20)	Nick Poushinsky (8)	Nan Lin (38)
Nick Poushinsky (5)	Nick Poushinsky (18)	Mark Granovetter (7)	Mark Granovetter (29)
Sam Leinhardt (4)	Joel Levine (17)	Sam Leinhardt (6)	Joel Levine (24)
Joel Levine (4)	Sam Leinhardt (10)	Joel Levine (5)	Sam Leinhardt (17)
John Sonquist (4)	John Sonquist (9)	John Sonquist (4)	John Sonquist (9)
Brian Foster (3)	Brian Foster (7)	Gary Coombs (4)	Brian Foster (9)
Ev Rogers (2)	Gary Coombs (6)	Brian Foster (4)	Gary Coombs (9)
Gary Coombs (3)	Ed Laumann (6)	Ev Rogers (2)	Ed Laumann (6)
Ed Laumann (2)	Ev Rogers (4)	Ed Laumann (2)	Ev Rogers (6)

7 Conclusion

To understand the scientific community data and to explore the possible applications and tools which could be developed for the end user, detail analysis with the methods from text mining, social network analysis and temporal analysis was applied on the data. This report gives the summary of results of this analysis. The results show broad range of possibilities, from semi-automatic ontology construction to classification. Some of the results (like cohesion and brokerage of the network) indicate that the techniques suitable for practical application. On the other hand, some of the results (like centrality measure with integrating content of nodes) revile nice potential and space for further research work.

8 Bibliography

Artificial Intelligence Laboratory - Institute Jozef Stefan. (n.d.). *Artificial Intelligence Laboratory*. Retrieved 11 15, 2011, from Artificial Intelligence Laboratory: http://ailab.ijs.si/tools/text-garden/

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*, 273-297.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of ECML-98, 10th European Conference on Machine Learning.* Chemnitz.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*.

Lu, Q., & Getoor, L. (2003). Link-based classification. *In Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, (pp. 496–503).

Macskassy, S. A., & Provost, F. (2003). A simple relational classifier. *Multi-Relational Data MiningWorkshop (MRDM) at the Ninth ACM SIGKDD International Conferenceon Knowledge Discovery and Data Mining.*

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 245-251.

Atlas slovenske znanosti

RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

R41 Visualization and interactive analysis

Ljubljana, 15.9.2012

Table of content

Introduction

In this report tools developed for interactive visualizations of data are presented. The tools are divided into two categories: visualizations of collaboration and visualizations of competences including the temporal data.

1 Visualizing collaboration

To know who works with whom, can tell us a lot about a field of science. The number of collaboration of a researcher can be a nice indicator of her/his research activity. We can look deeper to see what kind of interactions produce the best research results, or discover potential collaborations which could be very beneficial. All this is motivations for constructing the tools for visualizing collaboration, which are presented in this chapter.

1.1 All-to-all visualization of researchers collaboration

Collaboration diagram visualizes a set of researchers as vertices. The connection between vertices represents the collaboration on project between two researchers. The set of the researchers to be visualized can be constructed in more ways. One option is to use a search term to retrieve all the researchers who have their descriptions related to the search term. For instance, the researchers from the Figure 1 are obtained by search term “artificial intelligence”.

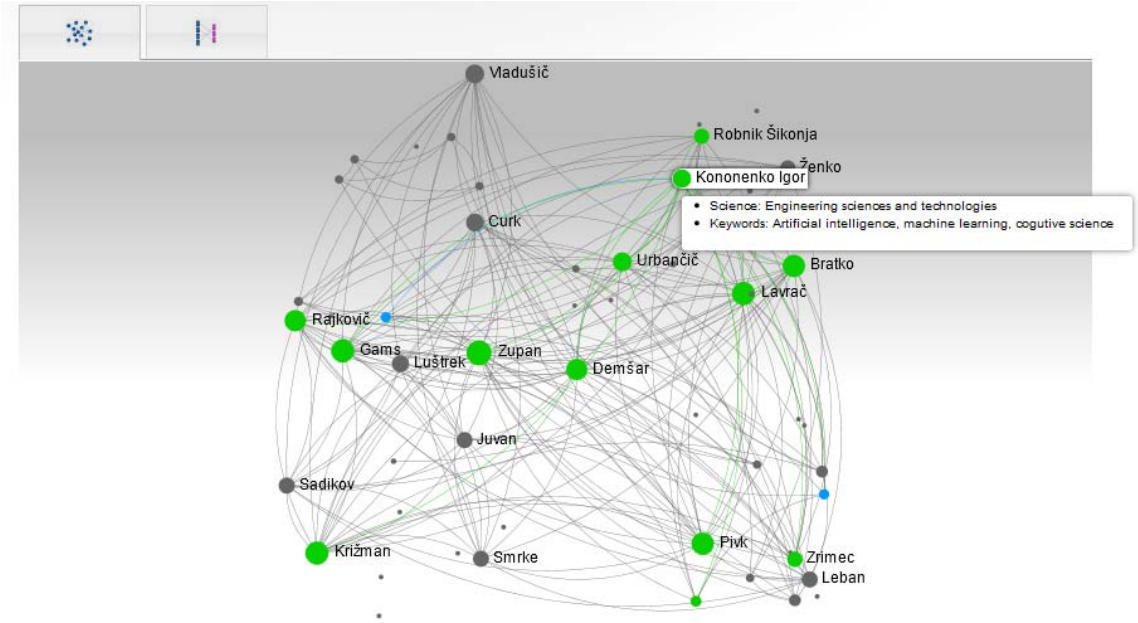


Figure 1 - Collaboration of researchers retrieved with the search term "artificial intelligence", visualized with the all-to-all type of collaboration map

Second possible way of retrieving the set of researcher could be by selecting all the researchers who collaborate with a particular researcher – visualization in Figure 2.

The vertices of the graph can be of different sizes and of colors. The sizes of the vertices are proportional to the vertex degree of the vertex on the visualization. The different size of the vertex controls the displaying of the name associated to the vertex. Figure 2 is visualization with many

vertices, but the names of only the biggest vertices (the most important) are displayed. This enables us to immediately spot the most important vertices of the graph (Mladenic and Grobelnik on figure 1).

Different color of the vertex represents the science group to which the researcher belongs to. The information about association to science group is already present in the database (the ARRS science categorization is used: http://www.arrs.gov.si/en/gradivo/sifranti/sif-vpp.asp). Most of the vertices on the Figure 1 are in green color, which represents the engineering sciences and technologies. Some of the vertices are light blue; they represent the researchers from the medical sciences. There are also two vertices in dark blue and one in dark red color, these color represent natural sciences and mathematics and social sciences respectively. The dominant color of the diagram immediately gives the information about which science group(s) is/are most important for some concept (on the visualization on Figure 1, the green color is dominant, hence the engineering sciences and technologies is the most important science group for the artificial intelligence).

On the Figure 1, there are many vertices and connections in grey color. This is due to the state of the visualization which appears when the cursor is moved over one of the vertices. When this event is triggered, only the vertices which are connected with the selected vertices keep their original color, while all others become gray. In addition to that, the full name of the selected researcher is displayed, together with his science group and keywords (as shown for Kononenko Igor on the Figure 1).

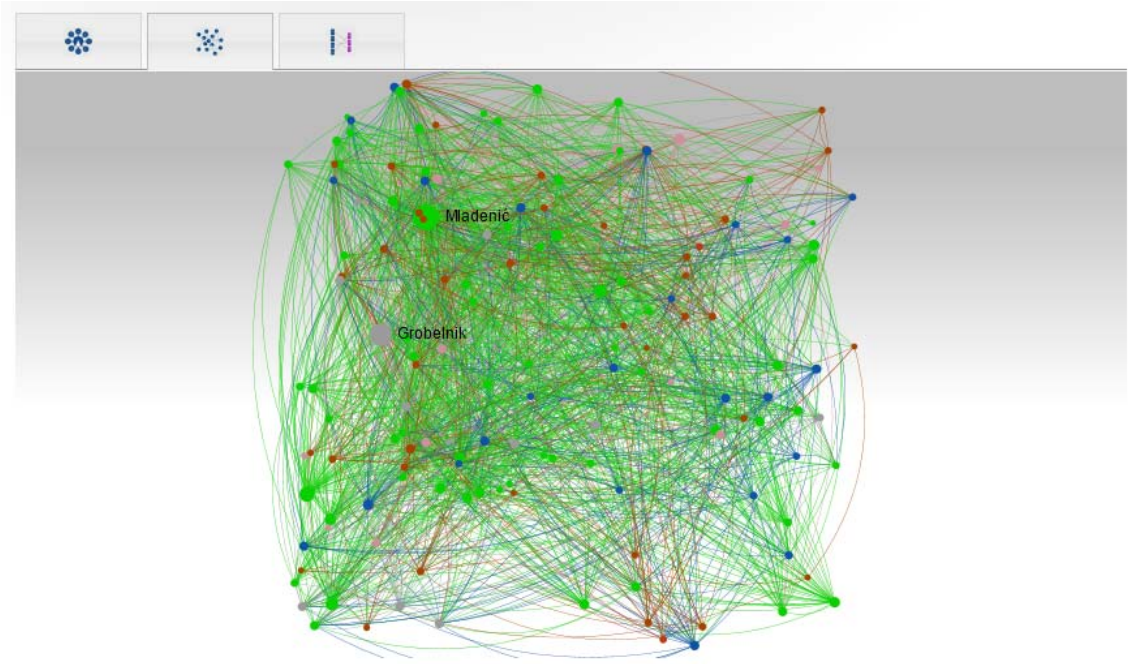


Figure 2. Collaboration map of a researcher

Other interactive action available on the visualization is zooming. The diagram can be zoomed in and out, by the simple mouse scroll action. When the diagram is zoomed, the labels of smaller vertices are displayed. Zoomed state of the visualization from Figure 2, with a selected researcher is show in

Figure 3. Each vertex on the visualization is clickable; clicking on a vertex will result with loading the same type of the visualization for the clicked researcher.

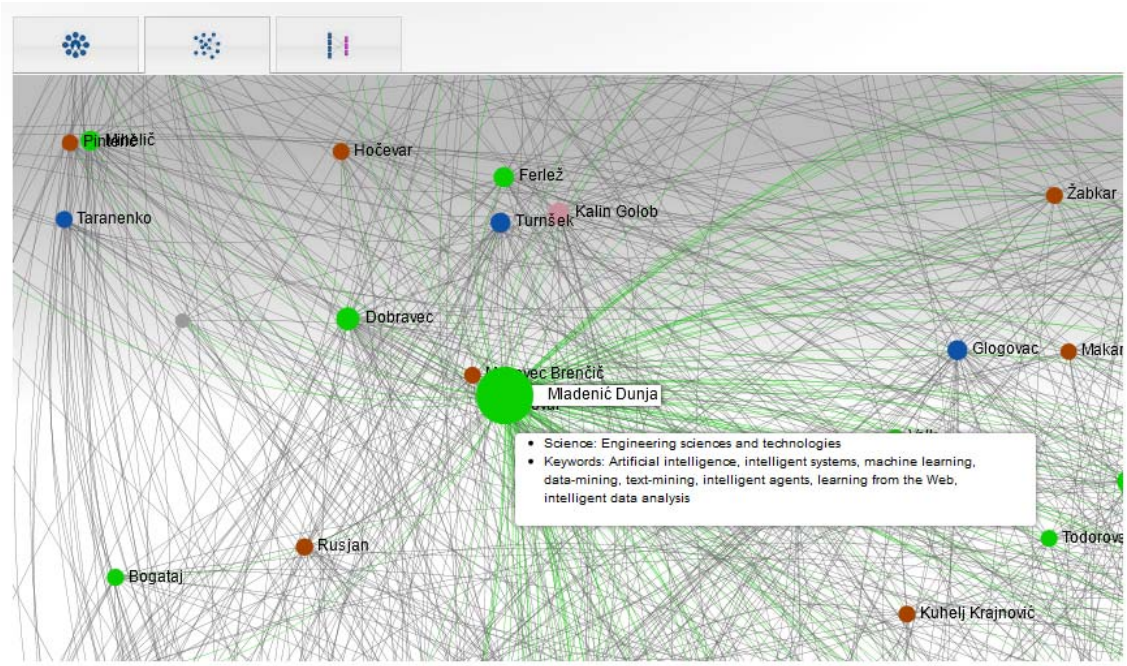


Figure 3. Collaboration map of a researcher Dunja Mladenic after zooming in

1.2 One-to-all visualization of researcher's collaboration

One-to-all type of collaboration map is an ego centric type of collaboration diagram. This means that the focus is always on one selected researchers, while the others are displayed according to the relationship with the selected one. The layout we used to construct the one-to-all type of collaboration map can be called circular. The selected researcher is placed in the middle of the diagram. The researchers with whom the selected researcher is collaborating are placed around him in concentric circles, with the radius inversely proportional to the strength of collaboration. This results with the placement of the researchers who are more important for the collaboration to the center of the diagram. The Figure 4 shows this type of collaboration map for Dunja Mladenic. The researchers closer to the center have stronger connection to the central researcher, these are: Grobelnik, Bojadziev, Fortuna, Brank, Skranjc and Jermol.

The diagram has a vertex display feature called fisheye. With this feature, the nodes around the mouse pointer are slightly zoomed, so that their names are displayed. With some user interaction, this enables visibility of labels for even the densest parts of the graph, while initially keeping the graph clear from too numerous labels.

The colors of the vertices are like with other types of vertices associated with different science groups. Most of the vertices on the Figure 4 are green (engineering sciences and technologies), but there are also some dark red (social sciences) and dark blues (natural sciences and mathematics). There is also few pink vertices (humanistic sciences) and grey vertices which represent researchers without specified science field.

The interactive features of the diagram include zooming; mouse over which displays full name, science and keywords of the selected researcher; and vertex click which opens generates the diagram for the selected researcher.

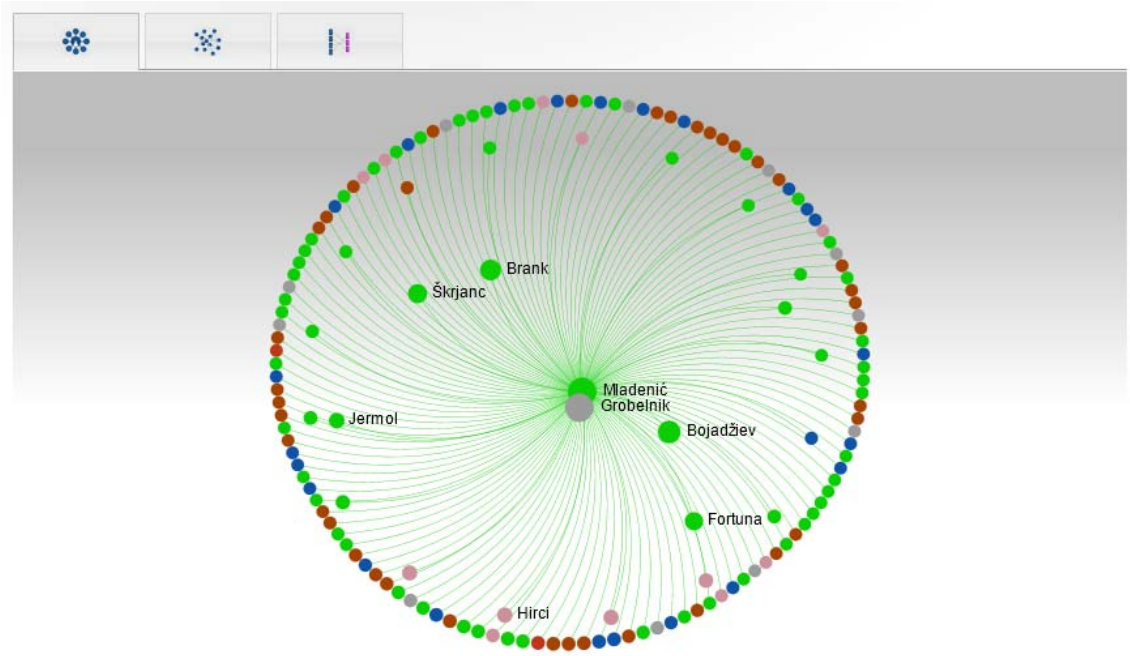


Figure 4. One-to-all type of collaboration map of Dunja Mladenic

1.3 Collaboration between organizations

Next visualization shows the collaboration between organizations, with emphasis on the collaboration between research organization and companies.

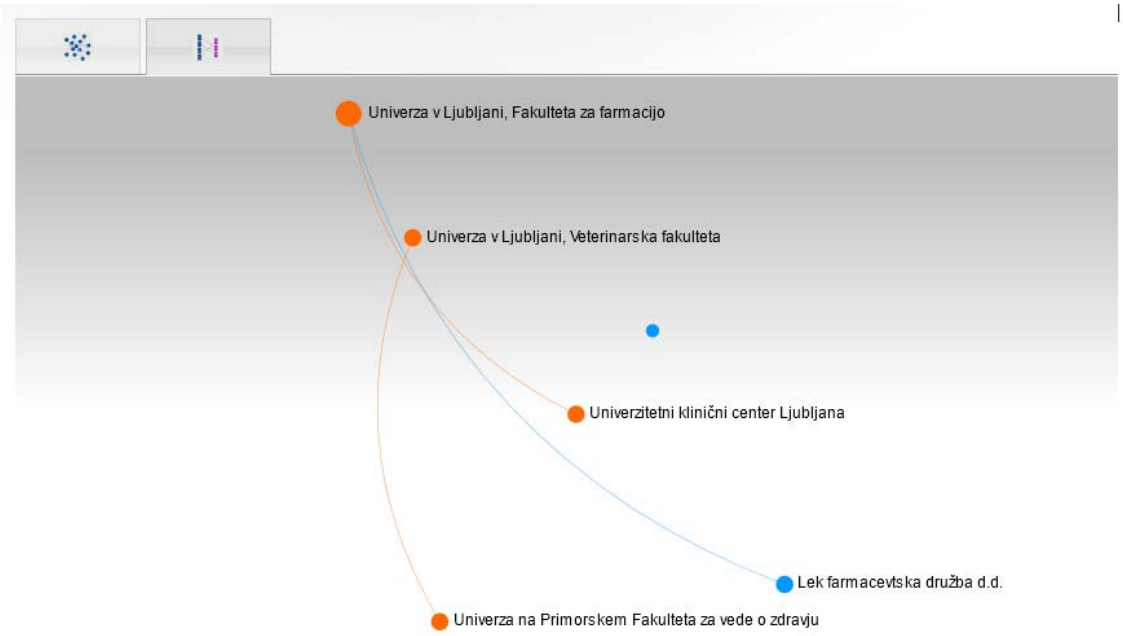


Figure 5. Collaboration between organizations retrieved with search teram "pharmacy"

The set of organizations to be visualized are obtained from a set of researchers. Researchers can be obtained either by searching by a search term or by selecting researchers who collaborate with an individual. Researchers collaborate on projects and since they belong to organization, the diagram of collaboration between organizations can be constructed.

