Онтология универсального подпространства Единого цифрового пространства научных знаний

Н.Е. Каленов [1], С.А. Власова [1], А.Н. Сотников [1]

[1] Межведомственный суперкомпьютерный центр Российской академии наук – филиал Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук» (МСЦ РАН – филиал ФГУ ФНЦ НИИСИ РАН)

Аннотация. Работа является развитием исследований, проводимых авторами в области создания Единого цифрового пространства научных знаний (ЕЦПНЗ). В рамках предыдущих исследований была предложена унифицированная структура представления онтологии элементов ЕЦПНЗ (подпространств, классов и атрибутов объектов, связей между объектами или атрибутами). В данной работе приводится предлагаемый авторами вариант онтологии универсального подпространства (УПП) ЕЦПНЗ, построенный в соответствии с разработанной структурой. В описываемой модели УПП выделено 10 предметных и 10 вспомогательных классов объектов. Среди предметных классов – «персоны», «публикации», «музейные предметы», «события» и др. Среди вспомогательных – «форматы», «единицы измерения», «языки» и др. В работе приводятся справочники каждого класса, построенные в соответствии с моделью структуры онтологии, перечень атрибутов объектов, их справочники и примеры статических словарей.

Ключевые слова: цифровое пространство научных знаний, онтология, классы объектов, атрибуты, структуризация, связанные данные.

Ontology of the Universal Subspace of Common Digital Space of Scientific Knowledge

N.E. Kalenov[1], S.A. Vlasova[1], A.N. Sotnikov[1]

[1] Joint SuperComputer Center of the Russian Academy of Sciences – Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences"

Abstract. The work is a development of research conducted by the authors in the field of creating a Common Digital Space of Scientific Knowledge (CDSSK). In the framework of previous studies, a unified structure for representing the ontology of the elements of the CDSSK (subspaces, classes and attributes of objects, relationships between objects or attributes) was proposed. This paper presents a variant of the ontology of the universal subspace (AXES) of the CDSSK, proposed by the authors, built in accordance with the developed structure. There are defined 10 subject and 10 auxiliary classes of objects In the described model of USS. Among the subject classes are “persons”, “publications”, “museum objects”, “events”, etc. Among the auxiliary ones are “formats”, “units of measurement”, “languages”, etc. The paper contains reference book of each class, is built in accordance with the ontology structure model, a list of object attributes, their directories and examples of static dictionaries.

Keywords: digital space of scientific knowledge, ontologies, structuring, linked data.

1. Введение

Работа, представленная в данном докладе, является продолжением исследований, проводимых в МСЦ РАН, связанных с созданием Единого цифрового пространства научных знаний (ЕЦПНЗ) как структурированной интегрированной информационной среды, отражающей достижения в различных областях науки [1, 2]. На предыдущих этапах работы была определена архитектура построения ЕЦПНЗ [3], рассмотрены вопросы сетевого обеспечения ЕЦПНЗ [4] и отражения в нем 3D- моделей мультимедийных объектов [5, 6].

В материалах конференции «Научный сервис в сети Интернет» в 2022 году был опубликован доклад, описывающий предложенную нами модель структуры онтологии ЕЦПНЗ [7]. Модифицированная версия модели представлена в [8]. Согласно этой модели ЕЦПНЗ представляется в виде 5-ти уровневой иерархической структуры (ЕЦПНЗ – подпространства – классы объектов – атрибуты объектов класса – значения атрибутов), дополненной связями трех типов – (универсальные, квазиуниверсальные и специфические). Информация обо всех элементах ЕЦПНЗ хранится в справочниках и словарях. Элементы каждого иерархического уровня описываются справочниками, фиксированной для каждого уровня структуры. Справочниками описываются также связи каждого из трех типов. Значения атрибутов и связей хранятся в словарях, информация о которых содержится в соответствующих справочниках. Словари подразделяются на две группы – статические и динамические. Первые содержат значения «стандартизованных» атрибутов (таких как перечень научных степеней, званий и должностей, рубрик классификационных систем и т.п.). Такие словари заполняются при первоначальной инсталляции ЕЦПНЗ или его фрагментов и корректируются администраторами, наделенными соответствующими правами.

