Архитектура единого цифрового пространства научных знаний

Н. Е. Каленов, Г. И. Савин, А. Н. Сотников

МСЦ РАН

Аннотация. Архитектура Единого цифрового пространства научных знаний (ЕЦПНЗ) определяется его функциями и задачами. ЕЦПНЗ включает совокупность подпространств, относящихся к различным научным направлениям. Единство подпространств обеспечивается унифицированными принципами построения подпространств и онтологическими связями между их объектами. Каждое подпространство включает цифровые объекты, метаданные, содержащие факты, связанные с объектами, и предметные онтологии, обеспечивающие развитый поиск и навигацию по пространству. Вся информация, отражается в ЕЦПНЗ по правилам «semantic WEB». Контент каждого подпространства включает ядро (проверенные временем достоверные научные результаты) и надстройку — новые научные результаты, прошедшие предварительную экспертизу.

Ключевые слова: научная информация, структура пространства знаний, онтологии, контент.

The Architecture of the Common Digital Space of Scientific Knowledge

N. E. Kalenov, G. I. Savin, A. N. Sotnikov

JSCC RAS

Abstract: The architecture of the Common Digital Space of Scientific Knowledge (CDSSK) is determined by its functions and objectives. CDSSK includes a set of subspaces related to various scientific fields. The unity of subspaces is provided by unified principles for constructing subspaces and ontological connections between their objects. Each subspace includes digital objects, metadata containing facts related to objects, and subject ontologies that provide advanced searches and navigation through space. All information is reflected in the CDSSK according to the rules of the «semantic WEB». The content of each subspace includes a core (time-tested reliable

scientific results) and a superstructure – new scientific results that have passed preliminary examination.

Keywords: scientific information, structure of knowledge space, ontologies, content

Современное информационное пространство содержит огромные объемы научной информации, отображаемой в публикациях и базах данных, представленных в цифровом виде. Эти объемы нарастают с увеличивающейся скоростью по всем областям науки. Если говорить об абсолютных цифрах, то в качестве примера можно привести данные о ежегодном количестве опубликованных статей в журналах и материалах конференций, отраженных в Российском индексе научного цитирования (РИНЦ) [1] по трем принципиально различным областям науки – информатике, микробиологии и языкознанию. Эти данные за последнее десятилетие по трем годам (2010, 2014 и 2018) приведены в табл. 1.

Таблица 1

Динамика роста количества публикаций

Область науки	Вид публикации	2010	2014	2018
Информатика	Статьи в журналах	24968	40293	36760
	Материалы конференций	7501	17591	36364
	Итого	32469	57884	73124
Микробиология	Статьи в журналах	10471	11607	11987
	Материалы конференций	999	2178	4213
	Итого	11470	13785	16200
Языкознание	Статьи в журналах	25900	47182	50064
	Материалы конференций	8047	20670	47745
	Итого	33947	67852	97809

Хотя динамика роста количества публикаций различна для различных научных направлений, и ее анализ не входит в задачи данной статьи, из приведенного примера видно, что во всех трех областях науки количество публикуемых статей

год от года возрастает и достаточно существенно. Чтобы быть в фарватере развития науки, исследователю необходимо обладать новейшей и точной информацией в его области исследований. Предоставление ученым той и только той информации, которая им необходима в каждый момент времени, традиционно являлось в России основной задачей научных библиотек и информационных центров. Развитие информатики, ее технической и алгоритмической базы, замена печатных изданий цифровыми породили тезис «в интернете всё есть», следствием чего явилось переключивание задачи информационного сопровождения научных исследований на плечи самих ученых. Этот подход может привести к тому, что исследователь будет тратить львиную долю своего времени и интеллектуальных усилий на ознакомление с информацией вместо того, чтобы заниматься исследованиями. С другой стороны, если он этого не сделает, отсутствие информации может привести к дублированию исследований и бессмысленной затрате времени на получение результатов, ранее уже полученных другими исследователями.

Создание Единого цифрового пространства научных знаний (ЕЦПНЗ) [2, 3], содержащего достоверную и не дублированную научную фактографическую и документальную структурированную информацию по различным разделам науки, является одним из направлений решения проблемы информационного сопровождения научных исследований. Но это — лишь одна из целей создания ЕЦПНЗ.

Общая цель создания ЕЦПНЗ — формирование и поддержка цифровой информационной среды, необходимой для решения комплекса задач развития общества, обеспечивающих:

- информационное сопровождение научных исследований;
- поддержку образовательных процессов, начиная от средней школы (тесно связанных с популяризацией нау-

ки) и кончая аспирантурой (напрямую связанных с информационным сопровождением науки);

- популяризацию науки (развитие мотивации для занятий наукой и получения соответствующего образования, одобрение финансирования науки со стороны общества).

- сохранность научных знаний;

- процессы мониторинга и управления наукой.