To emphasis the collaboration between research organizations and companies, different color is used for them and each group is positioned on a different part of the visualization.

By placing the cursor above a vertex representing an organization, the size of all the vertices connected to it is increased and their labels are displayed.

2 Visualizing competences

Visualizations of competences should provide a visual and interactive summary of competences of a researcher or a set of researchers created by some criteria.

2.1 Visualizing projects connected by keywords

Vertices of the visualization represent research projects. The set of projects can be obtained in more ways. One way of obtaining the projects could be by searching and retrieving the projects which are relevant for the search term. Different way of obtaining the list of projects can be selecting all the projects of some researcher. Another way could be selecting projects from some science field, or from some timeframe. The criteria for selecting the projects depend on our goal, i.e. competences of what or who we want to show. This can be – competences of a researcher, competences of a science subfield, competences in a point of time, or competences related to a particular research topic.

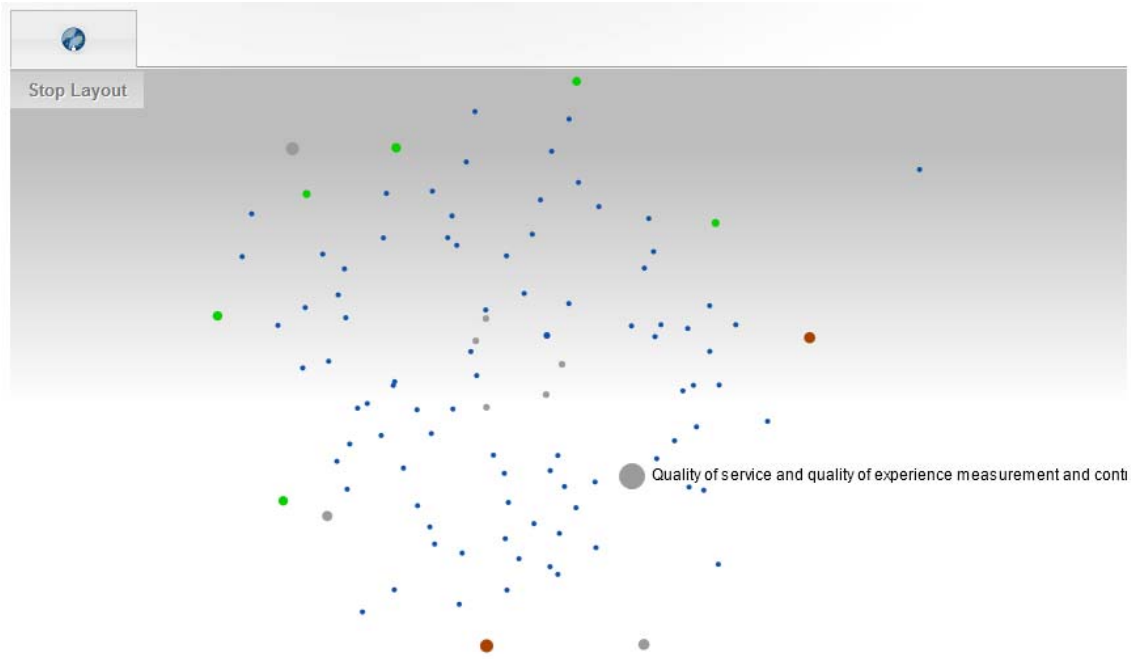


Figure 6 - First stage of the competence map of a researcher

Since most of the projects has keywords associated to them, the visualization is constructed in such a way that the keywords are separate small vertices connected to the projects which they describe. In case that more projects have the same keyword, only one vertex for the keyword will be created and it will be connected to more projects. In this way the connection between projects is made. While connecting the projects based on keywords, additional effort is needed to remove the effects of misspelling of the keywords, so that the projects do get connected if the difference between keywords are only caused by grammatical error (e.g. “artificial inteligence”, “artifical intelligence” and “artificial intelligence” should be considered as the same keyword). To do this, Levenshtein distance is computed between each pair of keywords, and if the distance was not exceeding some threshold, the keywords are considered equivalent. Some additional criteria were considered for connecting the keywords (e.g. synonyms, hypernyms, etc.), but they were not implemented for the

visualization.

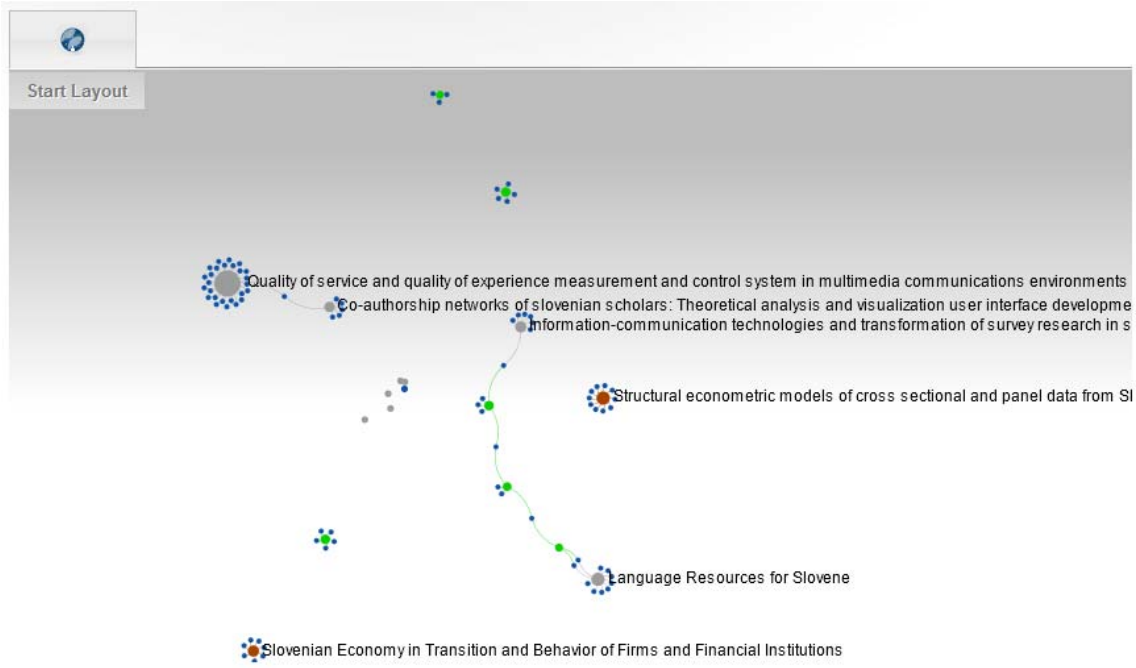


Figure 7. Layout of a competence map in latter stages

Since project are connected with keywords, a good layout of the graph would be to position the more connected projects closer to each other than the less connected ones, what would result with the groups of similar projects on the visualization. A class of fore-vector layout algorithms can perform this task, so one variation of Fruchterman Rheingold algorithm was applied for the visualization. The layouting is continues and it is displayed dynamically. This means that initially, all of the projects and keywords are positioned randomly, and then they are moved on the visualization to get positioned according to their connections. The Figure shows the competence map of a researcher in the early stages of layouting, when the positions of vertices are still quite random, whereas the Figure 7 shows the diagram after few seconds of layouting. After the visualization is loaded, the layouting automatically starts and after few seconds (for example 5 second) it stops. Otherwise it can be stopped and started at any time using the button in the upper left corner (Figures 6, 7 and 8).

The colors of the projects represent the science group to which the projects belong to, while the projects sizes are proportional to their vertex degree, i.e. to the number of keywords describing them.

The competence map in Figure 7 shows projects of a researcher. Most of the projects are from engineering sciences and technologies (green vertices), few are from social sciences (dark red vertices) and some of them are not categorized (gray vertices). In the middle of the visualization a group of 5 projects connected in a line can be observed. With some interaction of user with the diagram, it could be easily discovered what the common topic of this group of projects is. But even just the names of the two projects on the edges of this group indicate that they are about ICT and languages.

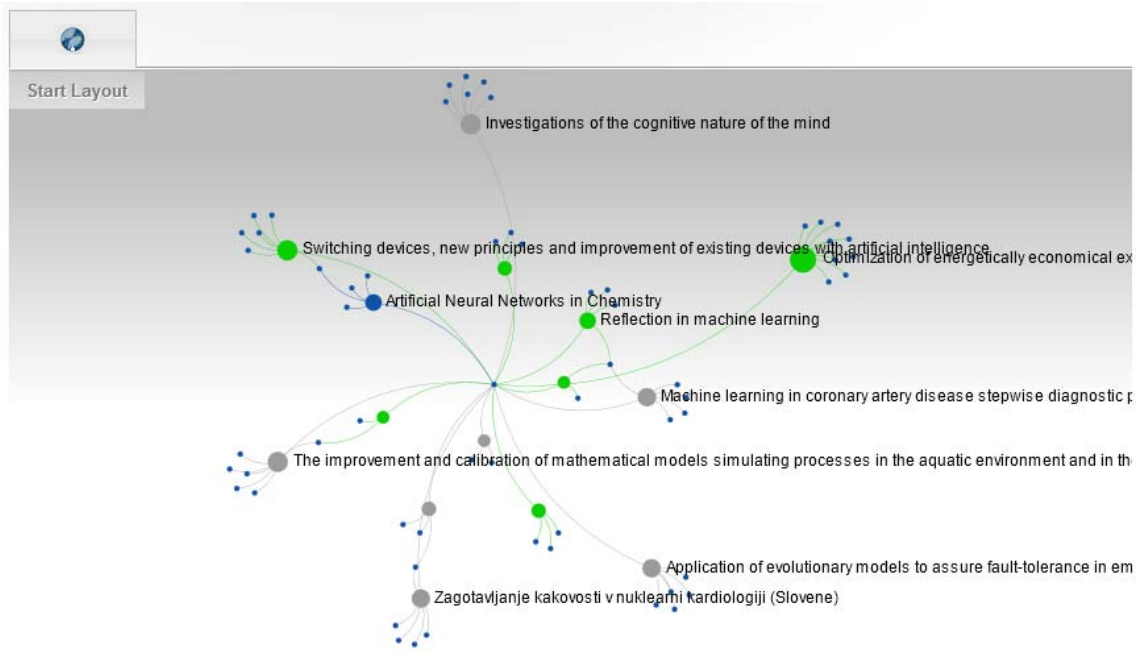


Figure 7. Competence map of the projects retrieved with the search term “artificial intelligence”

The Figure 8 shows the projects retrieved with the search for a research topic – artificial intelligence. All the projects on the visualizations are connected with one central keyword, which is obviously “artificial intelligence”. But some of the projects have some additional connections, so they are placed next to each other. For example projects “Switching devices, new principles and improvement of existing devices with artificial intelligence” and “Artificial Neural Networks in chemistry” are connected and placed in the same part of the visualization.

2.2 Visualizing temporal aspect of projects

The aim of the visualization is to show in a graphical way the dynamic of number of projects, selected by some criteria.

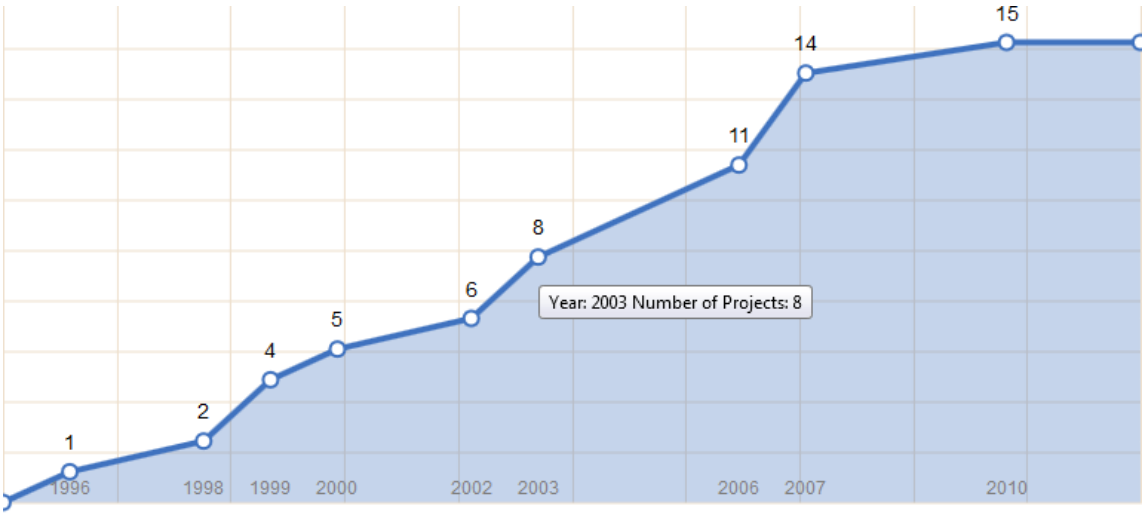


Figure 8. Visualization of number of projects of a researcher in time

The user should easily recognize the productive periods with big number of new projects (slopes) and the less productive periods with fewer number of projects (flat areas). The projects for the visualization can be selected in different ways, for example, projects retrieved with some keywords or all the projects of a researcher. The y axis of the diagram hold the number of projects, while the x axis shows the year in which the projects are initiated. The visualization is cumulative, meaning that each year contains the projects initiated in that year, plus all the projects from the previous years. The diagram is of a cumulative type, because it is used with the competence map to show the competences of a researcher, or a particular set of researchers. Since the competences do not simple disappeared after the projects is ended, but incrementally grow with new knowledge and experiences gained from the projects, the temporal diagram is cumulative. The interactive features of the visualization include displaying the number of projects in the particular year with dragging the pointer over a node of the diagram, while clicking the node causes filtering of the list of projects so that only the projects initiated at the selected year or earlier are displayed.

The example of the temporal diagram of a researcher is shown in the Figure 8. It can be observed that the most dynamic years were 1999, 2003 and 2007 (the biggest slopes in figure 8), when respectively 2, 2 and 3 new projects were initiated since the previous year.

Conclusion

This report presented the main characteristics of the developed tools for visualizing collaboration between researchers and organizations, and competences of different groups of researchers.

Atlas slovenske znanosti
RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

R42 Results of interactive visualizations evaluation

Ljubljana, 15.9.2012

Table of content

1 Introduction

We have performed evaluation with the user in order to get a feedback about the design and functionalities of the developed visualizations. The evaluation techniques was questionnaire, which design is described in the next section. In the second section, the categorical as well as descriptive results of the evaluation are carefully analyzed.

2 Design of the evaluation

Evaluation is a systematic way of reflecting on and assessing the value of what is being done (i.e. a project, a program, an event). Evaluation is commonly interpreted as an end product or an activity taking place at the end of a project. However evaluation should be considered as a process, taking place across all phases of a project, used to determine what has happened and whether the initial aims of the project have been carried out and achieved. Evaluation is more than assessing and measuring; it helps set the stage for a culture of learning, change and improvement. (University College London)

The evaluation method used for our tasks was Questionnaire. The questionnaire consisted of 4 scenarios and three questions for each of the scenario (see Appendix). The intention of four different scenarios was to evaluate the usage of the developed tools in different use cases. The scenarios were:

1. “Discovering researcher’s competences - you are considering collaborating with a particular researcher. Use the Science Atlas to discover the competences of the researcher. You are interested in which filed is he working in, what are his expertise and projects he is/was involved in.”
2. “Discovering researcher’s collaboration - you are considering collaborating with a particular researcher. Use the Science Atlas to discover the collaboration of the researcher. You are interested in number of collaborating researchers, with whom is he collaborating most intensively and what their competences are.”
3. “Searching for consultancies from research community – as a member of a company you are interested to find researchers and research institutions competent in your area of work. Use the Science Atlas and perform the search by keywords.”
4. “Searching for companies related to some research area - you want to find the companies which could implement some result of your research work. Use the Science Atlas to find the related companies.”

The questions given to the respondent were:

1. How good was the portal in meeting the requests presented in the scenario?
2. Do the visualizations improve searching and acquiring information compared to using just the traditional structured websites without the visualizations?
3. Please write any suggestions you may have related to tools covered by this scenario.

The first question could be answered with a single choice from the 5 options scale (excellent, very good, good, fair and poor), the second option had a similar 5 options scale (extremely, very, somewhat, slightly, and no improvements) and the third question could be answered with a textual description.

2.1 Analysis of the results

The evaluation was performed on 17 individuals, students and employees of a research institute. In the first part the analysis of the responds to the questions on scale from 1 to 5 were analyzed, while the second part deals with answers to the descriptive questions about suggestions for improvement.

2.1.1 Analysis of the ratings

Table 1 shows the results of evaluation as average ratings for different functionalities covered with the four scenarios. The overall grade is 4 out of 5, what corresponds to "very good" rating for the perceived value and for the perceived improvement over traditional methods it corresponds to answer "very".

Table 1 - Average ratings for different functionalities

Scenario and aspect	Average rating
Discovering researcher's competences - perceived value	**3.93**
Discovering researcher's competences - increase of efficiency	3.60
Discovering researcher's collaboration - perceived value	3.40
Discovering researcher's collaboration - increase of efficiency	3.20
Searching for consultancies from research community - perceived value	3.67
Searching for consultancies from research community - increase of efficiency	**4.07**
Searching for companies related to some research area - perceived value	3.53
Searching for companies related to some research area - increase of efficiency	3.73

The highest rating for the perceived value was for the functionalities covered by the "Discovering researcher's competences" scenario, while the biggest increase of efficiency over traditional methods was noted in the "Searching for consultancies from research community" scenario.

2.1.2 Analysis of suggestions

Table 2 contains suggestions grouped into four categories: 1) descriptions of elements, 2) graph layout b, 3) node size and position, and 4) Sources, retrieval and display of data. The number in the parenthesis next to each suggestion stands for the frequency of the suggestion in the evaluation results.