Словари второго типа наполняются по мере формирования контента ЕЦПНЗ конкретными данными (фамилии персон, наименования публикаций, ссылки на сетевые ресурсы и т.п.).

2. Классы универсального подпространства и их атрибуты

Реализация предложенной структуры моделировалась на примере формирования элементов универсального подпространства ЕЦПНЗ, содержащего объекты мультидисциплинарного характера и связи между ними. В первом приближении в универсальном подпространстве выделено 9 предметных и 10 вспомогательных классов объектов.

К предметным классам отнесены:

Персоны;
Публикации;
Квалификационные работы;
Документы;
Музейные предметы;
Изображения и мультимедийные объекты
События / мероприятия;
Организации;
Политематические базы данных, каталоги ресурсов;
Награды;

В качестве вспомогательных классов выступают:

Форматы данных;
Универсальные классификационные системы
Группы персон;
Местоположение (географические характеристики);
Временные характеристики;
Мировые константы;
Единицы измерения;
Числовые значения;
Языки;
Коллекции.

Для каждого из предметных и вспомогательных классов составлены перечни атрибутов и представлены фрагменты статических словарей.

При определении атрибутов объектов учитывался многолетний положительный опыт эксплуатации электронной библиотеки «Научное наследие России [9], современная версия которой построена на совокупности связанных разнородных данных и развивается как модель составляющей ЕЦПНЗ [10,11].

В качестве примеров приведем перечень атрибутов объектов ряда предметных и вспомогательных классов. В скобках после наименования

атрибута указано, является ли он обязательным (о) или факультативным (ф)

Атрибуты предметных классов.

Атрибуты объектов класса «Персоны»:

- фамилия (о),
- имя (о);
- отчество (ф);
- псевдоним (ф);
- дата рождения (о);
- место рождения (о);
- дата смерти (ф);
- место смерти (ф);
- ученая степень (ф);
- ученое звание (ф);
- биография (о);
- библиография персоны как автора (о),
- библиография о персоне (ф).

Дополнительная информация (область научных интересов, трудовая деятельность, научные открытия, идентификаторы в базах данных и т.д.) оформляются как именованные связи с соответствующими объектами.

Атрибуты объектов класса «Квалификационные работы»:

- наименование (о);
- вид работы (о);
- дата выпуска (защиты) (ф);
- URL полного текста (ф).

Дополнительная информация оформляется в виде связей с персонами (автор, научный руководитель, оппонент и т.д.), организациями (место выполнения, место защиты, ведущая организация и т.п.), классификационными системами и пр.

Атрибуты объектов класса «Музейные предметы»:

- наименование (о);
- наименование музея, где хранится оригинал (о);
- вид источника поступления (о);
- дата поступления в музей (о);

- дата обнаружения (создания) предмета1 (o);
- описание предмета (o).

Музейные объекты могут быть связаны специфическими связями с персоной (одним из значений возможной связи является «автор сбора» - для естественнонаучных коллекций), географическим объектом (место сбора), публикациями и т.д.

Атрибуты вспомогательных классов.

Важнейшую роль в структуре онтологии ЕЦПНЗ играет класс «Форматы представления данных». Его объекты представляют собой структурированный набор правил описания атрибутов других классов. Указание на формат представления данных присутствуют во всех справочниках атрибутов объектов и связей. Информация о форматах используется для формально-логического контроля вводимых данных при потоковой загрузке, а также в процессе диалога с оператором «ручного» ввода данных. Кроме того, этот класс объектов может использоваться в качестве источника информации о тех или иных форматах данных, с которыми встречаются пользователи ЕЦПНЗ.

Атрибуты объектов класса «Форматы представления данных»:

- тип представления данных (o); атрибут может принимать значения «текст», «целое число», «дата в формате гггг[.мм[.дд]]», «связи» и т.д.);
- вид формата (ф); атрибут конкретизирует формат, например, «pdf», «jpeg», «URL», ряд значений его словаря описывают структуру представления различных видов связей ЕЦПНЗ;
- обязательное (r) или факультативное (f) значение атрибута (o);
- уникальное (u) или множественное (m) значение атрибута (o);
- ограничения по структуре (ф) (атрибут содержит наименование конкретной структуры, которой должен соответствовать тот или иной атрибут, например, «ГОСТ 7-1.2003: Библиографическое описание2» или «алгоритм контроля ISBN3», или «необходимые требования к структуре адреса электронной почты4» и т.п.);
- ссылка на подробное описание формата (ф).

