Јелена Митровић
Филолошки факултет
Универзитет у Београду

UDC: 007:004
DOI: 10.18485/dh.2015.1.ch7

МОДЕЛ ГРУПНЕ РАСПОДЕЛЕ РАДА У ДИГИТАЛНОЈ ХУМАНИСТИЦИ

Сажетак

Дигитална хуманистика је интердисциплинарна област истраживања која повезује хуманистичке науке са информационо-комуникационим технологијама које су нам сваким даном све више доступне. Настанком интернета, а посебно развојем веба отворили су се многе могућности за истраживања из области хуманистике. Веб 2.0 пружа технолошке основе за развој и функционисање парадигме дигиталне хуманистике, јер је фокус овог концепта управо на сарадњи и дељењу, интеракцији и пружању информација.

Групна расподела рада, како се описно може превести енглески термин Crowdsourcing, пословни је модел који је нашао широку примену у области културе и науке. Показало се да је овај модел веома користан и уколико се примени на прави начин може да донесе веома добре резултате. Главна премиса групне расподеле рада и њене ефикасности заснива се на чињеници да је група људи који појединачно обављају неке релативно лаке задатке способна да оствари резултате који се могу мерити са резултатима добијеним од експерата у датој области. Тај феномен се још назива и „Групна мудрост“ (енг. Wisdom of the Crowd). У раду ће бити приказани примери добре праксе употребе овог модела у области дигитализације културне баштине и у области рачунарске обраде природних језика.

Кључне речи: Дигитална хуманистика, Групна мудрост, Групна расподела рада, Обрада природних језика, Дигитализација, Веб 2.0, Српски ворднет

1. Увод

Дигитална хуманистика је интердисциплинарна област истраживања која повезује хуманистичке науке са информационо-комуникационим технологијама које су нам сваким даном све више доступне. Ове нове технологије умногоме олакшавају и убрзавају

истраживања за која је раније било потребно много више времена и труда, али и подржавају и омогућавају рад научника, предавача, студената, ученика и свих заинтересованих за изучавање најразличитијих области људског деловања. Хуманистичке науке можемо описати као науке које се баве изучавањем обиља начина на које људи из целог света, из свих историјских периода, обрађују и документују, на неки начин бележе искуства. Филозофија, теологија, књижевност, уметност, историја, музикологија, лингвистика, театрологија – све су то науке које побољшавају наше разумевање света и дају нам алате за бележење тих сазнања. Групна расподела рада, како се описно може превести енглески термин „crowdsourcing", пословни је модел који се преселио у област културе и науке. Показало се да је овај модел веома користан и уколико се примени на прави начин може да донесе веома добре резултате у области дигиталне хуманистике.

2. Групна расподела рада (енг. Crowdsourcing)

Настанком интернета, а посебно развојем веба отвориле су се многе могућности за истраживања из области хуманистике. Тако је постало могуће да своја открића неупоредиво лакше поделимо са истраживачима и научницима који су заинтересовани за исте области истраживања, те да и сами дођемо до нових сазнања. Веб 2.0 [1] пружа технолошке основе за развој и функционисање парадигме дигиталне хуманистике, јер је фокус овог концепта управо на сарадњи и дељењу, интеракцији и пружању информација.

Главна премиса групне расподеле рада и њене ефикасности заснива се на чињеници да је група људи који појединачно обављају неке релативно лаке задатке способна да оствари резултате који се могу мерити са резултатима добијеним од експерата у датој области. Тај феномен се још назива и „Групна мудрост" (енг. Wisdom of the Crowd) (Arazy, Morgan / Patterson). У раду ће бити приказани примери добре праксе употребе овог модела у области дигитализације културне баштине и у области рачунарске обраде природних језика.

1 Understanding Web 2.0, Murugesan, San. http://hcotuk.etu.edu.tr/bil554/Understanding_web_2.pdf (приступљено 1.12. 2015).