Table 2. Suggestions obtained in the evaluation, grouped into 4 categories

Category	Suggestions
Descriptions of elements	Legend for colors and node size (30); Have a one-line description of what it shows (3); Show other competences of people and organizations in results (2); Add text to the icons for collaboration and competence maps (2); Include time dimension of the projects (2); Add text to tabs (1); Explanation meaning of layout and distances between nodes is missing (1); More obvious loading indicator (1); Links should bring some info (1); Show collaboration type (1);It is not obvious which dots are projects and which are keywords (1); Names missing even if enough space (1); Show all institution names on the visualization, they're not that many (1); Show more companies' info (1); Show an example (1)
Graph layout	Remove unnecessary lines (3); Connections between people on the same project leads to cliques (1); Use smart graph layout (1); Positioning should be calculated according to some relaxation to the links, not random (1); Make better use of the plane (1); Do not use alphabetical ordering (1); Graphs are too crowded, consider adding some filters (show top 10 collaborators) (1); Cluster researchers according to collaboration strength (1); Show only connections which exist in the absence of the main node (1); Text on the graph shouldn't cover each other (1); Make better use of the plane (1); Remove the fisheye (1); 3D Graph (1)
Node size and appearance	Increase minimum node size (4); Better visibility of keywords (4); Different sorting of the researchers (2); Bigger fonts (1); Sorting of researchers should be done based on importance (1); Order circles by similarities (1); Sort the nodes by color (1); List of dots should be based on the importance of researcher's connections, not just company connections; More intuitive sizes of nodes (1); Use number of projects for node size (1); Node size should be based on relevance to query or company size (1)
Sources, retrieval and display of data	Show more search results (4); Use non-national projects (2); Add the institutions (1); Table with projects and keywords (1); Issues with projects without keywords (1); Enable search in English language (1); Consider possible ranking of researchers, like number of projects and connections (1); Traditional view is also useful (1); Scrolling instead of paging (1); Show competence relevance score (1)

There were many suggestions about explanations and descriptions need for some elements of the visualizations. The most common suggestion in this category, as well as overall, was the need for a legend that would clarify the meaning of different colors and sizes of the nodes. The second most

frequent suggestions in this category was the need for a textual description for each visualization, so the user could know what exactly each visualization represents.

Many suggestions were related to the layout of the graph. There were mostly about removing lines eater by filtering by the connection weight, or by setting additional requirements for line to be drawn. Other suggestions were about positioning and clustering of the nodes on of the graph.

Next group contained suggestions about nodes of the visualizations. The main message from these suggestions was to set more intuitive criteria for node sizes, to sort and cluster the nodes and to make the appearance of the nodes and accompanied text clearer and better visible.

The last group of suggestions was about including additional data sources, retrieving and displaying the data.

3 Conclusion

This report show results of the evaluation of the developed visualizations. The evaluation was performed using the questionnaire technique which design was described in the first part of this report. In the second part the results of the questionnaire are presented. The ratings for different usage scenarios given to the users are in general very good, with the space for improvements. The biggest value was gained with the numerous suggestions obtained from the users. They were clustered in 4 categories and present a valuable compass for further design and development of the Science Atlas.

4 Bibliography

University College London. (brez datuma). *UCL*. Prevzeto 5. 9 2012 iz Evaluation Methods: http://www.ucl.ac.uk/public-engagement/research/toolkits/Methods/#

Appendix

Slovenian Science Atlas web portal evaluation

Do you agree to use this survey for a research purpose? YES or NO

Age: _____ Gender: | F | M | Occupation: _____

Scenario 1: Discovering researcher's competences

You are considering collaborating with a particular researcher. Use the Science Atlas to **discover** the **competences** of the researcher. You are interested in which filed is he working in, what are his expertise and projects he is/was involved in.

1.1. How good was the portal in meeting the requests presented in the scenario?

- ☐ Excellent
- ☐ Very good
- ☐ Good
- ☐ Fair
- ☐ Poor

1.2. Do the visualizations improve searching and acquiring information compared to using just the traditional structured websites like SICRIS?

- ☐ Extremely
- ☐ Very
- ☐ Somewhat
- ☐ Slightly
- ☐ No improvements

1.3. Please write any suggestions you may have related to tools covered by this scenario.

_____

Scenario 2: Discovering researcher's collaboration

You are considering collaborating with a particular researcher. Use the Science Atlas to **discover** the **collaboration** of the researcher. You are interested in number of collaborating researchers, with whom is he collaborating most intensively and what their competences are.

2.1. How good was the portal in meeting the requests presented in the scenario?

- ☐ Excellent
- ☐ Very good
- ☐ Good
- ☐ Fair
- ☐ Poor

2.2. Do the visualizations improve searching and acquiring information compared to using just the traditional structured websites like SICRIS?

- [ ] Extremely
- [ ] Very
- [ ] Somewhat
- [ ] Slightly
- [ ] No improvements

2.3. Please write any suggestions you may have related to tools covered by this scenario.

Scenario 3: Searching for consultancies from research community

As a member of a company you are interested to **find researchers** and research institutions competent in your area of work. Use the Science Atlas and perform the search by keywords.

3.1. How good was the portal in meeting the requests presented in the scenario?

- [ ] Excellent
- [ ] Very good
- [ ] Good
- [ ] Fair
- [ ] Poor

3.2. Do the visualizations improve searching and acquiring information compared to using just the traditional structured websites like SICRIS?

- [ ] Extremely
- [ ] Very
- [ ] Somewhat
- [ ] Slightly
- [ ] No improvements

3.3. Please write any suggestions you may have related to tools covered by this scenario.

Scenario 4: Searching for companies related to some research area

You want to **find** the **companies** which could implement some result of your research work. Use the Science Atlas to find the related companies.

4.1. How good was the portal in meeting the requests presented in the scenario?

- [ ] Excellent
- [ ] Very good
- [ ] Good
- [ ] Fair
- [ ] Poor

4.2. Do the visualizations improve searching and acquiring information compared to using just the traditional structured websites like SICRIS?

- [ ] Extremely
- [ ] Very
- [ ] Somewhat
- [ ] Slightly
- [ ] No improvements

4.3. Please write any suggestions you may have related to tools covered by this scenario.

_____

Atlas slovenske znanosti

RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

R51 – Server side and web client

Ljubljana, 15.3.2013

Table of content

1 Introduction

This report gives an overview of the server and client side of the Slovenian Science Atlas application, and serves as a technical documentation of the system. The goal of this report is to represent the architecture of the system, explain the individual parts of the architecture and the interaction between them. The main two parts of the system are the server side (backend) and the client side (front-end) of the application, which is the dichotomy applied also to the structure of this report.

2 Server side

The server side of the application represents the core that contains all the functionalities provided with the application. The server side of application is installed on the web server hosting machine, which is a powerful machine with big amounts of memory and processor power. The functionalities of the server are written in C# programing language. The architecture of the system is arranged is such a way that all the computational demanding tasks are performed on the server side. These functionalities are then accessed by the clients in the form of services which are independent on the device and software (operating system) of the users. Figure 1 represents the architecture of the system.

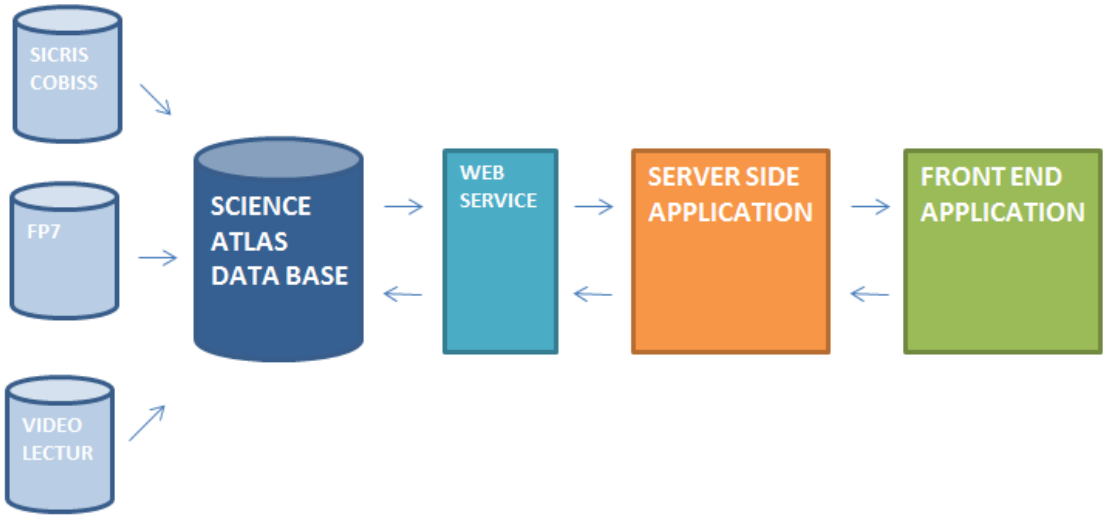


Figure 1. Architecture of the application

The data layer of the application is constructed from the database (Science Atlas Data Base) that integrates data from different sources (SICRIS, COBISS, FP& projects, VideoLectures.NET). Next, the web service is an interface between the data and the users. The request and the deployment of the data is mediated thru the module that performs various functionalities on the server side module. The last module is the front end application, with which all the functionalities are represented to user in the form of a client web application. In the rest of this report, the server side of the application will be analyzed in more detail.

2.1 Web service

Web service is the part of the application that enables the transferee of data from and to database. Using the web service, the data is transferred in a standard format (XML, JSON) and it is independent on the rest of the application which makes the provided data easily usable in other applications. The web service is written in C# programing language and .NET Framework 3.5.

2.1.1 Researchers data

2.1.1.1 ProfileOfRsrJson

The call returns the profile information of a researcher (with the specified id input parameter) that include first and last name of the researcher, classification of the researcher (codes and names of science group, field and subfield), organizations of the researcher and contact information. In addition to this, the call returns all the projects on which a researcher has worked on.

Input parameter: string id

Query:

```
select science from tblPrjClassification where prjId = id and weight =
1)as scienceCode,(select field from tblPrjClassification where prjId = id
and weight = 1) as fieldCode,(select subfield from tblPrjClassification
where prjId = id and weight = 1) as subfieldCode,(select
tblScienceCodes.description from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1))as science_en,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblPrjClassification where prjId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblPrjClassification
where prjId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblPrjClassification where
prjId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblPrjClassification where prjId = id and weight = 1)) as field_en,(select
tblSubfieldCodes.description from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield,(select
tblSubfieldCodes.description_en from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield_en from tblProjects where id
IN( select prjId from tblPrjOfOrg where orgId  = '"+id+"') order by name;
```

Output type: JSON

2.1.1.2 RsrByIdsJson

This function returns profile information of every researcher for which the id is given in the array of strings input parameter.

Input parameter: string[] ids (string of researcher ids)

Query:

```
select id, mstid, prjcoll, pubcoll, firstName, lastName, status, keyws,
keyws_en, tell, fax, email, url,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId) as orgId1,
(select name from tblOrganizations where id IN(select TOP 1
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId)) as orgName1,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId) order by tblRsrIsinOrg.orgId)as orgId2, (select name
from tblOrganizations where id  IN(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId)order by tblRsrIsinOrg.orgId))as orgName2,(select TOP
1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in (select TOP 2
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId) order by
tblRsrIsinOrg.orgId)as orgId2,(select name from tblOrganizations where id
IN(select TOP 1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in
(select TOP 2 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId)order
by tblRsrIsinOrg.orgId))as orgName3,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 2 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId))as orgId3,(select science from tblRsrClassification
where rsrId = id and weight = 1)as scienceCode,(select
tblScienceCodes.description from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblRsrClassification
where rsrId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblRsrClassification
where rsrId = id and weight = 1))as science_en,(select field from
tblRsrClassification where rsrId = id and weight = 1) as fieldCode,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblRsrClassification where rsrId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblRsrClassification
where rsrId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
```

```
tblFieldCodes.scienceId in (select science from tblRsrClassification where
rsrId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblRsrClassification where rsrId = id and weight = 1)) as field_en,(select
subfield from tblRsrClassification where rsrId = id and weight = 1) as
subfieldCode,(select  tblSubfieldCodes.description from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1))as
subfield,(select  tblSubfieldCodes.description_en from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1))as
subfield_en from tblResearchers where
for (int i = 0; i < ids.Length; i++){
                if (i > 0) queryText += " or ";
                queryText += "id = " + ids[i].ToString();
            }
queryText += " order by lastName;";
```

Output type: JSON

2.1.1.3 RsrOnPrjJson

The call returns the profile information of all the researchers who worked on a project specified with the input id.

Input parameter: string id (id of a project)

Query:

```
select id, prjcoll, pubcoll, mstid, firstName, lastName, status, keyws,
keyws_en, tell, fax, email, url,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId) as orgId1,
(select name from tblOrganizations where id IN(select TOP 1
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId)) as orgName1,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId) order by tblRsrIsinOrg.orgId)as orgId2, (select name
from tblOrganizations where id  IN(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId)order by tblRsrIsinOrg.orgId))as orgName2,(select TOP
1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in (select TOP 2
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId) order by
```

```
tblRsrIsinOrg.orgId)as orgId2,(select name from tblOrganizations where id
IN(select TOP 1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in
(select TOP 2 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId)order
by tblRsrIsinOrg.orgId))as orgName3,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 2 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId))as orgId3,(select science from tblRsrClassification
where rsrId = id and weight = 1)as scienceCode,(select
tblScienceCodes.description from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblRsrClassification
where rsrId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblRsrClassification
where rsrId = id and weight = 1))as science_en,(select field from
tblRsrClassification where rsrId = id and weight = 1) as fieldCode,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblRsrClassification where rsrId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblRsrClassification
where rsrId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblRsrClassification where
rsrId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblRsrClassification where rsrId = id and weight = 1)) as field_en,(select
subfield from tblRsrClassification where rsrId = id and weight = 1) as
subfieldCode,(select  tblSubfieldCodes.description from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1))as
subfield,(select  tblSubfieldCodes.description_en from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1))as
subfield_en from tblResearchers where id IN(select rsrId from tblRsrHasPrj
where prjId = '" + id + "');
```

Output type: JSON

2.1.1.4 RsrOnPubJson

The call returns the profile information of all the researchers who collaborated on a publication specified with the input id.

Input parameter: string id (id of a publication)

Query:

```
select id, prjcoll, pubcoll, mstid, firstName, lastName, status, keyws,
keyws_en, tell, fax, email, url,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId) as orgId1,
(select name from tblOrganizations where id IN(select TOP 1
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId)) as orgName1,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId) order by tblRsrIsinOrg.orgId)as orgId2, (select name
from tblOrganizations where id  IN(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId)order by tblRsrIsinOrg.orgId))as orgName2,(select TOP
1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in (select TOP 2
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId) order by
tblRsrIsinOrg.orgId)as orgId3,(select name from tblOrganizations where id
IN(select TOP 1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in
(select TOP 2 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId)order
by tblRsrIsinOrg.orgId))as orgName3,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId)as
orgId,(select science from tblRsrClassification where rsrId = id and
weight = 1)as scienceCode,(select tblScienceCodes.description from
tblScienceCodes where tblScienceCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblRsrClassification
where rsrId = id and weight = 1))as science_en,(select field from
tblRsrClassification where rsrId = id and weight = 1) as fieldCode,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblRsrClassification where rsrId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblRsrClassification
where rsrId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblRsrClassification where
rsrId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblRsrClassification where rsrId = id and weight = 1)) as field_en,(select
subfield from tblRsrClassification where rsrId = id and weight = 1) as
subfieldCode,(select  tblSubfieldCodes.description from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1))as
subfield,(select  tblSubfieldCodes.description_en from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
```

```
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1))as
subfield_en from tblResearchers where mstid in (select mstid from
tblRsrHasPub where cobissId = '"+id+"';
```

Output type: JSON

2.1.1.5 RsrInOrgJson

The call returns the profile information of all the researchers employed in an organization.

Input parameter: string id (id of the organization)

Query:

```
select science from tblPrjClassification where prjId = id and weight =
1)as scienceCode,(select field from tblPrjClassification where prjId = id
and weight = 1) as fieldCode,(select subfield from tblPrjClassification
where prjId = id and weight = 1) as subfieldCode,(select
tblScienceCodes.description from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1))as science_en,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblPrjClassification where prjId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblPrjClassification
where prjId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblPrjClassification where
prjId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblPrjClassification where prjId = id and weight = 1)) as field_en,(select
tblSubfieldCodes.description from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield,(select
tblSubfieldCodes.description_en from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield_en from tblProjects where id
IN( select prjId from tblPrjOfOrg where orgId  = '"+id+"') order by name;
```

Output type: JSON

2.1.1.6 CollaborationOfRsrOnPrj1Json

The call returns the profile information of the researchers that collaborated with a researcher specified with the input id, the collaboration is based on publications

Input parameter: string id (id of a researcher)

Query:

```
select id, prjcoll, pubcoll, mstid, firstName, lastName, status, keyws,
keyws_en, tell, fax, email, url,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId) as orgId1,
(select name from tblOrganizations where id IN(select TOP 1
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId)) as orgName1,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId) order by tblRsrIsinOrg.orgId)as orgId2, (select name
from tblOrganizations where id  IN(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId)order by tblRsrIsinOrg.orgId))as orgName2,(select TOP
1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in (select TOP 2
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId) order by
tblRsrIsinOrg.orgId)as orgId2,(select name from tblOrganizations where id
IN(select TOP 1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in
(select TOP 2 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId)order
by tblRsrIsinOrg.orgId))as orgName3,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId)as
orgId3,(select science from tblRsrClassification where rsrId = id and
weight = 1)as scienceCode,(select tblScienceCodes.description from
tblScienceCodes where tblScienceCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblRsrClassification
where rsrId = id and weight = 1))as science_en,(select field from
tblRsrClassification where rsrId = id and weight = 1) as fieldCode,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblRsrClassification where rsrId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblRsrClassification
where rsrId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblRsrClassification where
rsrId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblRsrClassification where rsrId = id and weight = 1)) as field_en,(select
subfield from tblRsrClassification where rsrId = id and weight = 1) as
subfieldCode,(select  tblSubfieldCodes.description from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1))as
```

```
subfield,(select  tblSubfieldCodes.description_en from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1))as
subfield_en, N from tblResearchers,(select rsrId, count(*) as N from
tblRsrHasPrj where prjId in( SELECT prjId FROM tblRsrHasPrj where rsrId =
'"+id+"') group by rsrId) as t2 where tblResearchers.id = t2.rsrId order
by N desc, CASE rsrId WHEN '"+id+"' THEN 1 ELSE 100 END, lastName;
```

Output type: JSON

2.1.1.7 CollaborationOfRsrOnPrjHeadJson

The call returns the profile information of the researchers that collaborated with a researcher specified with the input id, but it returns only the collaboration between the heads of the projects. The collaboration is based on projects.

