10. Ravi Sankar P., Srinivasa Rao B. K. N. Parallel Architecture for Implementation of Contrast Limited Adaptive Histogram Equalization. *International Journal of Advanced Engineering Sciences and Technologies (IJAEST)*, vol. 10, issue 1, pp. 047–051.

11. Rajesh kumar Rai, Puran Gour, Balvant Singh. Underwater Image Segmentation using CLAHE Enhancement and Thresholding. International Journal of Emerging Technology and Advanced Engineering, 2012, January, vol. 2, issue 1, pp. 118–123.

УДК 004.91

НЕКОТОРЫЕ АСПЕКТЫ СОЗДАНИЯ ИНФОРМАЦИОННЫХ СИСТЕМ ДЛЯ СБОРА И ХРАНЕНИЯ НАУЧНОЙ И НАУКОМЕТРИЧЕСКОЙ ИНФОРМАЦИИ[1]

Умаров Адам Сейлымович, студент, Астраханский государственный университет, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а, e-mail: mathmod@bk.ru

Попова Наталья Валерьевна, магистрант, Астраханский государственный университет, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а, e-mail: mathmod@bk.ru

Заболотина Виктория Андреевна, кандидат технических наук, доцент, Астраханский государственный университет, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а, e-mail: mathmod@bk.ru

Статья посвящена вопросам обеспечения эффективности использования информационных систем, позволяющих производить сбор, представление и анализ информации о результатах научной деятельности. Такими системами являются научные сети, системы для создания и информационной поддержки персональных страниц ученых в Интернете, Current Research Information Systems (CRIS). В работе описаны основные принципы, которые используются в концепциях систем рассматриваемого класса, анализируются причины, приводящие к нарушению концепций и появлению ряда эксплуатационных недостатков (повторный ввод данных о результате научной деятельности и появление дубликатов, неопределенность информации об авторе в описании результата научной деятельности, сложность классификации публикаций по принадлежности к базам цитирований). С учетом особенностей ручного ввода данных о результатах научной деятельности, реалий представления результатов научной деятельности в информационном пространстве и пр. в статье предлагаются способы решения возникающих проблем.

Ключевые слова: научная сеть, Research 2.0, Science 2.0, CRIS, нечеткий поиск, базы данных, научная деятельность

SOME ASPECTS OF THE DEVELOPMENT OF INFORMATION SYSTEMS FOR THE COLLECTION AND STORAGE OF SCIENTIFIC AND SCIENTOMETRIC INFORMATION

Umarov Adam S., student, Astrakhan State University, 20a Tatishchev St., Astrakhan, 414056, Russian Federation, e-mail: mathmod@bk.ru

Popova Natalya V., undergraduate student, Astrakhan State University, 20a Tatishchev St., Astrakhan, 414056, Russian Federation, e-mail: mathmod@bk.ru

[1] Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 12-07-31145 мол_а («Разработка логико-концептуальной модели информационной системы для обмена информацией внутри научного интернет-сообщества») и РГНФ в рамках проекта № 12-03-12000 («Разработка системы сбора, структурирования, анализа и представления научной и наукометрической информации на уровне научной организации (подразделения)»).

Zelepukhina Viktoriya A., Ph.D. (Engineering), Associate Professor, Astrakhan State University, 20a Tatishchev St., Astrakhan, 414056, Russian Federation, e-mail: mathmod@bk.ru

The article is devoted to the problems of creation and maintenance of scientific internet-networks and current research information systems (CRIS). Users of such information systems have the opportunity to add and store information about publications, patents, grants, awards etc. The paper describes the basic principles that are typically included into the conceptual model of such systems. However, there are some problems that can destroy the concept and cause the appearance of duplicates, the ambiguity of the authors names and the difficulty of classification of publications. The paper describes approaches that can solve these problems.

Keywords: scientific network, Research 2.0, Science 2.0, CRIS, fuzzy search, data bases, scientific activity

В настоящее время актуальным направлением в области информационных технологий (ИТ) является создание информационных систем (ИС) для сбора и хранения информации о результатах научной деятельности (РНД). Примерами таких систем являются научные сети (НС) [10, 15, 16, 18] и ИС текущих исследований (Current Research Information Systems (CRIS)) [2]. В РНД, как правило, включают информацию о публикациях и иных объектах интеллектуальной собственности, о полученных грантах, сведения о дипломах и наградах.