Термин групна расподела рада заправо је слободан превод оригиналног, енглеског термина „crowdsourcing“ – што је кованица настала од речи „crowd“ и „(out)sourcing“. Тај термин је у употребу увео Џеф Хауи (енг. Jeff Howe) и први пут га употребио у раду (Howe) у коме је овај феномен описао једноставно као начин да многи обављају послове и задатке које је раније обављало само неколико људи. Дефиниција аутора (Estellés-Arolas / González-Ladrón-de-Guevara) јесте једна од најпотпунијих дефиниција групне расподеле рада: „Групна расподела рада је врста учествовања у некој активности у којој особа, институција, непрофитна организација или компанија нуде групи особа различитих нивоа знања, хетерогености и броја, преко флексибилног отвореног позива, да се добровољно посвете неком задатку. Посвећивање задатку, који је разнолике комплексности и модуларности и у коме група људи учествује својим знањем и/или искуством, новцем, радом, увек укључује узајамну корист. Кориснику ће бити задовољена нека потреба, било да је то економска, друштвена, лична потреба, или развијање сопствених вештина, док ће особа, институција или организација која нуди задатак, оно чиме су корисници допринели, а чији облик ће зависити од врсте задатка искористити у своју корист“.

3. Пионирски подухвати групне расподеле рада у хуманистици

Један од првих пројеката групне расподеле рада у хуманистичким наукама, и то из области лексикографије, јесте пројекат изградње Оксфордског речника енглеског језика (енг. Oxford English Dictionary (OED)). Овај пројекат изградње свеобухватног енглеског речника, представља изврстан пример примене овог модела сарадње и један од најранијих претходника данашњих пројеката групне расподеле рада, који се највише заснивају на сарадњи путем интернета. Тако је већина историјских и лексичких информација које се налазе у ОЕД заснована на милионима цитата добијених из текстова на енглеском језику кроз такозвани Читалачки програм кроз који волонтери и плаћени учесници сакупљају цитате како би илустровали коришћење

речи. Читалачки програм Оксфордског речника започео је 1857. године, када су волонтери почели да прикупљају цитате за Нови енглески речник у издању Британског филолошког друштва[2]

Пројекат изградње веб сајта Википедија такође се може схватити као пројекат групне расподеле рада – један од првих, у дигиталном свету, и свакако један од најдужих. Википедија је покренута у фебруару 2003., и у трећој години постојања (26. августа 2005.) пројекат је већ имао преко 2 милиона чланака. Данас Википедија постоји на преко 150 језика, а поред Википедије постоје још и Викицитат, Викикњиге, Викизворник, Викиречник.

4. Групна расподела рада у дигитализацији културне баштине

Пројекат Digitalkoot одличан је пример још једног пројекта групне расподеле рада у дигиталној хуманистици. Покренут је 8. фебруара 2011. године са циљем исправљања грешака насталих у процесу дигитализације финских новина с краја 19. века, Aamulethi, које се чувају у Националној библиотеци Финске. Као први овакав програм у Европи, он кроз групну расподелу рада покреће велики број људи да помогну дигитализацију милиона страна архивских материјала. Овај програм комбинује забаву и добровољни рад, а спроводи се кроз две онлајн игрице. У „Лову на кртице" (енг. Mole Hunt, фин. Myyräjahti), пред играча се постављају две речи, за које, што је то брже могуће, треба утврдити јесу ли исте, при чему се мора обратити пажња на велика слова и на интерпункцијске знаке. Једна реч је оригинална реч извучена из скенираног текста, док је друга реч верзија те оригиналне речи добијена технологијом оптичког препознавања текста. После сваког одговора, кртица нестане са екрана, али тек на крају одиграног нивоа играч може да види колико тачних одговора је дао. На тај начин се откривају словне грешке у архивираним материјалима. Друга игрица носи назив „Мост за кртице" (енг. Mole Bridge, фин. Myyräsilta). Од играча се очекује да тачно откуцају речи које се појаве