Input parameter: string id (id of a researcher)

Query:

```
select *,(select TOP 1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId)as orgId1,(select name from
tblOrganizations where id IN(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId)) as
orgName1,(select TOP 1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in
(select TOP 1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId)
order by tblRsrIsinOrg.orgId) as orgId2, (select name from
tblOrganizations where id  IN(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId)order by tblRsrIsinOrg.orgId))as orgName2,(select name
from tblOrganizations where id  IN(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 2 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId)order by tblRsrIsinOrg.orgId))as orgName3,(select TOP
1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in (select TOP 2
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId))as orgId3,(select
science from tblRsrClassification where rsrId = id and weight = 1)as
scienceCode,(select tblScienceCodes.description from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblRsrClassification
where rsrId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblRsrClassification
where rsrId = id and weight = 1))as science_en,(select field from
```

```
tblRsrClassification where rsrId = id and weight = 1)as fieldCode,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblRsrClassification where rsrId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblRsrClassification
where rsrId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblRsrClassification where
rsrId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblRsrClassification where rsrId = id and weight = 1)) as field_en,(select
subfield from tblRsrClassification where rsrId = id and weight = 1)as
subfieldCode,(select  tblSubfieldCodes.description from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1)) as
subfield,(select  tblSubfieldCodes.description_en from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1)) as
subfield_en from tblResearchers,(select head, count(*) as N from
tblProjects where id in(SELECT prjId FROM tblRsrHasPrj where rsrId =
'"+id+"')  group by head) as t2 where tblResearchers.id = t2.head order by
N desc, CASE id WHEN '"+id+"' THEN 1 ELSE 100 END, lastName;
```

Output type: JSON

2.1.1.8 CollaborationOfRsrOnPubJson

The call returns the profile information of the researchers that collaborated with a researcher specified with the input id, the collaboration is based on publications.

Input parameter: string id (id of a researcher)

Query:

```
select id,prjcoll, pubcoll,tblResearchers.mstid, firstName, lastName,
status, keyws, keyws_en, tell, fax, email, url,(select TOP 1
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId) as orgId1, (select name from tblOrganizations where
id IN(select TOP 1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId)) as orgName1,(select TOP 1
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in (select TOP 1
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId) order by
tblRsrIsinOrg.orgId)as orgId2, (select name from tblOrganizations where id
IN(select TOP 1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in
(select TOP 1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId)order
```

```
by tblRsrIsinOrg.orgId))as orgName2,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 2 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId) order by tblRsrIsinOrg.orgId)as orgId2,(select name
from tblOrganizations where id  IN(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 2 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId)order by tblRsrIsinOrg.orgId))as orgName3,(select TOP
1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in (select TOP 2
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId))as orgId3,(select
science from tblRsrClassification where rsrId = id and weight = 1)as
scienceCode,(select tblScienceCodes.description from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblRsrClassification
where rsrId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblRsrClassification
where rsrId = id and weight = 1))as science_en,(select field from
tblRsrClassification where rsrId = id and weight = 1) as fieldCode,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblRsrClassification where rsrId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblRsrClassification
where rsrId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblRsrClassification where
rsrId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblRsrClassification where rsrId = id and weight = 1)) as field_en,(select
subfield from tblRsrClassification where rsrId = id and weight = 1) as
subfieldCode,(select  tblSubfieldCodes.description from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1))as
subfield,(select  tblSubfieldCodes.description_en from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1))as
subfield_en, N from tblResearchers,(select tblRsrHasPub.mstid, count(*) as
N from tblRsrHasPub where cobissId in(SELECT cobissId FROM tblRsrHasPub
where tblRsrHasPub.mstid in (select mstid from tblResearchers where
id='"+id+"')) group by tblRsrHasPub.mstid)as t2 where tblResearchers.mstid
= t2.mstid ORDER BY N desc,CASE id WHEN '"+id+"' THEN 1 ELSE 100 END,
lastName;
```

Output type: JSON

2.1.2 Deploying projects data

2.1.2.1 PrjOfOrgJson

The call returns all the projects of an organization.

Input parameter: string id (id of a organization)

Query:

```
select *,(select science from tblPrjClassification where prjId = id and
weight = 1)as scienceCode,(select field from tblPrjClassification where
prjId = id and weight = 1) as fieldCode,(select subfield from
tblPrjClassification where prjId = id and weight = 1) as
subfieldCode,(select tblScienceCodes.description from tblScienceCodes
where tblScienceCodes.scienceId in(select science from
tblPrjClassification where prjId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1))as science_en,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblPrjClassification where prjId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblPrjClassification
where prjId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblPrjClassification where
prjId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblPrjClassification where prjId = id and weight = 1)) as field_en,(select
tblSubfieldCodes.description from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield,(select
tblSubfieldCodes.description_en from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield_en from tblProjects where id
IN( select prjId from tblPrjOfOrg where orgId  = '"+id+"') order by name;
```

Output type: JSON

2.1.2.2 PrjOfRsrJson

The call returns all the projects of a researcher.

Input parameter: string id (id of a researcher)

Query:

```
select *,(select science from tblPrjClassification where prjId = id and
weight = 1)as scienceCode,(select field from tblPrjClassification where
prjId = id and weight = 1) as fieldCode,(select subfield from
tblPrjClassification where prjId = id and weight = 1) as
subfieldCode,(select tblScienceCodes.description from tblScienceCodes
where tblScienceCodes.scienceId in(select science from
tblPrjClassification where prjId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1))as science_en,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblPrjClassification where prjId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblPrjClassification
where prjId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblPrjClassification where
prjId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblPrjClassification where prjId = id and weight = 1)) as field_en,(select
tblSubfieldCodes.description from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield,(select
tblSubfieldCodes.description_en from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield_en from tblProjects where id
IN( select prjId from tblRsrHasPrj where rsrId IN( select id from
tblResearchers where id = '"+id+"'));
```

Output type: JSON

2.1.2.3 PrjAll

The call returns the ids of all the projects in the database.

Input parameter: string id (id of a researcher)

Query:

```
select id from tblProjects;
```

Output type: XML

2.1.2.4 PrjByRsrIdsJson

This call returns all the projects of a set of researchers. Set of researchers is an input parameter in the form of array of ids. This ids are putt into SQL query using a loop that combines them with OR logical statement.

Input parameter: string[] id (array of researcher ids)

Query:

```
select *,(select science from tblPrjClassification where prjId = id and
weight = 1)as scienceCode,(select field from tblPrjClassification where
prjId = id and weight = 1) as fieldCode,(select subfield from
tblPrjClassification where prjId = id and weight = 1) as
subfieldCode,(select tblScienceCodes.description from tblScienceCodes
where tblScienceCodes.scienceId in(select science from
tblPrjClassification where prjId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1))as science_en,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblPrjClassification where prjId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblPrjClassification
where prjId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblPrjClassification where
prjId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblPrjClassification where prjId = id and weight = 1)) as field_en,(select
tblSubfieldCodes.description from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield,(select
tblSubfieldCodes.description_en from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield_en from tblProjects where id
IN( select prjId from tblRsrHasPrj where rsrId IN( select id from
tblResearchers where "

for (int i = 0; i < ids.Length; i++)

            {

                if (i > 0) queryText += " or ";

                queryText += "id = " + ids[i].ToString();

            }

            queryText += ") ) order by name;";
```

Output type: JSON

2.1.2.5 PrjByIdJson

The call returns the information about the project specified by the project id.

Input parameter: string id (id of a project)

Query:

```
select *,(select science from tblPrjClassification where prjId = id and
weight = 1)as scienceCode,(select field from tblPrjClassification where
prjId = id and weight = 1) as fieldCode,(select subfield from
tblPrjClassification where prjId = id and weight = 1) as
subfieldCode,(select tblScienceCodes.description from tblScienceCodes
where tblScienceCodes.scienceId in(select science from
tblPrjClassification where prjId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1))as science_en,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblPrjClassification where prjId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblPrjClassification
where prjId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblPrjClassification where
prjId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblPrjClassification where prjId = id and weight = 1)) as field_en,(select
tblSubfieldCodes.description from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield,(select
tblSubfieldCodes.description_en from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield_en from tblProjects where id
= '" + ids + "'"
```

Output type: JSON

2.1.2.6 PrjByIdsJson

This call returns projects specified by the array of project ids given as the input parameter. This ids are putt into SQL query using a loop that combines them with OR logical statement.

Input parameter: string[] id (ids of projects)

Query:

```
select *,(select science from tblPrjClassification where prjId = id and
weight = 1)as scienceCode,(select field from tblPrjClassification where
prjId = id and weight = 1) as fieldCode,(select subfield from
tblPrjClassification where prjId = id and weight = 1) as
subfieldCode,(select tblScienceCodes.description from tblScienceCodes
where tblScienceCodes.scienceId in(select science from
```

```
tblPrjClassification where prjId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1))as science_en,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblPrjClassification where prjId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblPrjClassification
where prjId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblPrjClassification where
prjId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblPrjClassification where prjId = id and weight = 1)) as field_en,(select
tblSubfieldCodes.description from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield,(select
tblSubfieldCodes.description_en from tblSubfieldCodes where
tblSubfieldCodes.scienceId in(select science from tblPrjClassification
where prjId = id and weight = 1) and tblSubfieldCodes.fieldId in (select
field from tblPrjClassification where prjId = id and weight = 1) and
tblSubfieldCodes.subfieldId in ( select subfield from tblPrjClassification
where prjId = id and weight = 1))as subfield_en from tblProjects where "
for (int i = 0; i < ids.Length; i++)

      {

         if (i > 0) queryText += " or ";

         queryText += "id = " + ids[i].ToString();

      }

      queryText += " order by name;";
```

Output type: JSON

2.1.3 Publications

2.1.3.1 PublicationsOfRsrJson

This call returns all the publications of a researcher specified by the researcher id.

Input parameter: string id (id of a researcher)

Query:

```
select id, mstid, prjcoll, pubcoll, firstName, lastName, status, keyws,
keyws_en, tell, fax, email, url,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId) as orgId1,
(select name from tblOrganizations where id IN(select TOP 1
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId)) as orgName1,(select TOP 1 tblRsrIsinOrg.orgId from
```

```
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId) order by tblRsrIsinOrg.orgId)as orgId2, (select name
from tblOrganizations where id  IN(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId)order by tblRsrIsinOrg.orgId))as orgName2,(select TOP
1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in (select TOP 2
tblRsrIsinOrg.orgId from tblRsrIsinOrg where tblResearchers.id =
tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId) order by
tblRsrIsinOrg.orgId)as orgId2,(select name from tblOrganizations where id
IN(select TOP 1 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId and tblRsrIsinOrg.orgId not in
(select TOP 2 tblRsrIsinOrg.orgId from tblRsrIsinOrg where
tblResearchers.id = tblRsrIsinOrg.rsrId order by tblRsrIsinOrg.orgId)order
by tblRsrIsinOrg.orgId))as orgName3,(select TOP 1 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId and
tblRsrIsinOrg.orgId not in (select TOP 2 tblRsrIsinOrg.orgId from
tblRsrIsinOrg where tblResearchers.id = tblRsrIsinOrg.rsrId order by
tblRsrIsinOrg.orgId))as orgId3,(select science from tblRsrClassification
where rsrId = id and weight = 1)as scienceCode,(select
tblScienceCodes.description from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblRsrClassification
where rsrId = id and weight = 1))as science,(select
tblScienceCodes.description_en from tblScienceCodes where
tblScienceCodes.scienceId in(select science from tblRsrClassification
where rsrId = id and weight = 1))as science_en,(select field from
tblRsrClassification where rsrId = id and weight = 1) as fieldCode,(select
tblFieldCodes.description from tblFieldCodes where tblFieldCodes.scienceId
in (select science from tblRsrClassification where rsrId = id and weight =
1) and tblFieldCodes.fieldId in (select field from tblRsrClassification
where rsrId = id and weight = 1)) as field,(select
tblFieldCodes.description_en from tblFieldCodes where
tblFieldCodes.scienceId in (select science from tblRsrClassification where
rsrId = id and weight = 1) and tblFieldCodes.fieldId in (select field from
tblRsrClassification where rsrId = id and weight = 1)) as field_en,(select
subfield from tblRsrClassification where rsrId = id and weight = 1) as
subfieldCode,(select  tblSubfieldCodes.description from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1))as
subfield,(select  tblSubfieldCodes.description_en from tblSubfieldCodes
where tblSubfieldCodes.scienceId in(select science from
tblRsrClassification where rsrId = id and weight = 1) and
tblSubfieldCodes.fieldId in (select field from tblRsrClassification where
rsrId = id and weight = 1) and tblSubfieldCodes.subfieldId in ( select
subfield from tblRsrClassification where rsrId = id and weight = 1))as
subfield_en from tblResearchers where id = '" + id + "';
```

Output type: JSON

2.1.3.2 PubById

This call returns the publication specified by the id of the publication.

Input parameter: string id (id of a publication)

Query:

```
select * from tblPublications where cobissId = '" + id + "';
```

Output type: JSON

2.1.4 Organizations

2.1.4.1 OrgByRsrIdsJson

This call returns all the organization of researchers specified by the array of researcher ids, specified by the input parameter. The ids are putt into SQL query using a loop that combines them with OR logical statement.

Input parameter: string[] id (ids of researchers)

Query:

```
select * from tblOrganizations,(select orgId, count(*) as N from
tblRsrIsinOrg where rsrId IN(select rsrId from tblRsrIsinOrg where ";

            for (int i = 0; i < ids.Length; i++)

            {

                if (i > 0) queryText += " or ";

                queryText += "rsrId = " + ids[i].ToString();

            }

            queryText += " ) group by orgId) as t2 where
tblOrganizations.id = t2.orgId;";
```

Output type: JSON

2.1.4.2 OrgByIdJson

This call returns projects specified by the array of project ids given as the input parameter. This ids are putt into SQL query using a loop that combines them with OR logical statement.

Input parameter: string[] id (ids of organizations)

Query:

```
select * from tblOrganizations where id = " + ids;
```

Output type: JSON

2.1.5 Lectures

2.1.5.1 LecOfRsrJson

This call returns all the lectures of a researcher specified by the researcher id input parameter

Input parameter: string id (id of a researcher)

Query:

```
select *, (select tblRsrHasLec.role from tblRsrHasLec where
tblLectures.url = tblRsrHasLec.lecUrl and tblRsrHasLec.rsrId = '"+id+"' )
as role from tblLectures where url IN(select lecUrl from tblRsrHasLec
where rsrId IN( select id from tblResearchers where id = '"+id+"')) order
by title;
```

Output type: JSON

2.2 Server side application functionalities

The function of the server side module is to perform demanding computations that include data retrieval and manipulation. This module enables the interaction between the client application from where the requests are formed and web service module where data is obtained from the database in a standardized form. The server side application makes all the computations that are needed for the web client (like calculating the coordinates of the data points, etc.), and also the transformations of the data into a format the web client can use without the need for additional processing on the client side. The main functions of the server side module of the application are presented in this chapter.

```
private string parseJsonCircle(string data)
```

The function takes the array of researchers in the form of JSON string as an input and computes the coordinates of the circular layout for the researcher's collaboration diagram. The first researcher in the array is always the central researcher. The data contains the information about the number of collaborations (projects or publications) each researcher has with central one, which is used for the computing coordinates. The output of the function is the array of researchers enriched with the coordinates for the circular layout.

```
private string[] parseJsonAlltoAllClosedOrg(string data1, string data2, string data3)
```

The function generates the graph (pair of nodes defined by ids) of organizations which are connected based on the collaboration. The input of the function is the array or organizations which are included in the set of nodes for which the graph is constructed. The output is an array of string, where first element of the array is the JSON string containing the coordinates of the nodes and the second is the JSON string containing the edge (pairs of nodes) and the weights between them.

```
private string parseJsonAlltoallClosedHead(string data)
```

The function returns a JSON string containing the edges (pairs of nodes) between the set or researchers given with the input JSON string. The edge between the two researchers from the set is constructed if two researchers collaborated on a project.

```
private string parseJsonAlltoallStringPubClosed
```

The function returns a JSON string containing the edges (pairs of nodes) between the set or researchers given with the input JSON string. The edge between the two researchers from the set is constructed if they collaborated on a publication.

```
private string parseJsonAlltoallClosed(string data)
```

The function generates graph with all the connections between nodes that collaborate with a chosen node. Ids of edges are array positions of researchers.

```
private string searchRsr(string data), private string searchPrj(string data)
private string searchPub(string data), private string searchOrg(string data)
private string searchOrgFP7(string data)
```

The functions are implementations of searches that are implemented with indexes.

3 Web client

Web client is a part of application that initializes on the clients device. Web client is an interface between the user of the system and the rest of the application. The main part of web client includes the web interface (front-end of the application). Elements of user interface; like buttons and input boxes represent an interface using which the user can form his requests.

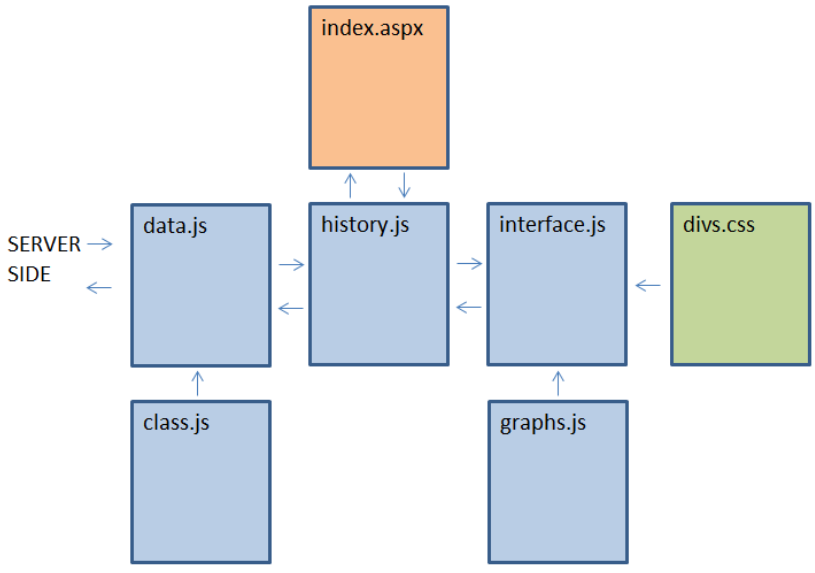


Figure 2. Architecture of the web client

These request are transformed into calls that are passed to the server side application. Once the server side receives the requests, it performs the operations needed to obtain, transform and send the data back to the client. The data returned to client is the format easily usable on the client side, without the need for computationally costly operations. The web client then uses the data to display it in the form useful to the user. Slovenian Science Atlas web client displays the data to the user in the graphical way using the visualizations.