Таким образом, ЕЦПНЗ должно выполнять научно-информационные, образовательные, общекультурные и управленческие функции.

Соответственно, ЕЦПНЗ должно содержать элементы, ориентированные на следующие категории пользователей:

- исследователи, которым должна быть предоставлена многоаспектная ретроспективная и текущая (отфильтрованная по различным критериям, обеспечивающим ее достоверность и новизну ¹) информация по соответствующему их интересам научному направлению;

- школьники и студенты, которые должны получать проверенную временем достоверную базовую информацию различного уровня; эта информация должна включать факты со ссылками на учебные материалы, собственно тексты классических учебников (отбор их также должен осуществляться на основе независимых критериев), цифровые модели тех или иных явлений и событий;

- специалисты — аналитики и представители управленческих структур, анализирующие состояние и тенденции развития различных областей науки.

- широкая общественность, которую необходимо знакомить с наиболее интересными результатами, полученными в той или иной области науки, а также с историей научных открытий и их авторами;

- специалисты и «любители», интересующиеся историей науки и ее творцами.

¹ Фильтрация может осуществляться на основе критериев отбора источников информации, по экспертным оценкам признанных авторитетных ученых и т. п.

Задачи, решаемые в рамках создания ЕЦПНЗ, определяют его архитектуру. Пространство должно включать совокупность подпространств (ПП) по различным областям науки, построенных по единым принципам, основанным на использовании онтологических стандартов, используемых в среде semantic WEB. Контент каждого входящего в ЕЦПНЗ ПП включает три составляющих — цифровые объекты, являющиеся отражением объектов реального мира; связанные между собой и с цифровыми объектами метаданные, отражающие различные свойства каждого экземпляра цифрового объекта; предметные онтологии (тезаурусы, дополненные индексами различных классификационных систем), описывающие данную научную область. Отдельные ПП объединяются в единое пространство благодаря (а) наличию междисциплинарных связей внутри предметных онтологий и (б) наличию универсальных классов объектов, связанных с различными областями науки.

К универсальным классам относятся объекты, онтологически не связанные с какой-то конкретной областью науки. Профили их метаданных, представленных в ЕЦПНЗ, не зависят от того, к какому тематическому подпространству они относятся. К универсальным принадлежат такие классы объектов как «событие», «персона», «публикация», «организация», «архивный документ» и др.

В каждом подпространстве, наряду с универсальными классами объектов, присутствуют свои (локальные) классы, специфичные для данного научного направления. Например, такие классы, как «теорема», «уравнение», «множество», характерны для математики; «химический элемент», «реакция» — для химии; «корпус текста», «язык» — для языкознания.

Выявление локальных классов объектов, относящихся к тому или иному подпространству ЕЦПНЗ, определение профилей метаданных объектов каждого класса и видов связей между объектами одного и различных классов являются основными задачами при проектировании любого тематического подпространства ЕЦПНЗ.

Исходя из задач, решаемых в рамках ЕЦПНЗ, каждое подпространство должно включать «базис» — фундамен-

тельную, проверенную временем, информацию, относящуюся к данному научному направлению, и «надстройку» — новые научные результаты, прошедшие предварительную экспертизу. Базис представляет собой совокупность законов, постулатов, основных результатов, полученных в данной научной области, с отсылками на источники, в которых они опубликованы, и полные тексты этих источников. Контент базиса включает три взаимосвязанных уровня — образовательный, научно-популярный и фундаментальный.

- Образовательный уровень, рассчитанный, в частности, на учащихся школ, формируется на основе информации, полученной из фундаментальных учебников, и включает, наряду с фактографической информацией и полными текстами учебников, верифицированные мультимедийные ресурсы и, в частности, оцифрованные музейные объекты. Контент этого уровня является, в основном, статическим — информация обновляется достаточно редко.

- Научно-популярный, рассчитанный на более «продвинутых» пользователей, но не специалистов в данной области. Этот уровень квазистатический. Он может изменяться при появлении новых существенных для данной области знания открытий или результатов. Формирование этого уровня строится на основе энциклопедической и справочной информации, а также монографий и учебников для высшей школы. На этом уровне наряду с извлечёнными метаданными, пользователь может получить ссылки на источники и их цифровые копии в соответствии с законом об охране авторских прав.

- Контент фундаментального уровня предназначен для специалистов в данной области науки. Этот уровень включает более глубокую, по сравнению с предыдущими двумя уровнями, информацию по данному научному направлению, ядро этого контента является статическим и содержит многоаспектную фактографическую информацию со ссылками на ее источники. Если источниками являются опубликованные материалы, их полные тексты (по возможности, первые публикации, представляющие историческую

ценность) также должны входить в контент базиса данного ПП.