На кафедре авторским коллективом накоплен некоторый опыт в создании подобных систем. В частности, были разработаны научная сеть для поддержки комплексных междисциплинарных исследований и CRIS Астраханского государственного университета [5]. Обе системы развиваются согласно следующей концепции.

А. **Использование подхода Web 2.0.** Информация в базе данных (БД) ИС формируется силами самих пользователей. Причем иерархическая структура сообщества (в случае НС) и организации (в случае CRIS) позволяет автоматизировать заполнение профиля отдельных исследовательских групп и подразделений за счет интеграции информации, входящей в состав нижестоящих структур (рис.).

В. **Хранение информации о результате научной деятельности в единственном экземпляре** для того, чтобы исключить появление дублирующейся информации, искажающей наукометрические показатели и рейтинги. Для отображения информации о РНД коллектива авторов на страницах всех соавторов, зарегистрированных в системе, используются связи вида «Публикация – Автор – Пользователь», «Интеллектуальная собственность – Автор – Пользователь», «Грант – Руководитель – Пользователь», «Сведения об обучении – Руководитель – Пользователь», «Сведения о защите диссертации – Оппонент – Пользователь», «Сведения о защите диссертации – Руководитель – Пользователь» и др.

Указанные принципы заложены и в крупные научные сети (например, ResearchGate [12]). Они имеют следующие преимущества:

- уменьшение трудоемкости формирования БД РНД (для общего РНД один пользователь заносит информацию для всех соавторов);
- устранение избыточности данных и обеспечение отсутствия дубликатов, что важно при подсчете количественных показателей РНД для формирования статистических отчетов;
- уменьшение трудоемкости для пользователей ИС работ, связанных с поиском и анализом информации о РНД.

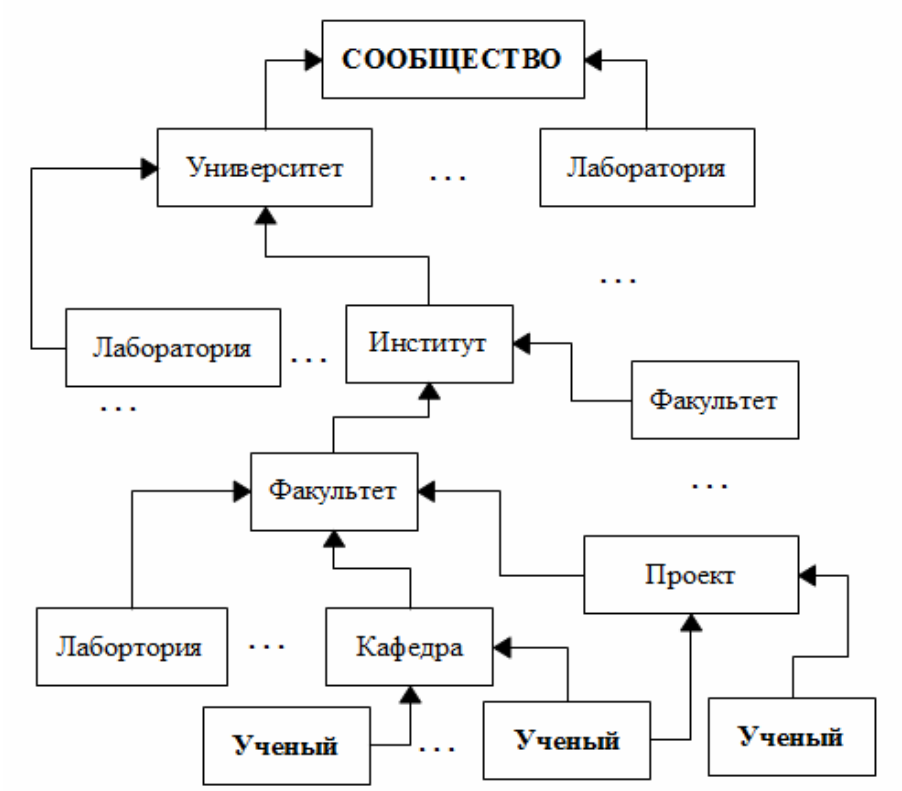

Рис. Типичная иерархическая структура данных в системах сбора и хранения научной и наукометрической информации

Однако в процессе эксплуатации и поддержки таких систем мы столкнулись с рядом трудностей, нарушающих заложенную концепцию. Эти сложности описываются с учетом практического опыта эксплуатации разработанной авторами ИС (см. выше).