3.1 Architecture of the web client

Figure 2 shows the architecture of the web client. The modules of the architecture are represented with the names of the main files that perform their functionalities. Main functionalities of the modules of the architecture are described in this chapter.

3.1.1 User interface

index.aspx is the file that the user is viewing. This is the part of application which is on the highest level of the system. The user views and interacts with the HTML code generated in this single web page document. The part of the user interface module is the divs.css file. This file contains the CSS stylish of the user interface.

3.1.2 History recording module

Below the index.aspx is the history.js file. This file represents the module that records all the actions of the user interactions with the system. This enables the standard back and forward navigation in the web application, even thou only one HTML (.aspx) page is used in the system. The JS lugin that was used for implementing this kind of navigation is the jQuery BBQ: Back Button & Query Library - v1.2.1 - 2/17/2010 (http://benalman.com/projects/jquery-bbq-plugin/).

3.1.3 Data manipulation module

After each request is recorded using the history.js, it is moved to the data.js file which represent the data manipulation module of the web client. Data manipulation module transforms the user's actions into request that are passed to the server side of the application. This module is also responsible for retrieving the data and transforming it into form suitable for displaying to the user.

3.1.4 Data display module

The obtained data has to be somehow displayed to the user, this is the function of the data display module that is represented with the interface.js file. This file contains all the function that performs actions on the user interface. The result of all this actions is directly visible in the index.aspx file. Some of this actions include displaying the lists of researchers, projects, organizations, publications and videos in paginated windows, attaching actions to the buttons and input boxes, changing display properties of various html elements etc. Very important functionality of this module is the generation and display of visualizations. The visualizations are implemented in the file called graph.js. The visualizations are generated using the Sigma.js (http://sigmajs.org/) HTML 5 plugin.

4 Conclusion

The report covers the main functionalities of the server side of the application and the web client. The server side of the application is responsible for obtaining the data from the database and performing transformations and other computationally demanding functionalities of the system. In order to be able to access the data layer in the standardized way, independent from the rest of the system, the data deployment is implemented using the web service, which is a separate module of the server side application.

Web Client is responsible for the interaction between the user and the system. Main priorities of this layer are: usability, user experience, visual appearance, speed and accuracy of response. The web client is implemented with different modules – visualizations generation, user interface manipulation, information displaying, history recording, data transfer and communication with other layers.

Atlas slovenske znanosti

RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

R52 – Testing of the system

Ljubljana, 15.3.2013

Table of content

1 Introduction

In this report, results of the testing of the Slovenian Science Atlas system are represented. The load testing method was applied. This method enables us analyzing the behavior of the system under different number of simulated users. We can examine different performance parameters (for example page load times, number of failures, bandwidth, etc.) in accordance with time and different number of users.

The load testing report is structured in the way that each performance parameter is firstly described; following by the chart show the parameter for Slovenian Science Atlas system. The report also contains few tables that summarize more performance parameters. The testing result reporting is divided into two sections – user level analysis and time based analysis.

2 Load testing

In order to test the systems performance, under different load conditions (with different number of users), load testing was performed. Using these tests, maximum operating capacity can be determined. Also, using load tests, elements which cause degradation of performance (bottlenecks) can be identified.

2.1 Test case summary

The summary of the test case contains high level metrics – duration of the test case recording, number of pages loaded, number of URLs, number of images, total size, average page size, total image size and average image size.

Table 1. Test case summary

Total Duration	01:07.89
Pages	2
URLs	28
Images	13
Total size	157.0 KB
Average page size	78.5 KB
Total image size	55.9 KB
Average image size	4.7 KB

2.1.1 Load test summary

Based on test case recording the Load Tests can be created. Load tests can be configured with different settings – starting number of users of the webpage, the number of new users that will be

added, the way and frequency of adding new users, the duration of the test. The load test configurations were as shown in the Table 2.

Table 2. Load test configuration

Plan duration	30 min
Peak users	19
Sample period	10 sec
Starting users	5
Users peer ramp	1
Do not exceed	25

In the table 2 are the summary facts of the performed load test.

Table 3. Load test summary

Start Time	12:40 PM 3/6/13
Duration	00:30:12
Completed Pages	2,452
Total hits	68,653
Peak hits/sec	82.2
Peak transfer speed	3.2 Mbps
Peak cases/min	180.0
Total Pages Failed	3

The table shows that the duration of the test was 30 min during which 2,452 pages were successfully completed The hit/sec ratio was at the peak level as high as 82.2, meaning that at most 82 hits on page in one second were simulated for the purpose of this test.

2.2 User level analysis

2.2.1 Page duration

The Page Duration chart shows the minimum, maximum and average page duration for all pages in the test that completed during the sample periods summarized for each user level. Note that the page duration includes the time required to retrieve all resources for the page from the server. It includes network transmission time but not browser rendering time. In a well-performing system, the page durations should remain below the required limits up to or beyond the required load (number of users), subject to the performance requirements set forth for the system. Figure 1 show the page duration.

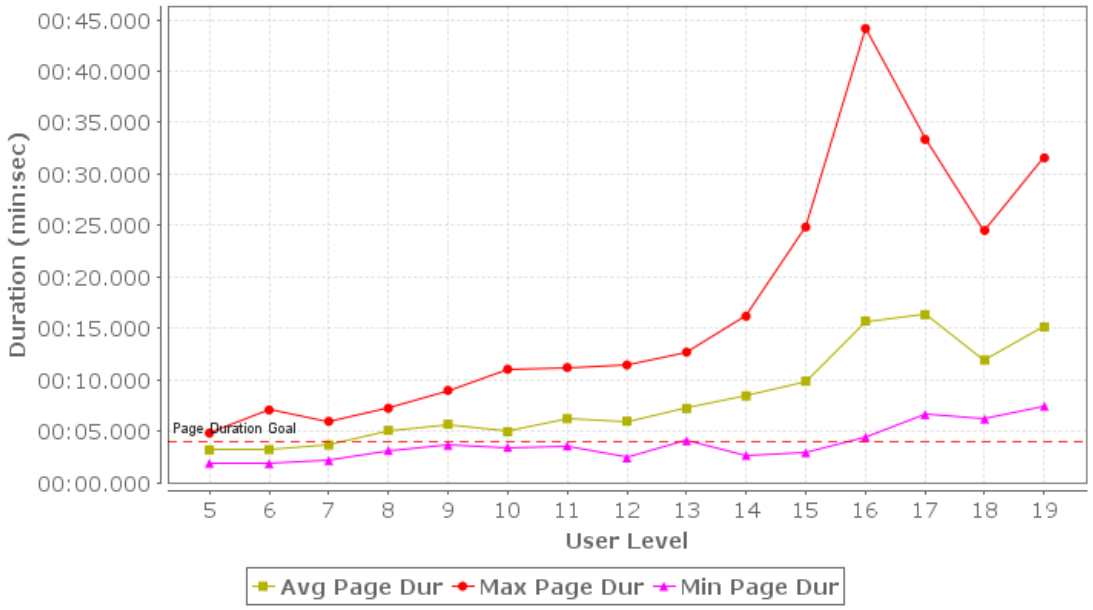


Figure 1. Page duration chart

The chart shows how the page duration grows with the increase of the number of users. While the minimum page duration stays stable, the maximum page duration grows excessively after the number of users increases above 14. This is caused by the extremes cases of requests that require huge amounts of data from the database. Since the number of this kind of requests is limited, the average page duration stays acceptable.

2.2.2 Page completion rate

The Page Completion Rate chart shows the total number of pages completed per second during the sample periods summarized for each user level. In a well-performing system, this number should scale linearly with the applied load (number of users). Figure 2 shows the page completion rate.

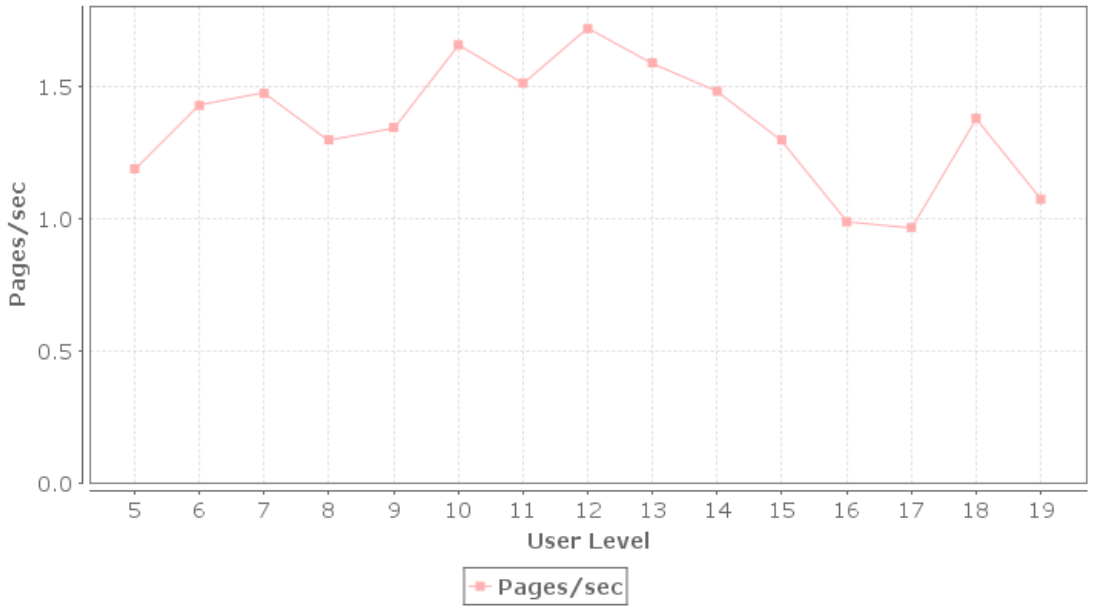


Figure 2. Page completion rate

From Figure 2 we can see that the page completion rate stays stable or drops sub linearly with the growth users, which is very good performance result.

2.2.3 Transaction (URL) Completion Rate

The Transaction (URL) Completion Rate chart shows the total number of HTTP transactions (URLs) completed per second during the sample periods summarized for each user level. In a well-performing system, this number should scale linearly with the applied load (number of users). The figure 3 shows the transaction completion chart.

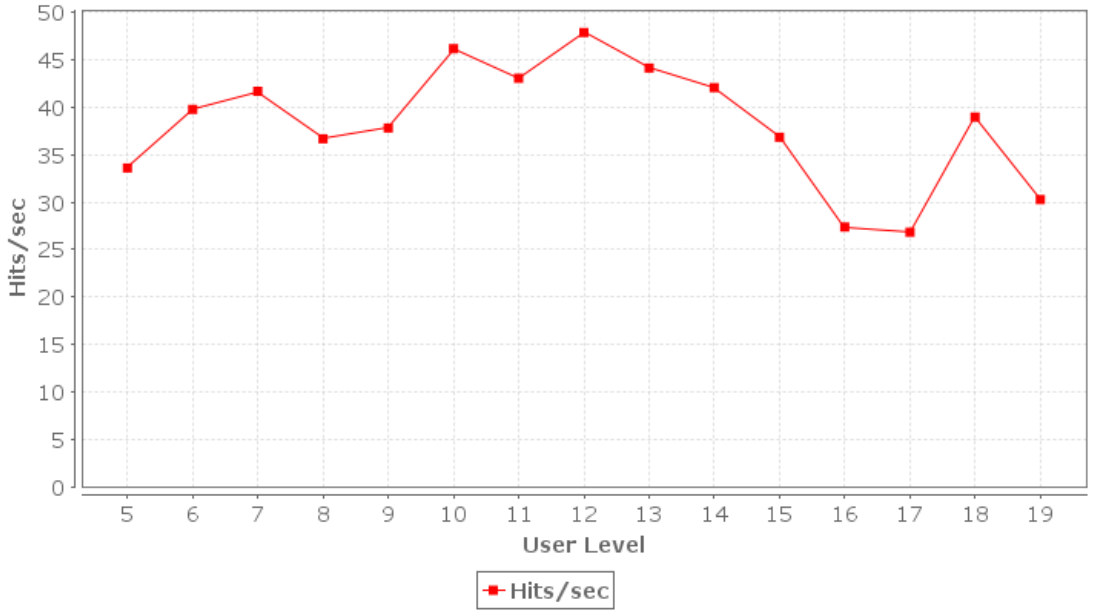


Figure 3. Transaction completion rate

Very similarly to the page completion rate, the transaction completion rate stays stable with the growth of the number of users. The hits peer second rate stays between 25 and 50.

2.2.4 Bandwidth

The Bandwidth chart shows the total bandwidth consumed by traffic generated directly by the load test engines during the sample periods summarized for each user level. In a system that is not constrained by bandwidth, this number should scale linearly with the applied load (number of users). The bandwidth consumption is described in terms of the servers; i.e. outgoing bandwidth refers to data sent by the server to the browser.

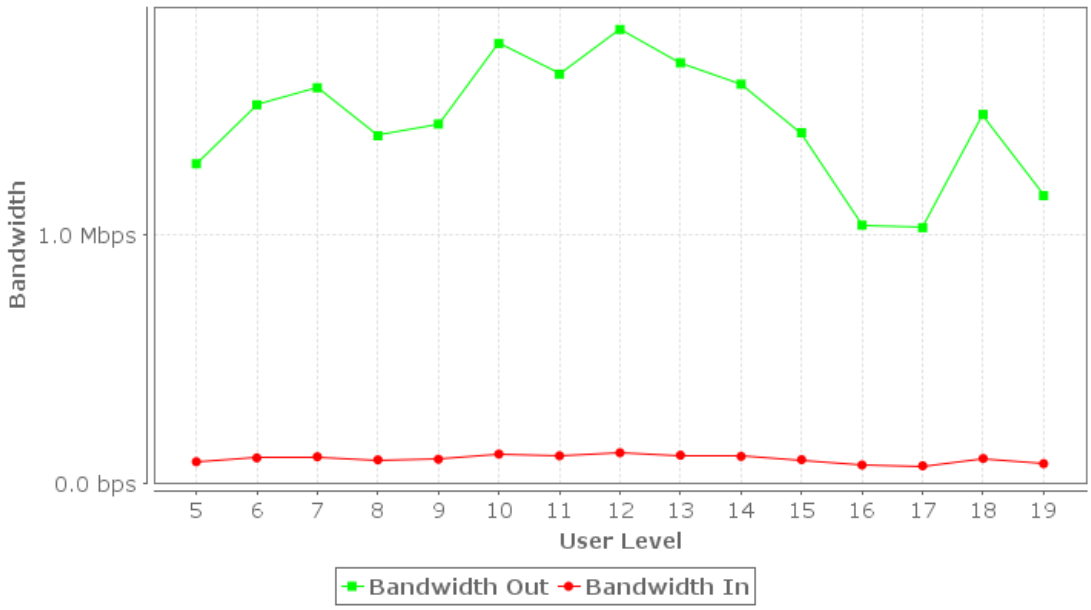


Figure 4. Bandwidth chart

Outgoing bandwidth of the system stays above 1 Mbps, but does not have significant changes correlated with the number of users.

2.2.5 Waiting Users and Average Wait Time

The Waiting Users and Average Wait Time metrics help diagnose certain types of performance problems. For example, they can help determine what pages users have stopped on when a server becomes non-responsive. The 'Waiting Users' metric counts the number of users waiting to complete a web page at the end of the sample periods summarized for each user level. The 'Average Wait Time' describes the amount of time, on average, that each of those users has been waiting to complete the page. Figure 6 shows the waiting users and average wait time metrics.

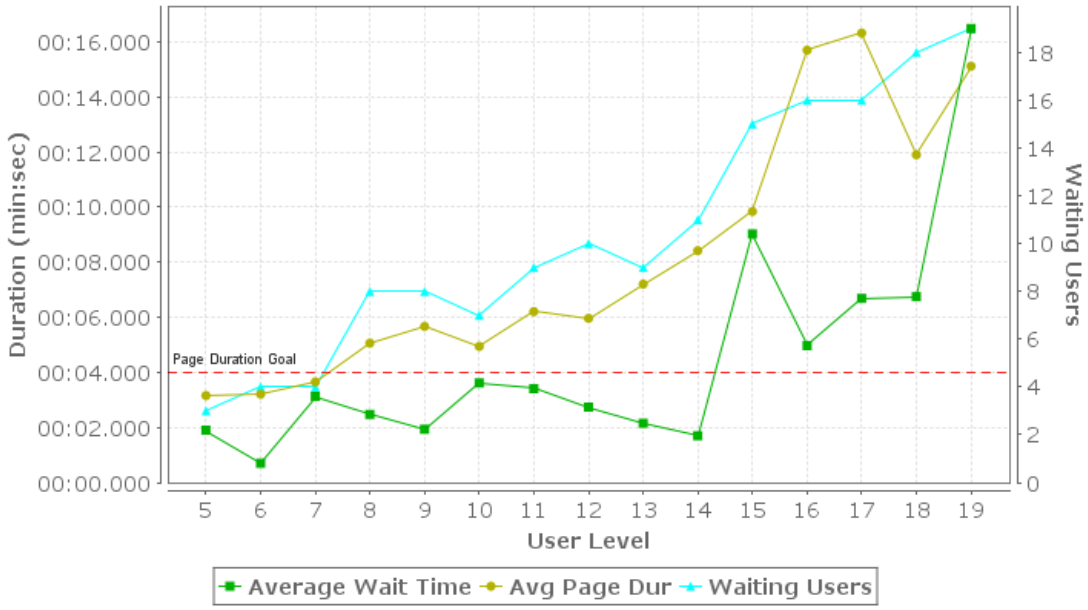


Figure 5. Waiting Users and Average Wait Time

The average wait time experiences growth peaks after the number of users grows above 14. The average page duration and number of waiting users grows linearly with the number of users.