Контент надстройки подпространства рассчитан на исследователей, работающих в данной области науки, и содержит новые научные результаты, опубликованные в авторитетных журналах или отраженные в авторских свидетельствах. По прошествии определенного времени, по результатам оценки экспертным сообществом, отдельные составляющие надстройки могут быть переведены в базис или удалены из ЕЦПНЗ.

Технологически ЕЦПНЗ должно включать следующие взаимосвязанные составляющие: центральное ядро, совокупность локальных ядер тематических подпространств и распределенный контейнер. Центральное ядро содержит метаданные объектов универсальных классов и обобщенную предметную онтологию; локальные ядра содержат метаданные объектов локальных классов, связанные между собой и с объектами универсальных классов. Распределенный контейнер содержит цифровые копии предметов реального мира (публикаций, архивных документов, музейных артефактов, научно-популярных фильмов), они могут храниться на серверах владельцев — библиотек, музеев, архивов и др.

Создание ЕЦПНЗ является проблемой национального масштаба. Для ее решения необходимо разработать общую онтологию ЕЦПНЗ; функциональную и организационную структуру пространства; проанализировать существующие возможные источники информации и сформулировать критерии отбора объектов для включения в контент пространства; разработать технологию наполнения и актуализации контента; проработать вопросы сервисной составляющей ЕЦПНЗ и правовые вопросы использования и предоставления пользователям элементов пространства; сформировать и реализовать требования к лингвистическим, техническим и программным средствам отдельных составляющих ЕЦПНЗ и пространства в целом.

Для реализации перечисленных задач у отечественных исследователей имеется серьезный задел. В качестве основы для дальнейшего развития работ в направлении создания

ЕЦПНЗ можно рассматривать исследования, связанные с созданием и поддержкой электронной библиотеки «Научное наследие России» [4, 5], проводимые в МСЦ РАН; исследования в области семантических цифровых библиотек, проводимые в ФИЦИУ [6, 7]; исследования в области развития лингвистических средств, выполняемые специалистами ВИНТИ РАН и ИНИОН РАН [8–10]. В качестве прототипов практической реализации отдельных составляющих ЕЦПНЗ могут выступать Общероссийский математический портал MathNet [11, 12], информационная система «Соционет» [13, 14], информационная система «История геологии и горного дела» [15, 16].

В качестве источников контента ЕЦПНЗ необходимо использовать энциклопедические данные, общенаучные документальные информационные системы, такие как WEB of Science, Scopus, РИНЦ, каталоги крупнейших научных библиотек, проблемно-ориентированные фактографические базы данных, информационные системы научных музеев и архивов.

Исследования в области формирования ЕЦПНЗ проводятся в МСЦ РАН — филиале ФГУ ФНЦ НИИСИ РАН в рамках госзадания 0580–2021–0016 и при поддержке РФФИ (проект 20–07–00773).

Литература

1. Российский индекс научного цитирования. <https://www.elibrary.ru/> (Дата обращения 9.11.2020)
2. Антопольский А. Б., Каленов Н. Е., Серебряков В. А., Сотников А. Н. О едином цифровом пространстве научных знаний // Вестник Российской академии, 2019. Т. 89. № 7. С. 728–735.
3. Антопольский А. Б. и др. Принципы построения и структура единого цифрового пространства научных знаний (ЕЦПНЗ) // Научно-техническая информация. Сер. 1. 2020. № 4. С. 9–17.
4. Каленов Н. Е., Савин Г. И., Серебряков В. А., Сотников А. Н. Принципы построения и формирования электронной библиотеки «Научное наследие России» // Программные продукты, системы и алгоритмы, 2012. Т. 4. № 100. С. 30–40.
5. Погорелко К. П. Динамика использования электронной библиотеки «Научное наследие России» // Информационное обеспечение

науки: новые технологии: Сборник научных трудов. М.: БЕН РАН, 2017. С. 192–200.

6. Атаева О. М., Серебряков В. А. Онтология цифровой семантической библиотеки LibMeta // Информатика и ее применения. 2018. Т. 12. № 1. С. 2–10.

7. Атаева О. М., Серебряков В. А., Тучкова Н. П. Расширение предметной области информационного запроса на основе онтологии знаний цифровой библиотеки LibMeta // Научный сервис в сети Интернет. 2019. № 21. С. 63–75.

8. Антопольский А. Б., Белоозеров В. Н., Каленов Н. Е., Маркова Т. С. О развитии терминологической базы данных в виде комплекса отраслевых информационно-поисковых тезаурусов // Информационные ресурсы России. 2018. № 5 (165). С. 22–30.

9. Белоозеров В. Н., Шапкин А. В., Щуко Ю. Н. Сеть классификационных систем ВИНТИ РАН // Программные продукты, системы и алгоритмы. 2018. № 4. С. 20–23.

10. Антопольский А. Б. О создании центра лингвистических ресурсов РАН // Известия Российской академии наук. Серия литературы и языка. 2019. Т. 78. № 4. С. 5–12.