А. **Повторный ввод данных и появление дубликатов.** В процессе добавления библиографической информации в БД пользователи не всегда устанавливают связи с другими участниками системы – соавторами по общему РНД. Таким образом, возможна ситуация, при которой информация, например, о публикации будет добавлена в БД и представлена в ИС столько раз, сколько соавторов данной работы зарегистрировано в качестве пользователей ИС. В частности, при первичном запуске разработанной нами ИС на один РНД в среднем приходился один случай повторного ввода. При этом в подавляющем большинстве случаев идентифицировать такие дубликаты с применением средств классической логики невозможно: ручной ввод данных искажает информацию о РНД за счет опечаток, использования сокращений и аббревиатур, указания неверного порядка следования соавторов для общего РНД и др. Например, «*Математическое моделирование*» и «*Мат. моделирование*» – это названия одного и того же журнала, но по-разному введенные пользователями в БД ИС. Заголовок публикации «*Компьютерный эксперимент в физи-ческом практикуме*» содержит лишний знак, получившийся вследствие копирования библиографической ссылки из текстового документа с включенным режимом автоматической расстановки переносов. Дополнительная информация о наиболее распространенных ошибках пользователей при формировании библиографических записей для научных статей доступна на сайте American Physical Society [11].

Неоднозначность идентификации авторов РНД. В процессе функционирования ИС случается так, что к моменту регистрации нового пользователя информация о части

его РНД уже внесена другими участниками. Следовательно, необходима правильная идентификация таких записей и создание соответствующих связей с профилем нового участника. Поставленная задача осложняется следующими факторами:

- для одной фамилии автора РНД может быть найдено несколько соответствий среди пользователей ИС. Количество таких соответствий зависит от объема БД ИС и, в частности, от числа зарегистрированных субъектов научной деятельности;
- возможны ошибки (опечатки) в написании фамилии автора РНД. Особенно часто такие ошибки возникают в процессе автоматического импортирования атрибутов для РНД из слабоструктурированных данных;
- могут быть указаны неполные данные (фамилия без инициалов);
- ученый может иметь публикации на разных языках, причем его имя может быть записано с использованием разных правил транслитерации. Например, Teplyi D. L., Tëplyĭ D. L., Teplyĭ D. L., Teplyi D. L., Teply D. L., Tyopliy D. L., Teplyj D. L. – это один и тот же автор;
- ученый ранее мог публиковаться под другими именами и псевдонимами. Например, в случае выхода женщины замуж и, соответственно, смены ею фамилии.

С. **Сложность автоматического присвоения опциональных характеристик публикациям.** Научным работам можно поставить в соответствие следующие характеристики: статья может быть опубликована в журнале, включенном Высшей аттестационной комиссией (ВАК) в перечень российских рецензируемых научных журналов, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученых степеней доктора и кандидата наук [4], индексироваться базами данных Scopus [14], Web of Science (WoS) [17] и др. Данная информация важна при проведении библиометрического анализа [1], однако автоматизация процесса верификации и актуализации таких сведений является сложной задачей. Это связано с тем, что известные системы, агрегирующие научную и наукометрическую информацию, либо не предоставляют стандартного Application Programming Interface (API), либо предоставляют, но только по подписке, которая имеется далеко не у всех организаций.

Для решения части вышеуказанных проблем нами был проведен целый комплекс дополнительных мероприятий, связанных с разработкой алгоритмов и созданием необходимого программного обеспечения.

а. Для реализации нечеткого сравнения строк были исследованы основные модели информационного поиска [3], в частности, алгоритм шинглов, алгоритмы расширения выборки, алгоритм *n*-грамм, алгоритмы, основанные на вычислении расстояния редактирования. Алгоритмы, основанные на использовании расстояния редактирования, требовательны к ресурсам, поэтому их целесообразно использовать при сравнении строк малой длины. Гораздо менее ресурсоемки алгоритмы, основанные на использовании *n*-грамм, которые представляют собой последовательность из *n* элементов. В области нечеткого информационного поиска *n*-грамма определяется как ряд слов или букв. Предполагается, что если большинство *n*-грамм двух строк совпадают, то строки можно считать похожими. Например, если разложить две фамилии «Иванов» и «Ивнов» на биграммы («**_и ив** ва ан **но ов в_**» и «**_и ив** вн **но ов в_**», соответственно), то с применением какой-либо меры (например, Джаккарда), можно сделать вывод о близости данных строк.