2 Oxford dictionary, http://blog.oxforddictionaries.com/2014/02/can-world-englishes-benefit-crowdsourcing/

на екрану, при чему се такође мора обратити пажња на велика слова и на интерпункцијске знаке. Тачан одговор, односно тачно укуцана реч, помаже кртицама да изграде мост преко реке. Са сваком новом укуцаном речју, мост за кртице добија нови део од дрвета. Систем онда утврђује јесу ли одговори тачни тако што се нови делови који су добијени за тачне одговоре претварају у челичне делове моста, док они делови који су мосту додати након нетачног укуцавања речи – експлодирају и нестају носећи са собом и неколико околних делова моста. Када играч успе да изгради мост и спасе кртице, ниво се завршава и израчунавају се добијени поени. У овој игрици, као и у претходној, добијају се позитивни поени за тачне одговоре, а негативни поени за нетачне одговоре. У игрици „Мост за кртице" постоји дугме на које се може кликнути ако играч не може да препозна реч, то јест ако не може да је укуца, такозвано Impossible дугме. Притиском на то дугме, у горњем делу екрана се, уместо речи коју играч није могао да препозна, појављује нова реч, а број стечених поена остаје непромењен, то јест играч не губи поене. Играње онлајн игрица у циљу дигитализације је релативно нов концепт, али свакако је то иновација која је привукла много пажње. На овај начин се врши и промоција дигитализованих збирки, јер тако много више људи има прилику да сазна за дигитализоване материјале.

DigiTalkoot пројекат је био веома успешан: скоро 110.000 учесника је обавило преко 8 милиона задатака исправљања грешака у дигитализованим текстовима – све то захваљујући снази модела групне расподеле рада.

5. Домаћи пројекат групне расподеле рада из области рачунарске лингвистике

Пројекат групне расподеле рада који је недавно спроведен у нашим условима, а у области хуманистичких наука, односи се на доградњу Српског ворднета (енг. Serbian WordNet – SWN) (Mladenović, Mitrović / Krstev) . Ова изузетно важна лексичко-семантичка мрежа и лексички ресурс који налази најразличитије примене у области Рачу-

нарске лингвистике, почео је да се развија у оквиру пројекта Балканет (енг. BalkaNet) (Stamou, Oflazer / Pala) као што је то био случај и са бугарским, чешким, румунским, грчким и турским ворднетом. SWN се заснива на структури Принстонског ворднета (енг. Princeton WordNet – PWN) (C. Fellbaum), те је изграђен у складу са принципом модела проширивања (енг. expand model), у складу с правилима која је налагао пројекат BalkaNet, то јест копирањем синсетова из принстонског ворднета у српски ворднет и преводећи их, уз очување хијерархијске структуре Принстонског ворднета. По завршетку пројекта BalkaNet, 2004. године, нови синсетови додавани су захваљујући удруженом раду професора, сарадника и студената Катедре за библиотекарство и информатику, Филолошког факултета у Београду и Групе за језичке ресурсе и технологије Универзитета у Београду [3], као и захваљујући бројним волонтерима (Крстев и др.) . Данас Српски ворднет броји нешто више од 21,000 синсетова, док Принстонски броји око 117,000 синсетова.

6. Групна расподела рада у циљу додавања нових семантичких веза у ворднету

Структура сваког ворднета заснива се на релацијама међу речима – основна релација међу речима у сваком ворднету јесте синонимија – на пример, као између речи „кола“ и „аутомобил“. Синоними су у ворднету груписани у скупове под називом синсетови (енг. Synsets). Најчешћа релација између синсетова је релација између надређеног и подређеног, то јест хиперонимија и хипонимија. Тако су, на пример, повезани један општији синсет {намештај, комад намештаја} и специфичнији као што су {кревет} и {кревет на спрат}. Релација хипонимије је транзитивна, то јест, ако кажемо да је фотеља врста столице, и ако је столица врста намештаја, следи да је фотеља врста намештаја. С друге стране релација под називом меронимија је однос између целине и дела те целине – столица и наслон су пример такве релације. Придеви у ворднету су организовани у смислу антонимије, то јест супротности – на пример, сув и мокар, млад и стар.