2.2.6 Key metrics

The table 4 shows some of the key metrics that reflect the performance of the test as a whole, summarized by the selected user levels.

Table 4. Summarized by the selected user levels, this table shows some of the key metrics that reflect the performance of the test as a whole

User Level	Pages/sec	Page Failure Rate	Hits/sec	Bandwidth Out	Min Page Dur (ms)	Avg Page Dur (ms)	Max Page Dur (ms)	Waiting Users	Average Wait Time
5	1.2	0	33.6	161033	1869	3157	4804	3	1897
6	1.4	0	39.8	190826	1838	3205	7049	4	698
7	1.5	0	41.7	199461	2130	3650	5942	4	3125
8	1.3	0	36.7	175241	3115	5061	7209	8	2473
9	1.3	0	37.8	181174	3623	5667	8944	8	1934
10	1.7	0	46.2	221869	3427	4939	11005	7	3613
11	1.5	0	43.1	206289	3575	6202	11171	9	3408
12	1.7	0	47.9	228898	2395	5957	11426	10	2735
13	1.6	0	44.1	212073	4059	7201	12667	9	2150
14	1.5	0	42.1	201239	2628	8417	16167	11	1696
15	1.3	0	36.8	176756	2905	9841	24832	15	9020

16	1	0.0111	27.4	130088	4385	15697	44141	16	4972
17	1	0.0085	26.8	128845	6588	16318	33360	16	6687
18	1.4	0	39	185964	6170	11896	24506	18	6749
19	1.1	0.0077	30.3	145007	7418	15116	31560	19	16483

2.3 Time-based Analysis

2.3.1 Page Duration

The Page Duration chart shows the minimum, maximum and average page duration for all pages in the test relative to the elapsed test time (sample period) in which they completed. Note that the page duration includes the time required to retrieve all resources for the page from the server. It includes network transmission time but not browser rendering time. In a well-performing system, the page durations should remain below the required limits up to or beyond the required load (number of users), subject to the performance requirements set forth for the system. Figure 7 shows the page duration chart.

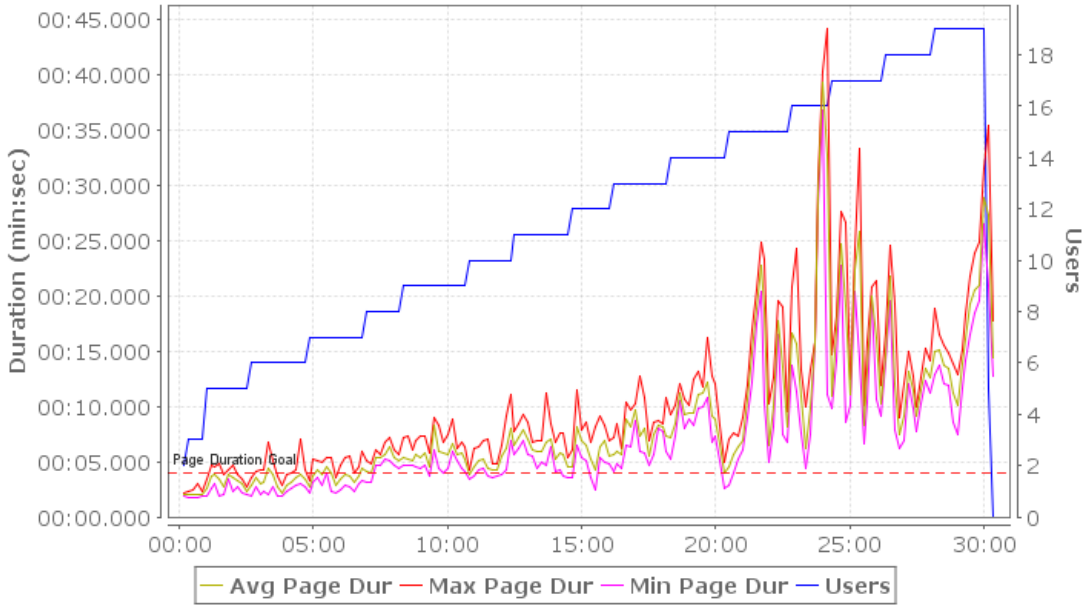


Figure 6. Page duration chart

The page duration grows linearly with some peaks emerging around 25th minute of the performance testing period.

2.3.2 Page completion rate

The Page Completion Rate chart shows the total number of pages completed per second relative to the elapsed test time (sample period) in which they completed. In a well-performing system, this number should scale linearly with the applied load (number of users). Figure 8 shows the page completion rate chart.

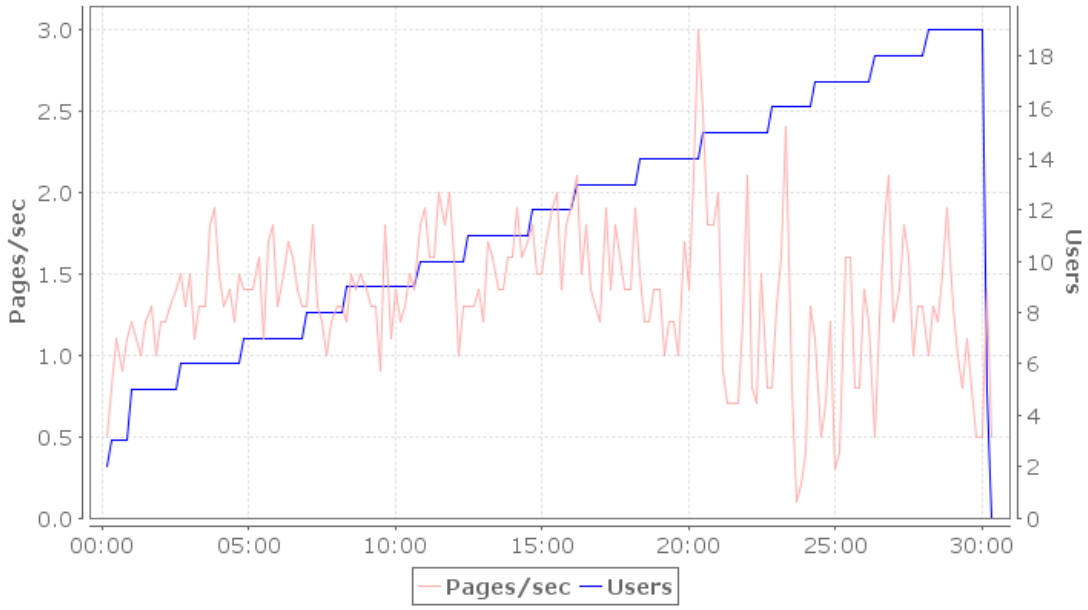


Figure 7. Page completion rate

The page completion rates contains many peaks, but does not very too much with the time.

2.3.3 Transaction (URL) Completion Rate

The Transaction (URL) Completion Rate chart shows the total number of HTTP transactions (URLs) completed per second relative to the elapsed test time (sample period) in which they completed. In a well-performing system, this number should scale linearly with the applied load (number of users). Figure 9 shows the page transaction completion rate chart.

Figure 8. Transaction completion rate

Similarly to the page completion rate chart, the Figure 9 shows the similar behavior for the transaction completion rate. The rate contains peaks, but is overall stable with the time.

2.3.4 Failures

The failures section chart illustrates how the total number of page failures and the page failure rate changed throughout the test relative to the elapsed test time (sample period) in which they occurred. A page can fail for any number of reasons, including failures in the network and servers (web, application or database). See the Failures section of the report for details on the page failures encountered. In a well-performing system, this number should be zero. Figure 10 shows the failures section chart.

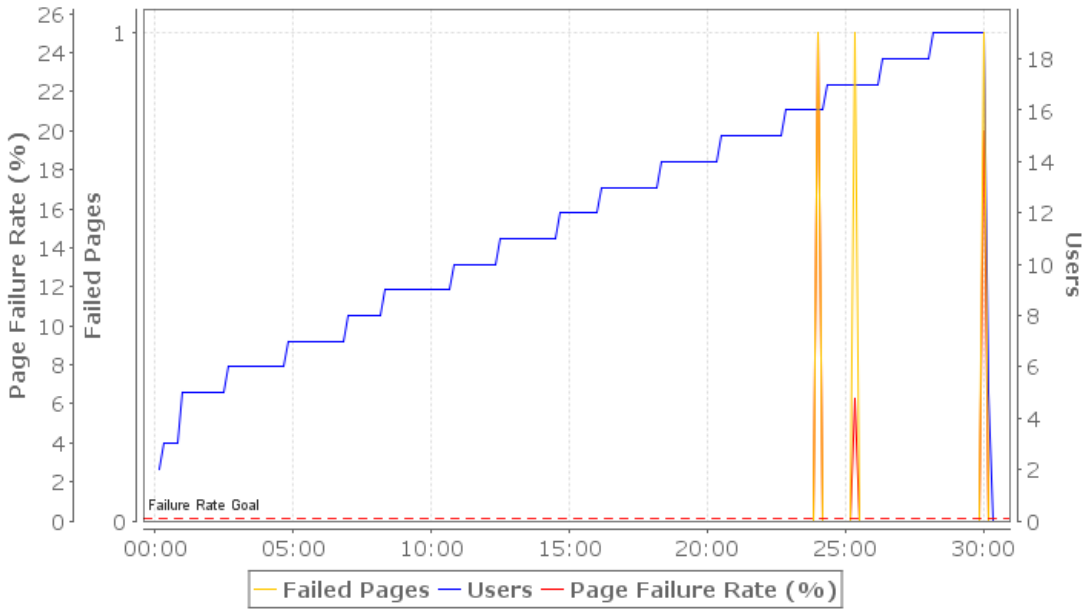


Figure 9. Failure section chart

From the chart on Figure 10, few emergences of failures can be seen, emerging at the end of the load testing period (around 25th and 30th minute).

2.3.5 Bandwidth

The Bandwidth chart shows the total bandwidth consumed by traffic generated directly by the load test engines throughout the test relative to the elapsed test time (sample period). In a system that is not constrained by bandwidth, this number should scale linearly with the applied load (number of users). Note that other sources of bandwidth may be active during a test and may even be caused indirectly by the load test but may not be included in this metric. The bandwidth consumption is described in terms of the servers; i.e. outgoing bandwidth refers to data sent by the server to the browser. Figure 11 shows the bandwidth chart.

Figure 10. Bandwidth chart

The Testcase Completion Rate chart shows the total number of testcases completed per minute relative to the elapsed test time (sample period) in which they completed. In a well-performing system, this number should scale linearly with the applied load (number of users). Figure 12 shows the page completion rate chart.

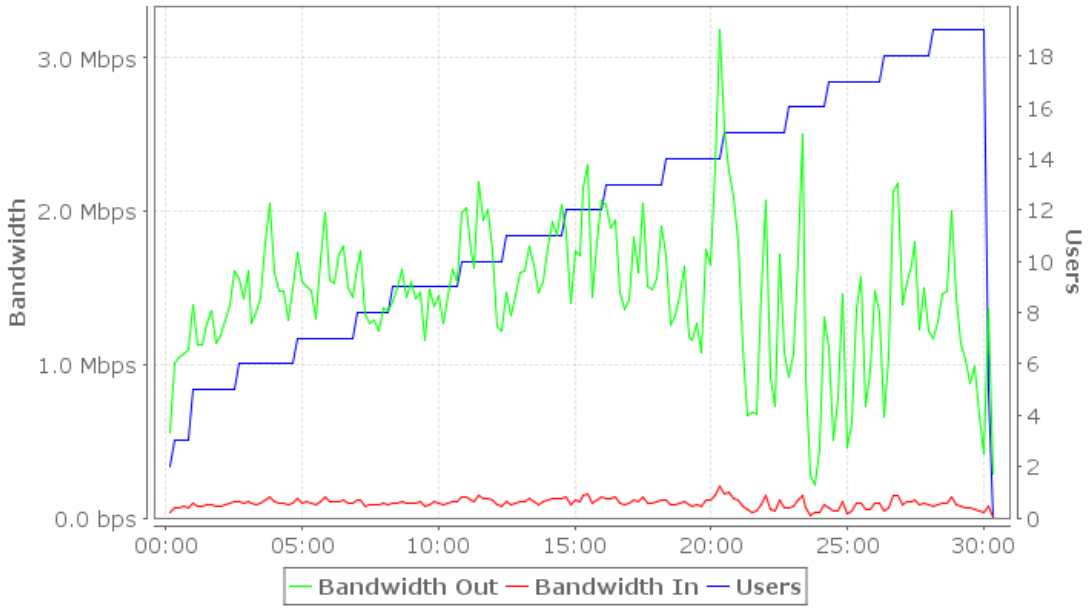


Figure 11. Completion rate

The input bandwidth stays low and stable during the whole testing period. The outgoing bandwidth stays mostly between 1 and 2 Mbps.

2.3.6 Waiting Users and Average Wait Time

The Waiting Users and Average Wait Time metrics help diagnose certain types of performance problems. For example, they can help determine what pages users have stopped on when a server becomes non-responsive. The 'Waiting Users' metric counts the number of users waiting to complete a web page at the end of the sample period. The 'Average Wait Time' describes the amount of time, on average, that each of those users has been waiting to complete the page. Figure 13 shows the page completion rate chart.

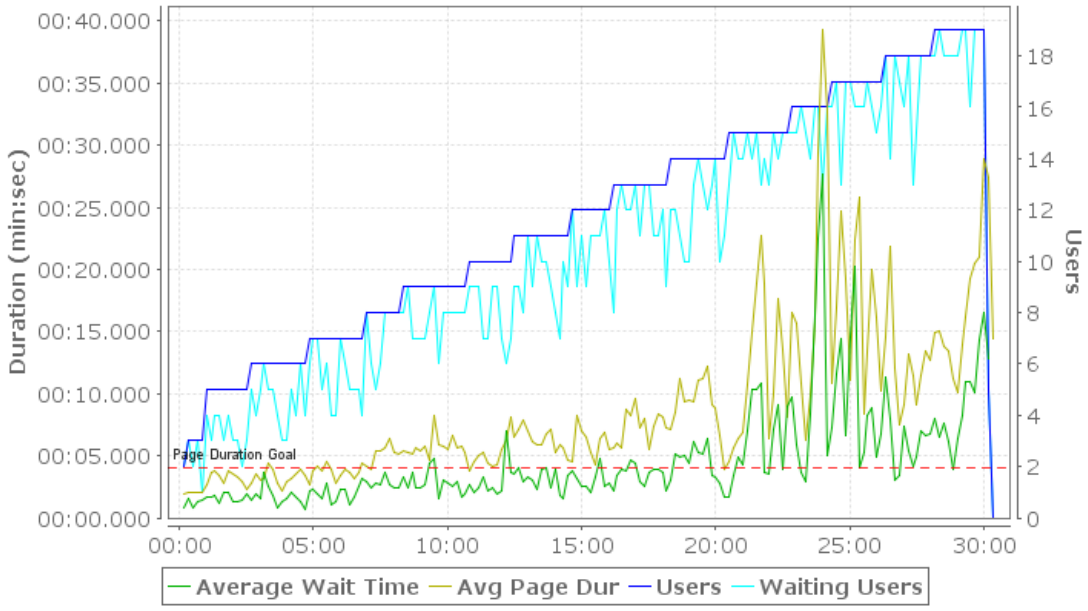


Figure 12. Waiting Users and Average Wait Time

The number of waiting users grows linearly in time. Average waiting time and average page duration were stable in most of the testing period, but experienced peaks and growth after 20th minute.

2.3.7 Test summary metrics

Sorted by the elapsed test time, this table shows some of the key metrics that reflect the performance of the test as a whole.

Table 5. Test summary metrics

Elapsed time (ms)	Users	Pages/sec	Page Failure Rate	Hits/sec	Bandwidth Out	Min Page Dur (ms)	Avg Page Dur (ms)	Max Page Dur (ms)	Waiting Users	Average Wait Time
10000	2	0.5	0	14.2	69764	1811	1971	2196	2	761
20000	3	0.8	0	26.9	125657	1800	1991	2291	3	1509
30000	3	1.1	0	26.9	130520	1779	2020	2427	2	794
100000	5	1.3	0	34.4	168243	2001	2736	3861	3	2089
200000	6	1.3	0	34.2	167764	2077	4383	6815	5	2594
300000	7	1.4	0	39.4	192601	3114	3757	5163	7	2283
1000000	13	1.4	0	37.8	183125	6456	8771	10391	12	3718
1010000	13	1.3	0	34.3	170179	6376	8201	9626	12	4593
1020000	13	1.2	0	36.4	176942	8703	9610	10243	13	4433
1030000	13	1.9	0	48.3	229453	5984	7273	12667	11	2883

1040000	13	1.4	0	41.9	200243	5768	8046	10726	13	2598
1050000	13	1.8	0	52.5	255973	4651	5494	6875	13	3604
1500000	17	0.3	0	10.4	57484	10145	11030	12446	17	14549
1800000	19	0.5	0.2	13.1	51817	26577	28882	31560	19	16483
1800000	19	0.5	0.2	13.1	51817	26577	28882	31560	19	16483

3 Conclusion

This report shows the results of the Slovenian Science Atlas load testing. With this testing we can see the inner picture of the system, so we know how it behaves under different circumstances, regarding to page completion, page failure, page load duration, user waiting, bandwidth, pages peer second rates and other important parameters for the system hosted as a web application. The system experiences above average maximum page duration times, but that is somehow expected, because of the nature of some web service calls (for some examples huge amounts of data are obtained from the server). The average page duration times are acceptable and the minimum page duration times are very low, what makes this application In the line with the requirements. From all the other aspects the application performs very well, since there are no cases where the increase in time or number of users results in a super linear growth of some parameter like delay.