К преимуществам алгоритмов, основанных на использовании *n*-грамм, можно отнести высокое быстродействие, к недостаткам – разрастание размера БД за счет необходимости предварительного создания и хранения *n*-грамм для всех записей. Алгоритмы, использующие *n*-граммы, показали наибольшую скорость поиска за счет возможности построения инвертированного индекса в БД и сведения задачи к обычному полнотекстовому поиску.

b. **Модуль автоматического импортирования библиографической информации из базы данных CrossRef** [8] посредством бесплатного API [7]. Подавляющему большинству научных работ, опубликованных за рубежом, присваивается постоянный идентификатор Digital Object Identifier (DOI) [9], который представляет собой последовательность символов, состоящую из двух частей. Первая часть – префикс издателя, определяемый при его первичной регистрации в агентстве CrossRef. Вторая часть – суффикс, формируемый издателем для принятой к публикации работы. Следует отметить, что российские издатели, как правило, в проекте CrossRef не участвуют. Поэтому данная подсистема используется в основном для добавления информации о зарубежных публикациях.

Чтобы найти дубликаты среди записей, полученных из CrossRef, достаточно сравнивать значения DOI. Это избавляет от необходимости нечеткого сравнения всех атрибутов публикации (тип, авторы, заголовок, название журнала, год, название сборника и т.д.). Весьма целесообразным было бы создание программы, которая аналогичным образом импортировала библиографические данные публикаций посредством ввода ISBN и ISSN пользователем ИС. Однако отечественные организации, которые осуществляют учет выпускаемой на территории Российской Федерации печатной продукции, такие возможности предоставляют только на платной основе.

c. **Подсистема для классификации библиографических записей о публикациях в полуавтоматическом режиме.** Для создания такого модуля мы использовали информацию из открытых источников, размещенных на сайтах ВАК [7] и Sciverse [13]. Информация была импортирована в БД разработанной авторами ИС. Таким образом, для определения того, входит ли журнал, в котором опубликована научная работа, в «список ВАК» или индексируется ли в Scopus, достаточно проверить наличие названия журнала добавляемой публикации в той или иной таблице БД.

Для реализации возможности нечеткого поиска по данным таблицам был сформирован инвертированный индекс по *3*-граммам названий. Названия предварительно прошли процесс нормализации: приведение к единому регистру, удаление стоп-слов и знаков пунктуации, стеммизация [16], генерация *3*-грамм.

Для учета сокращений и аббревиатур в названиях журналов были сгенерированы альтернативные варианты названий с учетом информации из ГОСТ Р 7.0.12-2011 «Система стандартов по информации, библиотечному и издательскому делу. Библиографическая запись, сокращение слов и словосочетаний на русском языке. Общие требования и правила» (раздел «Перечень особых случаев сокращений слов и словосочетаний на русском языке в библиографической записи») [6].

Поиск по списку Scopus производится разработанной нами системой без применения алгоритмов нечеткого поиска, т.к. для названий зарубежных журналов существуют списки стандартных сокращений [19]. Кроме того, возможность импортирования данных посредством системы CrossRef позволяет упростить процедуру поиска и ограничиться штатными средствами системы управления БД.

d. **Модуль синхронизации информации между персональными страницами ученых.**

Модуль работает в двух режимах: в виде фоновой программы и на уровне интерфейса с пользователем. Таким образом, указать связь автора с конкретным пользователем системы можно и в процессе ввода атрибутов РНД, и по требованию администратора системы. Для учета рассмотренных в статье особенностей в БД системы были добавлены отношения, в которых хранятся варианты фамилий пользователя и его псевдонимы. Кроме того, каждый вариант фамилии и инициалов автора проходит процесс транслитерации для идентификации связи с его зарубежными публикациями.

Нечеткий поиск по вариантам фамилий и псевдонимам пользователей использует *n*-граммы для ускоренной выборки «кандидатов на совпадение» и алгоритм нахождения расстояния Левенштейна для выбранных объектов с целью уточнения результата поиска.