3 Друштво за језичке ресурсе и технологије, http://jerteh.rs/

Већина релација у ворднету повезује речи које припадају истом делу говора (енг. part of speech – POS). Тако се сваки ворднет састоји од четири подскупа, од којих сваки садржи само именице, придеве, глаголе или прилоге, уз само неколико показивача (енг. pointers) између делова говора. Те, такозване Cross-POS релације (релације између синсетова различитих врста речи) подразумевају морфосемантичке везе између семантички сличних речи које имају корен истог значења. У Принстонском ворднету тако имамо observe (глагол), observant (придев) observation, observatory (именице). Када говоримо о паровима именица-глагол, одређена је семантичка улога именице у односу на глагол – {спаваћа_ кола} су локација за {спавати}, {сликар}је чинилац, извршилац (енг. agent) у односу на {сликање}, док је {слика} резултат.

У циљу проширења Српског ворднета и како бисмо овај ресурс могли да користимо у сврху семантичке анализе (енг. Semantic analysis) и анализе осећања (енг. Sentiment analysis), као и ради повезивања са другим вредним ресурсима српског језика, као што је Онтологија реторичких фигура за српски језик (Mladenović and Mitrović), одлучено је да у Српски ворднет буду убачене нове семантичке везе. Недавно спроведено истраживање (Mladenović et. al) је тако резултирало додавањем нових релација у Српски ворднет, које се могу додати и у друге светске ворднетове, и то релације између именичких и придевских синсетова, а све то на основу реторичке фигуре поређења.

У сврху евалуације аутоматске методе којом је претходно описано проширење остварено, коришћена је групна расподела рада, и то са веома добрим резултатима. Преко друштвене мреже Фејсбук (енг. Facebook) [4] било је омогућено дистрибуирање упитника које смо креирали помоћу сервиса Google Forms [5]. Током пет дана, учесници у овом пројекту, волонтери који желе да помогну да се српски језик очува у дигиталном окружењу, попуњавали су упитнике који су садржавали комбинације именица и придева добијених из Корпуса савременог српског језика (Vitas / Krstev). Задатак учесника је био да се изјасне да ли те комбинације користе у свакодневном језику. Тако

4 Facebook, www.facebook.com

5 Google Forms http://bit.ly/1OuDYBC

су, на пример, неке од понуђених комбинације биле: „Бео као снег“, „Црвен као булка“, „Чист као апотека“ итд. У овом истраживању учествовало је 434 особе, а њихови одговори проверени су статистичким методама које се у сличним истраживањима обично користе. Резултат самог истраживања било је додавање нових семантичких веза у Српски ворднет што је изузетно важно за најсавременије примене овог непроцењивог менталног лексикона и семантичке мреже за српски језик.

7. Закључак

Модел групне расподеле рада је свакако веома користан у дигиталној хуманистици, али је суштински важно да се пројекти које спроводимо ослањајући се на овај модел спроведу узимајући у обзир ко учествује у пројекту и то колико те особе уствари могу да нам помогну. У случају доградње Српског ворднета, већина учесника биле су особе које подржавају Фејсбук страницу Друштва за лексичке ресурсе и технологије, Универзитета у Београду, па се може претпоставити да су то особе које желе да подрже и унапређивање једног од најважнијих ресурса које имамо за рачунарску обраду српског језика. И други значајни пројекти могу се спровести уз коришћење модела групне расподеле рада, а свакако се планира и унапређивање других ресурса и рачунарских апликација за обраду српског језика.