11. Жижченко А. Б., Изаак А. Д. Информационная система Math-Net.Ru. Современное состояние и перспективы развития. Импакт-факторы российских математических журналов // Успехи математических наук. 2009. Т. 64. № 4 (388). С. 195–204.

12. Общероссийский математический портал <http://www.mathnet.ru/> (дата обращения 9.11.2020).

13. Parinov S., Lyapunov V., Puzyrev R., Kogalovsky M. Semantically Enrichable Research Information System Socionet // Communications in Computer and Information Science. 2015. Т. 518. С. 147–157.

14. Соционет. <https://socionet.ru/> (дата обращения 9.11.2020).

15. Каленов Н. Е., Малахова И. Г. Интегрированный общедоступный информационный ресурс «История геологии и горного дела» // Информационные ресурсы России, 2017. № 1. С. 19–23.

16. Информационная система «История геологии и горного дела». <http://higeo.ginras.ru/> (дата обращения 9.11.2020).

References

1. Rossijskij indeks nauchnogo citirovaniya. <https://www.elibrary.ru/> (Data obrashheniya 9.11.2020)

2. Antopol'skij A. B., Kalenov N. E., Serebryakov V. A., Sotnikov A. N. O edinom cifrovom prostranstve nauchny`x znaniy // Vestnik Rossijskoj akademii, 2019. Т. 89. № 7. S. 728–735.

3. Antopol'skij A. B. i dr. Principy` postroeniya i struktura edinogo cifrovogo prostranstva nauchny`x znaniy (ECzPNZ) // Nauchno-tekhnicheskaya informatsiya. Ser. 1. 2020. № 4. S. 9-17.
4. Kalenov N. E., Savin G. I., Serebryakov V. A., Sotnikov A. N. Principy` postroeniya i formirovaniya e`lektronnoj biblioteki «Nauchnoe nasledie Rossii» // Programmny`e produkty`, sistemy` i algoritmy`, 2012. T. 4. № 100. S. 30-40.
5. Pogorelko K. P. Dinamika ispol'zovaniya e`lektronnoj biblioteki «Nauchnoe nasledie Rossii» // Informacionnoe obespechenie nauki: novy`e tekhnologii: Sbornik nauchny`x trudov. M.: BEN RAN, 2017. S. 192-200.
6. Ataeva O. M., Serebryakov V. A. Ontologiya cifrovoy semanticheskoy biblioteki LibMeta // Informatika i ee primeneniya. 2018. T. 12. № 1. S. 2-10.
7. Ataeva O. M., Serebryakov V. A., Tuchkova N. P. Rasshirenie predmetnoj oblasti informacionnogo zaprosa na osnove ontologii znaniy cifrovoy biblioteki LibMeta // Nauchny`j servis v seti Internet. 2019. № 21. S. 63-75.
8. Antopol'skij A. B., Beloozerov V. N., Kalenov N. E., Markarova T. S. O razvitii terminologicheskoy bazy` danny`x v vide kompleksa otraslevy`x informacionno-poiskovy`x tezaurusov // Informacionny`e resursy` Rossii. 2018. № 5 (165). S. 22-30.
9. Beloozerov V. N., Shapkin A. V., Shhuko Yu. N. Set` klassifikatsionny`x sistem VINTI RAN // Programmny`e produkty`, sistemy` i algoritmy`. 2018. № 4. S. 20-23.
10. Antopol'skij A. B. O sozdanii centra lingvisticheskix resursov RAN // Izvestiya Rossijskoj akademii nauk. Seriya literatury` i yazy`ka. 2019. T. 78. № 4. S. 5-12.
11. Zhizhchenko A. B., Izaak A. D. Informacionnaya sistema Math-Net.Ru. Sovremennoe sostoyanie i perspektivy` razvitiya. Impakt-factory` rossijskix matematicheskix zhurnalov // Uspexi matematicheskix nauk. 2009. T. 64. № 4 (388). S. 195-204.
12. Obshherossijskij matematicheskij portal <http://www.mathnet.ru/> (data obrashheniya 9.11.2020).
13. Parinov S., Lyapunov V., Puzyrev R., Kogalovsky M. Semantically Enrichable Research Information System Socionet // Communications in Computer and Information Science. 2015. T. 518. S. 147-157.
14. Socionet. <https://socionet.ru/> (data obrashheniya 9.11.2020).
15. Kalenov N. E., Malaxova I. G. Integrirovanny`j obshhedostupny`j informacionny`j resurs «Istoriya geologii i gornogo dela» // Informacionny`e resursy` Rossii, 2017. № 1. S. 19-23.
16. Informacionnaya sistema «Istoriya geologii i gornogo dela». <http://higeo.ginras.ru/> (data obrashheniya 9.11.2020).