е. **Модуль идентификации нечетких дубликатов в БД РНД.** Модуль использует *3*-граммы для нечеткого сравнения атрибутов публикаций. Причем для повышения скорости сравнения публикаций каждый раз при добавлении новой публикации или обновлении характеристик существующей автоматически генерируются *3*-граммы и сохраняются в БД. В процессе сравнения атрибутов библиографической записи набираются некоторые баллы: если в результате всех сравнений их сумма оказывается больше некоторого порогового значения, то программа «принимает решение» о том, что записи действительно являются описанием одного и того же объекта. Найденные потенциальные дубликаты группируются в кластеры, информация о которых также записывается в БД.

Все описанные программно-алгоритмические решения имеют полуавтоматическую реализацию и носят в большей степени рекомендательный характер. Таким образом, пользователь системы или ее администратор (в случае с потенциальными дубликатами) должен принимать окончательное решение о подтверждении или отклонении рекомендаций программы.

Учитывая рассмотренную в настоящей статье концепцию и особенности созданной нами ИС для сбора и хранения научной и наукометрической информации, целесообразно выделить следующие компоненты в качестве основных составляющих таких систем:

- БД, хранящая информацию о субъектах научной деятельности (ученый, коллектив ученых);
- БД, хранящая информацию об объектах научной деятельности (публикации, иные объекты интеллектуальной собственности, финансирование и гранты, проекты);
- средства для анализа и идентификации связей между субъектами и объектами;
- средства устранения избыточности в БД на основе идентификации нечетких дубликатов записей о РНД;
- средства для сопряжения с внешними источниками научной и наукометрической информации.

Применение совокупности указанных компонентов позволит оперативно производить наполнение БД и будет способствовать своевременной синхронизации и верификации научной и наукометрической информации.

Список литературы

1. Брумштейн Ю. М. Публикационная политика регионального вуза в контексте управления его научным имиджем / Ю. М. Брумштейн, А. Б. Кузьмина, Л. В. Яковлева // Прикаспийский журнал: управление и высокие технологии. – 2013. – № 2 (22). – С. 99–109.

2. Зелепухина В. А. Концепция информационно-аналитической системы для сбора и анализа научной и наукометрической информации в организации / В. А. Зелепухина, Ю. Ю. Тарасевич // Информатизация образования и науки. – 2013. – № 2(18) – С. 133–144.

3. Кузнецов М. А. Математические модели информационного поиска web-ресурсов / М. А. Кузнецов, Т. Т. Нгуен // Прикаспийский журнал: управление и высокие технологии. – 2013. – № 2 (22). – С. 25–30.

4. Перечень российских рецензируемых научных журналов. – Режим доступа: http://vak.ed.gov.ru (дата обращения 09.08.2013), свободный. – Заглавие с экрана. – Яз. рус.

5. Результаты научной деятельности. – Режим доступа: http://science.aspu.ru (дата обращения 09.08.2013), свободный. – Заглавие с экрана. – Яз. рус.

6. Система стандартов по информации, библиотечному и издательскому делу. Библиографическая запись. Сокращение слов и словосочетаний на русском языке. Общие требования и правила. – Режим доступа: http://protect.gost.ru/document.aspx?control=7&id=179586 (дата обращения 09.08.2013), свободный. – Заглавие с экрана. – Яз. рус.

7. CrossRef Search API. – Available at: http://search.crossref.org/help/api (accessed 9 August 2013).

8. CrossRef. – Available at: http://crossref.org (accessed 9 August 2013).
9. Digital Object Identifier System. – Available at: http://www.doi.org (accessed 9 August 2013).
10. Duffy A. Shifting the Research Grant Collaboration Paradigm with Research 2.0 / A. Duffy // e-Research Collaboration. Theory, Techniques and Challenges. – Berlin – Heidelberg : Springer-Verlag Berlin Heidelberg, 2010. – P. 219–232.
11. Examples of Common Errors in References. – Available at: http://publish.aps.org/authors/examples-errors-references (accessed 9 August 2013).
12. ResearchGate. – Available at: http://researchgate.net (accessed 9 August 2013).
13. SciVerse. – Available at: http://www.info.sciverse.com/documents/files/scopus-training/resourcelibrary/xls/title_list.xlsx (accessed 9 August 2013).
14. Scopus. – Available at: http://www.scopus.com (accessed 9 August 2013).
15. Tacke O. Open Science 2.0: How Research and Education can benefit from Open Innovation and Web 2.0. / O. Tacke // On Collective Intelligence. – Berlin : Springer Berlin Heidelberg, 2011. – P. 37–48.
16. The Porter stemming algorithm. – Available at: http://snowball.tartarus.org/algorithms/porter/stemmer.html (accessed 9 August 2013).
17. Thomson Reuters Web of Science. – Available at: http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/ (accessed 9 August 2013).
18. Ullmann T. D. Components of a Research 2.0 Infrastructure / T. D. Ullmann, F. Wild, P. Scott et al. // Sustaining TEL: From Innovation to Learning and Practice : Proceedings of the 5th European Conference on Technology Enhanced Learning, EC-TEL 2010. Barcelona, Spain, September 28 – October 1, 2010. – Berlin – Heidelberg : Springer-Verlag Heidelberg, 2010. – P. 590–595.
19. WOS, Journal Title Abbreviations. – Available at: http://images.webofknowledge.com/WOK46/help/WOS/A_abrvjt.html (accessed 9 August 2013).