Литература

Arazy, Ofer, Wayne Morgan / Raymond Patterson. „Wisdom of the Crowds: Decentralized Knowledge Construction in Wikipedia.“ *6th Annual Workshop on Information Technologies & Systems*. Milwaukee, USA., n.d.

Estellés-Arolas, Enrique / Fernando González-Ladrón-de-Guevara. „Towards an Integrated Crowdsourcing Definition.“ *Journal of Information Science* 38.2 (2002): 189-200.

Fellbaum, Christiane. „ WordNet and wordnets.“ Brown, Keith et al. Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670. *Encyclopedia of Language and Linguistics*. Second edition. Oxford: Elsevier, 665-670, 2005. 665-670.

Fellbaum, Christiane. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press, 1998.

Howe, Jeff. *Crowdsourcing: How the Power of the Crowd is Driving the Future of Business*. London: Random House Business Books, 2008.

Mladenović, Miljana / Jelena Mitrović. „Ontology of Rhetorical Figures for Serbian." *Lecture Notes in Computer Science and Artificial Intelligence*. Springer-Verlag Berlin Heilderberg, 2013. 383-393.

Mladenović, Miljana, Jelena Mitrović. „Semantic Networks for Serbian – New Tools for Developing and Maintaining a WordNet." *35th Annivarsary of Computational Linguistics in Serbia*. Belgrade, 2013.

Mladenović, Miljana, Jelena Mitrović / Cvetana Krstev. „A Language-independent Model for Adding New Semantic Relations to a WordNet." *Global WordNet Conference*. Bucharest, 2016.

Mladenović, Miljana,, Jelena Mitrović / Cvetana Krstev. „Developing and Maintaining a WordNet: Procedures and Tools." *Proceedings of the Global WordNet Conference*. Tartu, Estonia, 2014. 55-62.

Stamou, Sofia,, и други. „Balkanet: A Multilingual Semantic Network for Balkan Languages." *1st International Global WordNet Conference*. "Balkanet: A Multilingual Semantic Network for Balkan Languages" Paper presented at the 1st International Global WordNet Conference, Mysore, India, January 21-25 2002., 2002.

Vitas, Duško / Cvetana Krstev. „ Processing of Corpora of Serbian Using Electronic Dictionaries." *Prace Filologiczne*. Warszawa, 2012. 279-292.

Крстев, Цветана, и други. „Кооперативан рад на доградњи Српског Wordneta." *Infotheca IX(1-2): 57-75.* IX.(1-2) (2008): 57-75.

Jelena Mitrović
University of Belgrade
Faculty of Philology

CROWDSOURCING IN DIGITAL HUMANITIES

Summary

Digital Humanities is an interdisciplinary field of research that connects Humanities with information-communication technologies (ICT) which are more and more available to us every day. These technologies facilitate and speed up research in areas where much more effort and time was needed before, and generally support the work of scientists, lecturers, students and all persons who are interested in studying different areas of human endeavors.

With the emergence of Internet, and especially with the development of the Web, new possibilities for research in the Humanities have appeared. It has become possible for us to share our findings with other scientists and researchers much more easily and to acquire new knowledge. Web 2.0 gives us the technical basis for the development and functioning of the Digital Humanities paradigm, because the focus of this concept is on cooperation, sharing, interaction and providing information.

Crowdsourcing is a business model which found its way in the areas of culture and science. This model can be very successful and it can bring very good results if it is implemented in the right way. The main premise of crowdsourcing is based on the fact that a group of people that separately perform some relatively simple tasks is able to accomplish results that are comparable to the results obtained from experts in the same field of research. This phenomenon is also known as the "Wisdom of the Crowd". In this paper we will present best practice examples of using crowdsourcing in digitization of cultural heritage and in Natural Language Processing.

Key words: Digital Humanities, Group Wisdom, Crowdsourcing, Natural Language Processing, Digitization