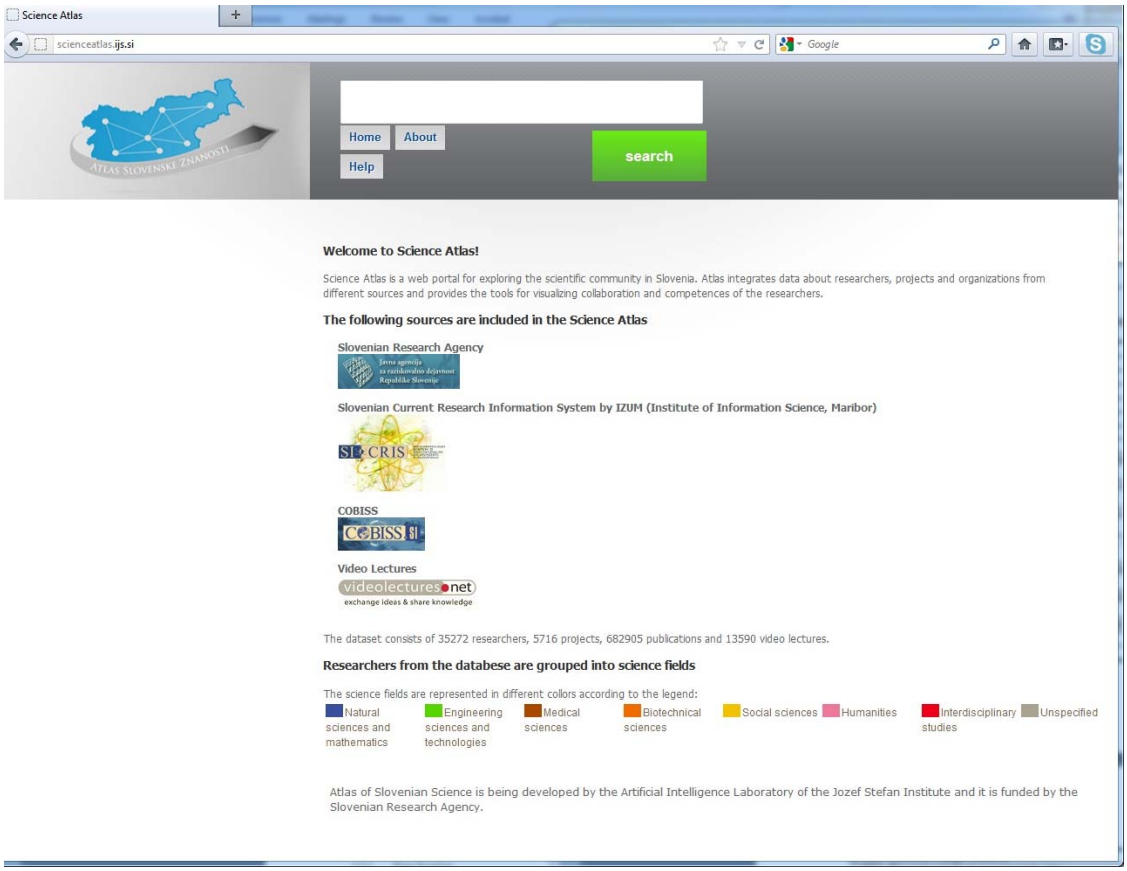
Atlas slovenske znanosti

RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

R53 –Final Prototype System

Ljubljana, 15.3.2013

The final prototype system is available at http://scienceatlas.ijs.si/, where you get the following introductory screen:



This application is based on HTML5 technology, which is supported by latest versions of all major browsers. Please update your browser to the latest version. It is recommended to use Firefox or Google Chrome.

Atlas of Slovenian Science: R53 – Prototype system

The system enables searching by researcher name or content keywords.

For instance, for »Dunja Mladenic« we get the following screen:

By selecting visualization on collaborations of the researcher, the following is returned by the system:

Searching on the content, “umetna inteligenca” and selecting visualization of collaborations, the following is returned by the system:

More information about the system including a demo video is also publicly available at http://ailab.ijs.si/tools/atlas-of-slovenian-science/

Atlas slovenske znanosti

RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

R6.1 Diseminacijski in promocijski načrt

Vsebina

1 Cilji diseminacijskih in promocijskih aktivnosti

1.1 Kvantificirani cilji

Glavni cilj diseminacije je doseganje večje prepoznavnosti in dosegljivosti znanstveno-raziskovalne dejavnosti ter obveščanje in povezovanje določenih ciljnih skupin z razvojnimi in raziskovalnimi dosežki . Na ta način se bo povečal dostop do informacij, izkušenj in rezultatov raziskovalcev in raziskovalnih skupin. Omogočen bo dostop do sicer težko dosegljivih ali razpršenih informacij. Vzpostavljeni bodo kanali komunikacije, preko katerih se bo najlažje doseglo ciljno skupino. Preko komunikacijskih kanalov se bodo informacije širile med različnimi ciljnimi skupinami. Povečala se bo časovna in lokacijska fleksibilnost, vzpostavila se bo trajna prisotnost preko različnih diseminacijskih kanalov. Ocenjujemo, da bodo s projektom doseženi naslednji cilji.

Cilj	Ocena
Dostop do informacij, izkušenj in rezultatov	> 50%
Dostop do težko dosegljivih ali razpršenih informacij	> 50%
Vzpostavitev kanalov komunikacije	100%
Povečanje časovne in lokacijske fleksibilnosti	> 30%
Trajna prisotnost preko različnih diseminacijskih kanalov	100%

1.2 Osnovna strategija

V Sloveniji trenutno obstaja 870 raziskovalnih organizacij in 1398 raziskovalnih skupin (po evidenci SICRIS, 2010). Preko skupne informacijske točke in različnih diseminacijskih kanalov bo omogočeno objavljanje sprotnih in končnih rezultatov. Uporabljeni bodo diseminacijski kanali: spletni portali in web kanali, socialna omrežja, tradicionalni mediji, tradicionalni dogodki ter spletna stran projekta.

2 Ciljne skupine

Skupina 1: **Raziskovalci**, različne raziskovalne skupine na inštitutih in drugih raziskovalnih organizacijah. Predstavljajo skupino z enakimi interesi ter združuje znanstvenike in raziskovalce, ki se ukvarjajo z raziskovanjem problemov, vezanih na določeno raziskovalno skupino. Opravljajo temeljne in uporabne raziskave.

Skupina 2: **Razvojniki v podjetjih**, raziskovalno razvojna dejavnost vezana na predkonkurenčne raziskave in industrijske raziskave.

Vrste raziskav, raziskovalni projekti in rezultati raziskav obeh skupin so zbrani in objavljeni v različnih medijih, zbirkah in evidencah. Atlas slovenske znanosti bo omogočal odprt dostop do razvojno raziskovalnih javnih podatkov različnih modalnosti, diseminacija bo potekala po različnih diseminacijskih kanalih.

Skupina 3: **Raziskovalci v visokošolskih organizacijah** (podiplomski študentje, predavatelji, …) razvilo se bo partnersko sooblikovanje visokošolskega in raziskovalnega prostora, povečal se bo raziskovalno-razvojni potencial v visokošolskem sektorju.

Skupina 4: **Mladi raziskovalci na osnovnih in srednjih šolah**, ki sodelujejo pri izdelavi raziskovalnih in seminarskih nalog na šolskem, regijskem in nacionalnem nivoju.

Skupina 5: **Bibliotekarji** v knjižnicah in ponujanju uslug in tudi pri lasnem raziskovalnem delu.

Ciljne skupine bomo naslovili s pomočjo različnih diseminacijskih in promocijskih kanalov, kot je prikazano v tabeli.

	Diseminacijski kanal			
	Portali in web kanali	Socialna omrežja	Obstoječe baze in seznami z e-naslovi	Tradicionlani dogodki in mediji
Skupina 1	x	x	x	x
Skupina 2	x		x	x
Skupina 3	x	x	x	x
Skupina 4	x	x	x	
Skupina 5	x			x

2.1 Portali, web kanali, in socialna omrežja

Diseminacija bo potekala preko portalov in spletnih strani: domača spletna stran projekta, http://www.arrs.gov.si/sl/, http://videolectures.net/ , http://www.zdruzenje-manager.si/, http://www.gzs.si/slo/, http://www.ozs.si/, (http://www.ist-world.org/, http://www.rtd.si/slo/ in drugih.

Preko socialnih omrežij *facebook* http://facebook.com/ in *twitter* *http://twitter.com/* bodo ustvarjene spletne strani z informacijami in rezultati projekta. Upamo, da bomo s tem dodatno podprli sodelovanje med institucijami, soavtorstvo ipd.

2.2 Obstoječe baze in seznami z e-naslovi

Potekala bodo redna obveščanja preko obstoječih seznamov e-naslovov (SICRIS, videolectures, ARRS). Pripravljena bo baza registriranih uporabnikov, obveščanje bo potekalo preko personaliziranih sporočil (*newsletter.si)

2.3 Tradicionalni dogodki

Organizirani bodo sestanki s potencialnimi uporabniki in nosilci vsebin tako po podjetjih, združenjih kot tudi inštitutih in drugih raziskovalnih organizacijah. Potekale bodo predstavitve na seminarjih in srečanjih določenih ciljnih skupin (Dnevi IJS, Akademski in raziskovalni dnevi). Pri tem se bomo osredotočeni na sledeče nacionalne in mednarodne dogodke, kot npr. Eureka, EUROCRIS 2012, Dnevi Jožefa Stefana, IKT konference 2012.

V sodelovanju s zainteresiranimi organizacijami (n.pr. ARRS, IZUM, GZS) bodo organizirane delavnice, kjer bo predstavljena enovita odprto dostopna informacijska točka, ki bo omogočala ne samo pregled informacij, temveč vrsto uporabnih aplikacij in storitev za potrebe raziskovalcev, organizacij in

podjetij. Prikazane bodo aplikacije in njihova uporaba. Uporabnikom bodo predstavljene vse možnosti uporabe orodja.

2.4 Tradicionalni mediji

Preko že obstoječe baze medijev (e-naslovi novinarjev in uredništev, faksi) bodo potekale sprotne promocijske in diseminacijske aktivnosti. Rezultati projekta in obveščanje o novostih portala *Atlasa slovenske znanosti* bodo posredovani vsem večjim slovenskim medijskim hišam. Predvidevamo objavo člankov, intervjujev in krajših prispevkov ter sodelovanja na okroglih mizah.

3 Grobi plan aktivnosti

Dejavnost	**Izvedba**
Izdelava promocijskega in diseminacijskega plana	M0 – M6
Izvajanje diseminacijskih aktivnosti in promocije	M0 – M30
Instalacija delov aplikacij portala na portalu ARRS in RTD ter portalu IJS	M24 – M30
Diseminacija končnih rezultatov	M 30
Izdelava kratkih promocijskih videov	M 30

Promocijski in diseminacijski plan bo izvajan skladno z načrtom izvedbe raziskovalnega projekta. Promocijske in diseminacijske aktivnosti bomo izvajali tako sprotno kot tudi ob koncu projekta po izbranih diseminacijskih kanalih. Za doseganje večje prepoznavnosti in povezljivosti različnih spletnih storitev za znanstveno-raziskovalno dejavnost, bomo v zadnjem delu projekta inštalirali dele storitev (analize, vizualizacije, ipd.) na nekaj izbranih portalov: ARRS, IJS in RTD.
Promocijski videi bodo namenjeni predstavitvi funkcionalnosti spletnega portala *Atlas slovenske znanosti.*

4 Podroben plan aktivnosti za naslednje leto

Dejavnost	**Izvedba**
izdelava promocijskega in diseminacijskega plana	februar 2011
izdelava baze ciljnih uporabnikov	marec 2011
sprotne diseminacijske in promocijske aktivnosti	marec – december 2011

V letu 2011 se bomo pri diseminaciji in promociji osredotočili na postavitev baze s kontaktnimi podatki ciljnih uporabnikov, izdelavi promocijskega materijala v elektronski obliki (newsletter, spletna stran projekta *Atlas slovenske znanosti)*, ter promociji preko tradicionalnih medijev.

5 Zasnova zunanje podobe

Namen projekta je vzpostaviti enotno spletno informacijsko točko. V ta namen bo zasnovana zunanja podoba: ime, domena, logotip. Izdelali bomo spletno stran projekta in pripravili krajši opis projekta za objavo v medijih.

6 Zaključek

Z izvedbo diseminacijskega in promocijskega načrta bomo vzpostavili sistem obveščanja, širjenja, in povezovanja različnih ciljnih skupin preko enotnega sistema za enostaven in odprt dostop do razvojno-raziskovalnih podatkov s ciljem spodbujanja novih idej, sodelovanja med domačimi institucijami in industrijo, promocije znanstveno-raziskovalnih dosežkov doma in v tujini ter vzpostavitvijo okolja inovativne in kreativne kulture.

7 Reference

http://www.arrs.gov.si/sl/, http://videolectures.net/ , http://www.zdruzenje-manager.si/, http://www.gzs.si/slo/, http://www.ozs.si/, (http://www.ist-world.org/, http://www.rtd.si/slo/, http://sicris.izum.si/, http://www.rtd.si/slo/ , http://www.ijs.si/ , http://www.eurekanetwork.org/ , http://eurocris.infoscience.cz/ , http://videolectures.net/dnevi_ijs/

Atlas slovenske znanosti

RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

Poročilo o izvedenih diseminacijskih in promocijskih dejavnostih

Z izvedbo diseminacijskega in promocijskega načrta smo vzpostavili sistem obveščanja, širjenja, in povezovanja različnih ciljnih skupin preko enotnega sistema *http://scienceatlas.ijs.si/* za enostaven in odprt dostop do razvojno-raziskovalnih podatkov s ciljem spodbujanja novih idej, sodelovanja med domačimi institucijami in industrijo, promocije znanstveno-raziskovalnih dosežkov doma in v tujini ter vzpostavitvijo okolja inovativne in kreativne kulture.

Promocijski in diseminacijski plan je bil izvajan skladno z načrtom izvedbe raziskovalnega projekta. Promocijske in diseminacijske aktivnosti smo izvajali tako sprotno kot jih bomo tudi ob koncu projekta po izbranih diseminacijskih kanalih. Za doseganje večje prepoznavnosti in povezljivosti različnih spletnih storitev za znanstveno-raziskovalno dejavnost, bomo v zadnjem delu projekta inštalirali dele storitev (analize, vizualizacije, ipd.) na nekaj izbranih portalov: ARRS, IJS in RTD. Promocijski videi so namenjeni predstavitvi funkcionalnosti spletnega portala *Atlas slovenske znanosti: http://scienceatlas.ijs.si/*.

Glavni cilj diseminacije je bil doseganje večje prepoznavnosti in dosegljivosti znanstveno-raziskovalne dejavnosti ter obveščanje in povezovanje določenih ciljnih skupin z razvojnimi in raziskovalnimi dosežki . Na ta način se je povečal dostop do informacij, izkušenj in rezultatov raziskovalcev in raziskovalnih skupin. Omogočen je dostop do sicer težko dosegljivih ali razpršenih informacij. Vzpostavljeni so bili kanali komunikacije, preko katerih se bo najlažje doseglo ciljno skupino. Preko komunikacijskih kanalov se tako informacije širijo med različnimi ciljnimi skupinami. Povečala se je časovna in lokacijska fleksibilnost, vzpostavila se je trajna prisotnost preko različnih diseminacijskih kanalov.

V Sloveniji trenutno obstaja 870 raziskovalnih organizacij in 1398 raziskovalnih skupin (po evidenci SICRIS, 2010). Preko skupne informacijske točke http://scienceatlas.ijs.si/ in različnih diseminacijskih kanalov je omogočeno objavljanje sprotnih in končnih rezultatov. Uporabljeni so bili različni diseminacijski kanali: spletni portali in web kanali, socialna omrežja, tradicionalni mediji, tradicionalni dogodki ter spletna stran projekta.

Preko spletnega portala http://scienceatlas.ijs.si/ smo dosegli različne ciljne skupine:

- raziskovalci in raziskovalne skupine, ki opravljajo temeljne in uporabne raziskave.,
- razvojniki v podjetjih, raziskovalno razvojna dejavnost vezana na predkonkurenčne raziskave in industrijske raziskave.
- raziskovalci v visokošolskih organizacijah (podiplomski študentje, predavatelji, …)
- mladi raziskovalci na osnovnih in srednjih šolah, ki sodelujejo pri izdelavi raziskovalnih in seminarskih nalog na šolskem, regijskem in nacionalnem nivoju
- bibliotekarji

Ciljne skupine smo naslovili s pomočjo različnih diseminacijskih in promocijskih kanalov, kot je prikazano v tabeli.

	Diseminacijski kanal			
	Portali in web kanali	Socialna omrežja	Obstoječe baze in seznami z e-naslovi	Tradicionlani dogodki in mediji
Skupina 1	x	x	x	x
Skupina 2	x		x	x
Skupina 3	x	x	x	x
Skupina 4	x	x	x	
Skupina 5	x			x

Diseminacija je potekala preko portalov in spletnih strani: domača spletna stran projekta http://scienceatlas.ijs.si/ , http://www.arrs.gov.si/sl/, http://videolectures.net/ , http://www.zdruzenje-manager.si/, http://www.gzs.si/slo/, http://www.ozs.si/, (http://www.ist-world.org/, http://www.rtd.si/slo/ in drugih.

Organizirani so bili sestanki s potencialnimi uporabniki in nosilci vsebin tako po podjetjih, združenjih kot tudi inštitutih in drugih raziskovalnih organizacijah. Potekale so predstavitve na seminarjih in srečanjih določenih ciljnih skupin (Dnevi IJS, Akademski in raziskovalni dnevi). Pri tem smo se osredotočili na sledeče nacionalne in mednarodne dogodke, kot npr. Eureka, EUROCRIS 2012, Dnevi Jožefa Stefana, IKT konference 2012.

Izvedli smo predavanje za knjižničarje, študente in srednješolce, ki je potekalo v Univerzitetni knjižnici Maribor. Promocijsko predavanje smo izvedli tudi v okviru gibanja InCo v Državnem svetu, posneta je bila oddaja na tretjem programu Radia Slovenija, povezali smo se z Združenjem manager ter vzpostavili kanal, kjer predstavljamo raziskovalne in podjetniške dosežke. Na konferenci "Drugačna pot do znanja" smo predstavili portal in dosežke znanosti osnovno in srednješolskim učiteljem s področja tehnike in tehnologije. Promocijsko predavanje smo izvedli tudi na posvetovanju "Prosti dostop do dosežkov slovenskih znanstvenikov", ki sta ga organizirala Sekcija za specialne knjižnice in Sekcija za visokošolske knjižnice pri Zvezi bibliotekarskih društev Slovenije. Na Institutu »Jožef Stefan« smo izvedli predavanje in predstavitveno delavnico za 60 kadrovnikov slovenskih podjetij. Imeli smo tudi več delovnih sestankov na Gospodarski zbornici Slovenije.