References

1. Brumshteyn Yu. M., Kuzmina A. B., Yakovleva L. V. Publikatsionnaya politika regionalnogo vuza v kontekste upravleniya ego nauchnym imidzhem [Publication policy of regional university in context of its scientific image management]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2013, no. 2 (22), pp. 99–109.
2. Zelepukhina V. A., Tarasevich Yu. Yu. Kontseptsiya informatsionno-analiticheskoy sistemy dlya sbora i analiza nauchnoy i naukometricheskoy informatsii v organizatsii [The concept of data-processing system for the collection and analysis of scientific and scientometric information in the organization]. *Informatizatsiya obrazovaniya i nauki* [Education and Science Informatization], 2013, no. 2(18), pp. 133–144. 3. Kuznetsov M. A., Nguen T. T. Matematicheskie modeli informatsionnogo poiska web-resursov [Mathematical models of web resources search]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2013, no. 2 (22), pp. 25–30.
4. The list of Russian peer-reviewed scientific journals. Available at: http://vak.ed.gov.ru (accessed 9 August 2013). (In Russ.).
5. The results of scientific activity. Available at: http://science.aspu.ru (accessed 9 August 2013). (In Russ.).
6. The system of standards on information, librarianship and publishing. Bibliographic record. Abbreviations of words and phrases in Russian. General requirements and rules. Available at: http://protect.gost.ru/document.aspx?control=7&id=179586 (accessed 9 August 2013). (In Russ.).
7. CrossRef SearchAPI. Available at: http://search.crossref.org/help/api (accessed 9 August 2013).
8. CrossRef. Available at: http://crossref.org/ (accessed 9 August 2013).
9. Digital Object Identifier. Available at: http://www.doi.org (accessed 9 August 2013).
10. Duffy A. Shifting the Research Grant Collaboration Paradigm with Research 2.0. *e-Research Collaboration. Theory, Techniques and Challenges*. Berlin – Heidelberg, Springer-Verlag Berlin Heidelberg, 2010, pp. 219–232.
11. Examples of Common Errors in References. Available at: http://publish.aps.org/authors/examples-errors-references (accessed 9 August 2013).
12. ResearchGate. Available at: http://researchgate.net (accessed 9 August 2013).
13. SciVerse. Available at: http://www.info.sciverse.com/documents/files/scopus-training/resourcelibrary/xls/title_list.xlsx (accessed 9 August 2013).

14. Scopus. Available at: http://www.scopus.com (accessed 9 August 2013).

15. Tacke O. Open Science 2.0: How Research and Education can benefit from Open Innovation and Web 2.0. *On Collective Intelligence*. Berlin, Springer Berlin Heidelberg, 2011, pp. 37–48.

16. The Porter stemming algorithm. Available at: http://snowball.tartarus.org/algorithms /porter/stemmer.html (accessed 9 August 2013).

17. Thomson Reuters Web of Science Available at: http://thomsonreuters.com/products_services /science /science_products/a-z/web_of_science (accessed 9 August 2013).

18. Ullmann T. D., Wild F., Scott P.et al. Components of a Research 2.0 Infrastructure. *Sustaining TEL: From Innovation to Learning and Practice* : Proceedings of the 5th European Conference on Technology Enhanced Learning, EC-TEL 2010. Barcelona, Spain, September 28 – October 1, 2010. Berlin – Heidelberg, Springer-Verlag Heidelberg, 2010, pp. 590–595.

19. WOS, Journal Title Abbreviations. Available at: http://images.webofknowledge.com /WOK46/help/WOS/A_abrvjt.html (accessed 9 August 2013).