Атрибуты объектов класса «Единицы измерения»:

---

[1] Очевидно, что во многих случаях дата может быть определена лишь приблизительно, но, в любом случае, соответствующая информация должна быть указана.

[2] со ссылкой на официальное описание ГОСТа

3 с приведением формулы контроля

4 с формулировкой требований

- наименование единицы измерения;
- предмет измерения;
- обозначения (аббревиатура);
- дополнительная информация.

Для автоматического формирования справочников атрибутов и словарей их значений разработана диалоговая программа, с помощью которой сформированы справочники вышеперечисленных классов объектов и атрибутов, а также элементы статических словарей их значений.

Ниже приведен ряд справочников классов, фрагментов справочников атрибутов и статических словарей.

Class.1: Персоны; UN; UNPS; A_UNPS; ; информация о персонах, в той или иной мере связанных с научными исследованиями;

A_UNPS.1: фамилия; UNFT.10; N_A_UNPS.1; D; фамилия выбирается из словаря, при отсутствии она вводится и проверяется на эквивалентность с другими написаниями;

A_UNPS.4: дата рождения; UNFT.4; N_A_UNTC.2; D; дата выбирается из словаря временных характеристик, при отсутствии вводится в словарь в соответствии с указанным форматом;

A_UNPS.8: квалификация (ученая степень); UNFT.10; N_A_UNPS.8; S; выбирается из словаря;

Фрагмент статического словаря значений атрибута «квалификация (ученая степень)»:

N_A_UNPS.8.1: доктор физ.-мат. наук

N_A_UNPS.8.2: доктор техн. наук

N_A_UNPS.8.3: доктор хим. наук

Class.3: Квалификационные работы UN; UNDS; A_UNDS; диссертации, авторефераты и т.п.

A_UNDS.1: наименование; UNFT.1; N_A_UNDS.1; D; ;

A_UNDS.2: дата выпуска (защиты) работы; UNFT.4; N_A_UNTC.2; D; ;

A_UNDS.3: вид работы; UNFT.10; N_A_UNDS.3; S; ;

A_UNDS.4: URL полного текста; UNFT.17; N_A_UNDS.4; D; заполняется в случае отсутствия текста в ЕЦПНЗ, при наличии полного текста в ЕЦПНЗ формируется специфическая связь «квалификационная работа - документ» или «квалификационная работа – изображение»;

A_UNDS.5: Дополнительная информация; UNFT.12; N_A_UNDS.5; D; ;

Фрагмент статического словаря значений атрибута «Вид работы»:

N_A_UNDS.3.1: диссертация докторская

N_A_UNDS.3.2: диссертация кандидатская

N_A_UNDS.3.3: диссертация PhD

Class.6: Музейные предметы; UN; UNMS; A_UNMS; цифровые копии предметов, хранящихся в музеях

A_UNMS.1: наименование предмета; UNFT.1; N_A_UNMS.1; D; ; ;

A_UNMS.2: наименование музея, где хранится в настоящее время оригинал; UNFT.1; N_A_UNOR.1; D; выбирается из словаря организаций, при отсутствии вводится в словарь;

A_UNMS.3: вид источника поступления; UNFT.3; N_A_UNMS.3; S; ;

A_UNMS.4: дата поступления в музей; UNFT.11; N_A_UNTC.2; D; ;

A_UNMS.5: дата обнаружения (создания) объекта; ; N_A_UNTC.2; D; проверяется по словарю временных характеристик. при отсутствии вводится в словарь;

A_UNMS.6: описание музейного предмета; UNFT.12; N_A_UNMS.6; D; ;

Фрагменты статического словаря значений атрибута «Вид источника поступления»

N_A_UNMS.3.1: приобретен музеем

N_A_UNMS.3.2: получен в дар

N_A_UNMS.3.3: изготовлен для музея

Class.16: Форматы представления данных; UN; UNFT; A_UNFT; форматы представления атрибутов объектов (число. время, дата, текст и т.п.).