Za potrebe motivacijske vertikale smo v sodelovanju z ARRS in razvojno-raziskovalnimi oddelki in inštituti pripravljali gradivo za promocijske videe, ki raziskovalcem, gospodarstvu in širši javnosti

predstavljajo dobre raziskovalne skupine, posameznike ter raziskovalne dosežke. Rezultat naloge je serija 70 kratkih promocijskih videov v slovenskem in angleškem jeziku na portalu, na naslovu http://videolectures.net/promogram/ .

Začeli smo projekt Izjemnih znanstvenih dosežkov v letih 2011, kjer s serijo posnetih dogodkov, ki jih organizira Javna agencija za raziskovalno dejavnost Republike Slovenije (ARRS) v sodelovanju s Slovensko akademijo znanosti in umetnosti (SAZU) promoviramo in objavljamo znanstvene dosežke. Gre za predstavitev dosežkov, ki so jih po pregledu zaključnih in letnih poročil raziskovalnih projektov ter programov za leto 2011, financiranih s strani ARRS, izbrale članice in člani posameznih Znanstvenoraziskovalnih svetov ved ARRS. Predstavljeni so bili znanstveni dosežki s področij naravoslovnih, tehniških, medicinskih, biotehniških, družboslovnih in humanističnih ved ter interdisciplinarnih raziskav.

Predstavitve znanstvenih dosežkov omogočajo vpogled v delo slovenskih znanstvenic in znanstvenikov ter zajemajo bistvene znanstvene rezultate in opisujejo potek raziskovalnega dela. Namenjene so različnim javnostim: strokovni in laični, znanstveni in gospodarski, slovenski in tuji. Z izvedbo predstavitev ter z objavo na portalu videolectures. net želi ARRS gledalce seznaniti z novostmi, učinki in rezultati na vseh področjih znanstvenoraziskovalne dejavnosti v Sloveniji.

Tako je na delovnem naslovu http://videolectures.net/znanstveni_dosezki2011/, že postavljneih71 videov.

Tudi po zaključku projekta bodo organizirane delavnice, kjer bo predstavljena enovita odprto dostopna informacijska točka, ki bo omogočala ne samo pregled informacij, temveč vrsto uporabnih aplikacij in storitev za potrebe raziskovalcev, organizacij in podjetij. Prikazane bodo aplikacije in njihova uporaba. Uporabnikom bodo predstavljene vse možnosti uporabe orodja.

Rezultati projekta in obveščanje o novostih portala *Atlasa slovenske znanosti* bodo po zaključku projekta posredovani vsem večjim slovenskim medijskim hišam. Predvidevamo objavo člankov, intervjujev in krajših prispevkov ter sodelovanja na okroglih mizah.

Reference

*http://scienceatlas.ijs.si/*http://www.arrs.gov.si/sl/, http://videolectures.net/ , http://www.zdruzenje-manager.si/, http://www.gzs.si/slo/, http://www.ozs.si/, (http://www.ist-world.org/, http://www.rtd.si/slo/, http://sicris.izum.si/, http://www.rtd.si/slo/ , http://www.ijs.si/ , http://www.eurekanetwork.org/ , http://eurocris.infoscience.cz/ , http://videolectures.net/dnevi_ijs/ , http://videolectures.net/znanstveni_dosezki2011/ , http://videolectures.net/promogram/

Atlas slovenske znanosti

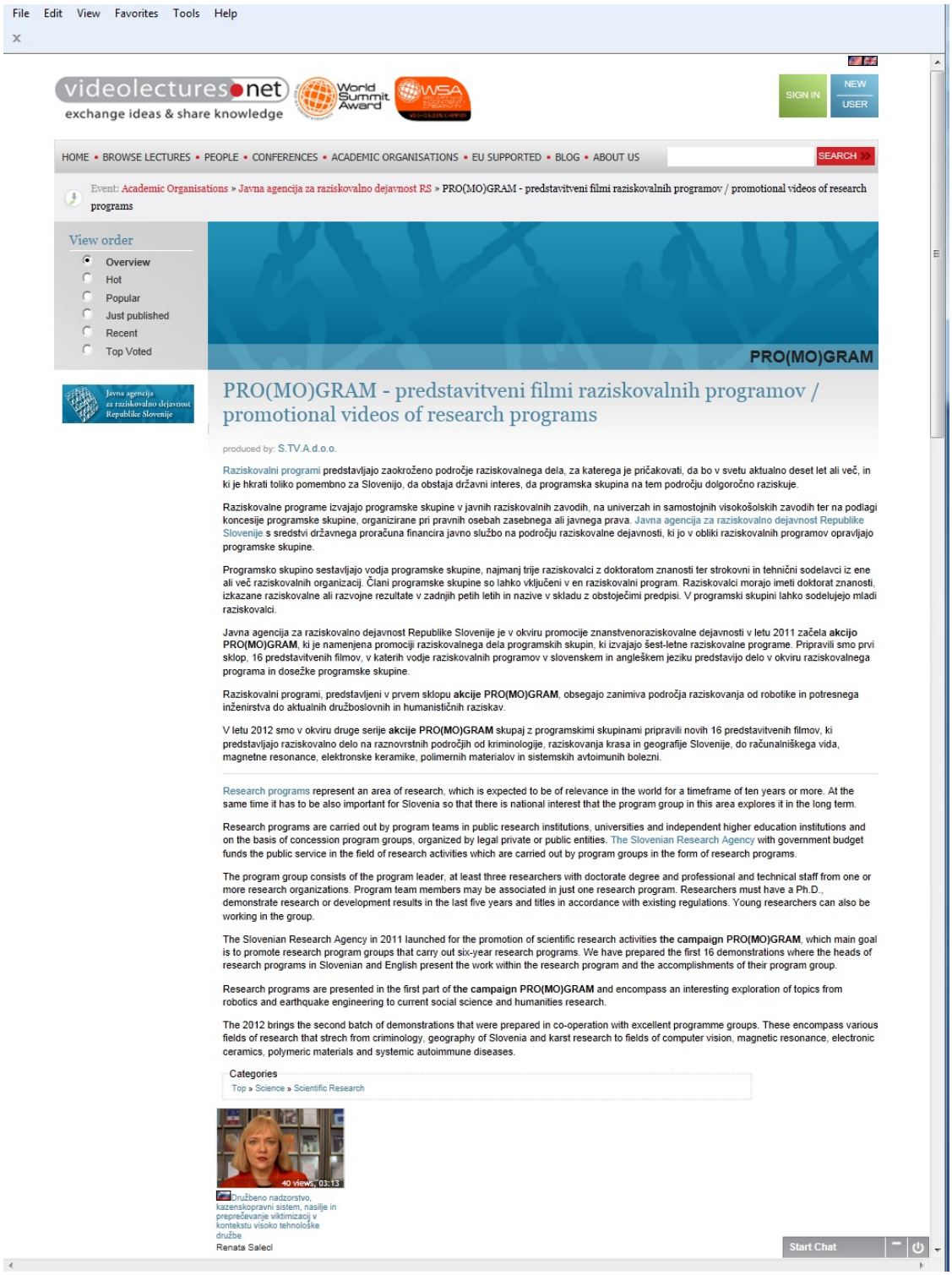
RAZISKOVALNI PROJEKT V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA "KONKURENČNOST SLOVENIJE 2006-2013" v letu 2010

R63 – Promocijski videi

Ljubljana, 15.3.2013

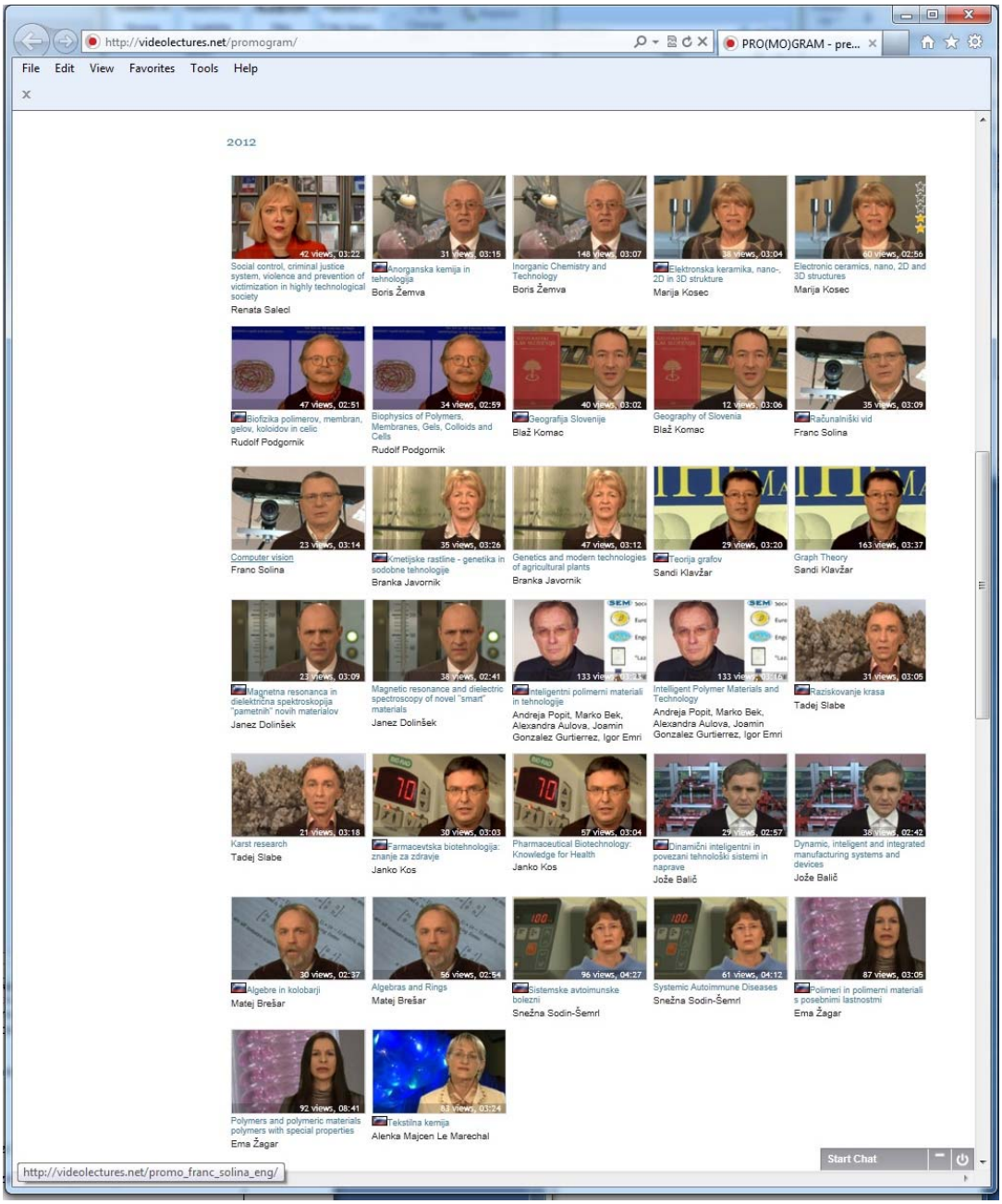
Atlas of Slovenian Science: R63 – Promocijski videi

Promocijski videi so skupaj z opisom dostopni na http://videolectures.net/promogram/:



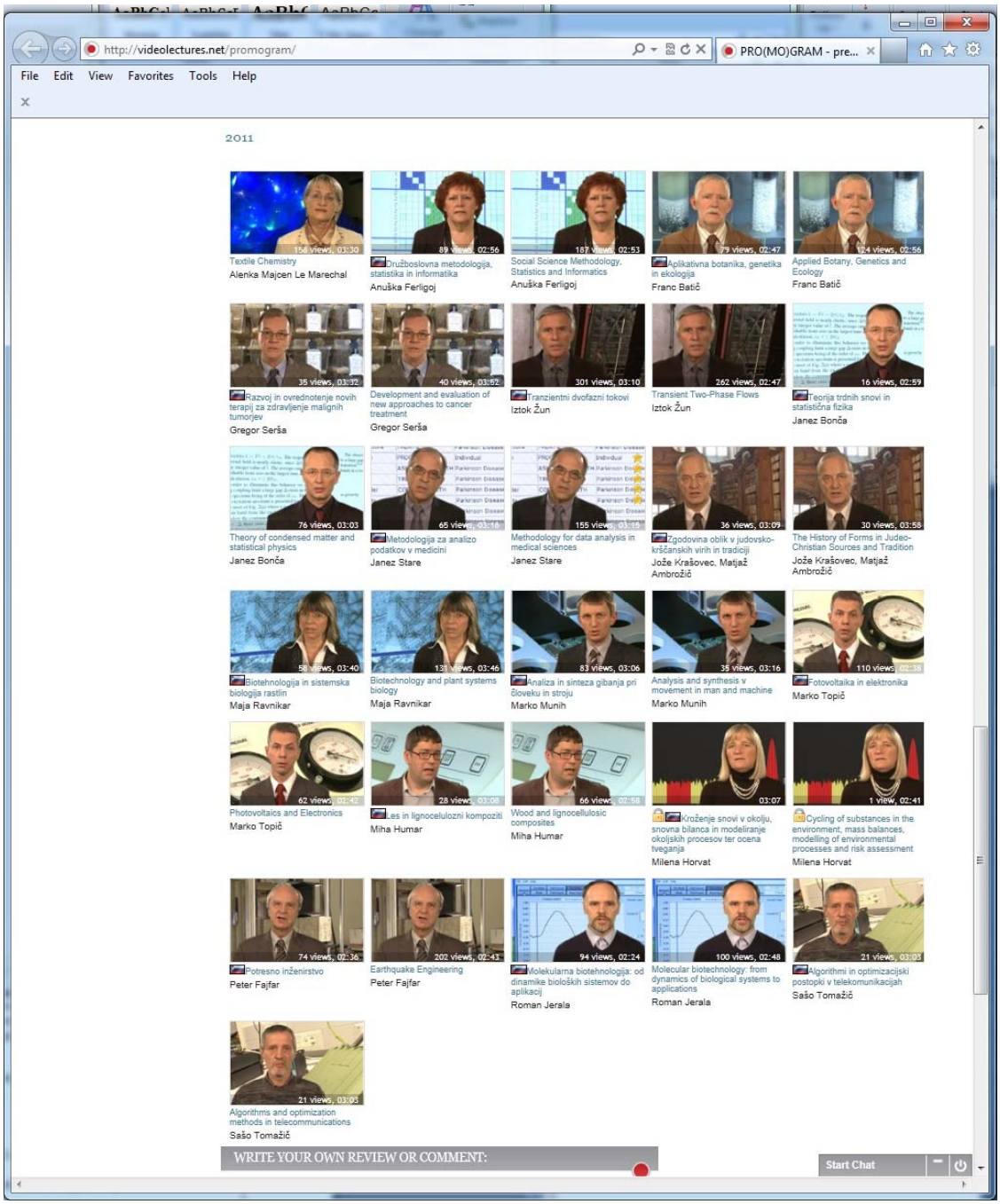
Atlas of Slovenian Science: R63 – Promocijski videi

Snemali smo v letu 2012:



Atlas of Slovenian Science: R63 – Promocijski videi

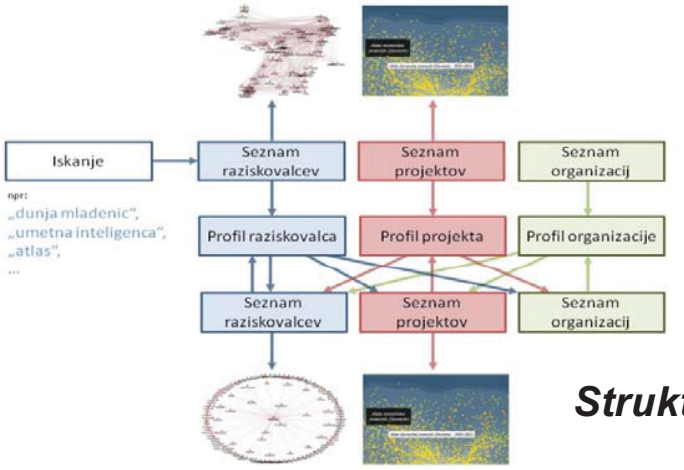
In v letu 2011:



Več informacij je na voljo v poročilu R62.

NARAVOSLOVJE
Področje: 1.02 Računalništvo in informatika
Atlas slovenske znanosti, Vir: http://ailab.ijs.si/tools/atlas-of-slovenian-science/

Portal: http://scienceatlas.ijs.si/



Struktura sistema

Enotnen sistem za enostaven dostop do razvojno-raziskovalnih podatkov s ciljem spodbujanja novih idej, sodelovanja med domačimi institucijami in industrijo, promocije znanstveno-raziskovalnih dosežkov doma in v tujini ter vzpostavljanje okolja inovativne in kreativne kulture. Razvit sistem je dostopen kot spletna informacijska točka na osnovi podatkov in baz, ki beležijo dosežke slovenskih raziskovalcev.

Modeliranje podatkov z analizo omrežij: Kohezija (angl. cohesion, lat. cohaerere - držati se skupaj) merjenje gostote ter povprečnega števila povezav med raziskovalci v omrežju in podomrežij. Posredništvo (angl. brokerage) – možnosti za izmenjavo informacij v omrežju raziskovalcev. Glavne komponente so posredniki in mostovi.

Modeliranje podatkov z analizo besedil: Odkrivanje raznih lastnosti besedila – število in frekvence besed, korelacije med besedama. Nenadzorovano strojno učenje – razvrščanje raziskovalcev/projektov v nove skupine na podlagi njihove vsebine. Nadzorovano strojno učenje – klasifikacija ne kategoriziranih raziskovalcev/projektov po obstoječi klasifikaciji na podlagi njihove vsebine in ostalih kategoriziranih raziskovalcev/projektov.

Modeliranje časovnega razvoja omrežij: Razvoj obsega in gostote omrežja – število raziskovalcev/projektov in povezav med njimi čez čas. Razvoj premera omrežja – koliko raziskovalcev posreduje med vsaka dva raziskovalca v omrežju in kako se ta odnos menja čez čas. Razvoj povezane komponente omrežja – velikost skupine med seboj povezanih raziskovalcev skozi čas.

Modeliranje časovnega razvoja tematik: Top-down pristop – analiza celotne vsebine in razbijanje v posamezne časovne okvirje. Bottom-up pristop – proučevanje vsebine posameznih časovnih okvirjev in sinteza v celotno razumevanje razvoja tematik.