A_UNFT.1: тип представления данных; ; N_A_UNFT.1; S; ;;

A_UNFT.2: вид формата; ; N_A_UNFT.2; S; ;

A_UNFT.3: обязательное (r) или факультативное (f) значение атрибута; ; N_A_UNFT.3; S; ;;

A_UNFT.4: уникальное (u) или множественное (m) значение атрибута; ; N_A_UNFT.4; S; ;

A_UNFT.5: ограничения по структуре; ; N_A_UNFT.5; D; ;

A_UNFT.6: ссылка на описание формата; ; N_A_UNFT; D; ;

Фрагменты словарей значений атрибутов объектов класса «Форматы»:

N_A_UNFT.1.1: текст

N_A_UNFT.1.2: изображение

N_A_UNFT.1.3: видео

N_A_UNFT.1.10: время в формате чч[.мм[.сек]]
N_A_UNFT.1.11: формула
N_A_UNFT.1.12: связи

N_A_UNFT.2.1: звук MP3
N_A_UNFT.2.2: текст PDF
N_A_UNFT.2.3: таблицы Excel, csv
N_A_UNFT.2.4: изображение JPG
N_A_UNFT.2.5: видео MP4
N_A_UNFT.2.6: простая связь URNc первого типа между объектами, атрибутами или значениями O1 и O2 вида <URNc>:<URNO1><URNO2>, где URNc – URN конкретной связи. Пример: наименование языка эквивалентно его коду; фамилия «Петров» эквивалентна «Petrov»; статья входит в состав энциклопедии и т.д.
N_A_UNFT.2.7: простая связь второго типа, указывающая на субъект, объект, URN связи и URN значения связи. Формат представления связи имеет вид: <URNc>:<URN субъекта><URN объекта>=<URN элемента словаря значений соответствующего атрибута связи>. Пример: персона P1 является сотрудником организации O1 в должности инженера (значение атрибута).

N_A_UNFT.3.1: r
N_A_UNFT.3.2: f

N_A_UNFT.4.1: u
N_A_UNFT.4.2: m

N_A_UNFT.5.1: арабские цифры
N_A_UNFT.5.2: библиографическое описание по ГОСТ https://docs.cntd.ru/document/1200034383
N_A_UNFT.5.3: структура адреса электронной почты – буквенно-цифровая строка, содержащая внутри символ «@», по крайней мере, одну точку не в начале и не в конце строки и не содержащая спецсимволов
N_A_UNFT.5.4: только буквы
N_A_UNFT.5.5: структура URL – строка символов, начинающаяся с http:// или https://

N_A_UNFT.6.1: https://habr.REm/ru/post/454944 [описание формата JPEG]
N_A_UNFT.6.2: https://open-file.ru/types/mp4 [описание формата mp4]

Фрагмент словаря форматов

UNFT.3: N_A_UNFT.1.1; ; N_A_UNFT3.2; N_A_UNFT.4.1; ; ; текст, только буквы, атрибут необязательный, значение уникальное]

UNFT.4: N_A_UNFT.1.7; ; N_A_UNFT3.1; N_A_UNFT.4.1; ; ;[дата в формате гггг[.мм[.дд]], атрибут обязательный, значение уникальное]

UNFT.17: N_A_UNFT.1.15; ; N_A_UNFT.3.2; N_A_UNFT.4.2; N_A_UNFT.5.5; ; [ссылка на внешний ресурс, атрибут необязательный, повторяющийся]

Class.17: Единицы измерения; UN; UNMU; A_UNMU; C_UNMU; стандартные единицы измерения различных физических величин.

A_UNMU.1: наименование единицы измерения; UNFT.14; N_A_UNMU.1; S; ;

A_UNMU.2: предмет измерения; UNFT.14; N_A_UNMU.2; S; ;

A_UNMU.3: обозначения (аббревиатура); UNFT.16; N_A_UNMU.3; S;;

A_UNMU.4: дополнительная информация; UNFT.18; N_A_UNMU.5; S; ;

Фрагменты словарей значений атрибутов класса «Единицы измерения»:

N_A_UNMU.1.1: секунда

N_A_UNMU.1.2: метр

N_A_UNMU.2.1: время

N_A_UNMU.2.2: длина

N_A_UNMU.3.1: с.

N_A_UNMU.3.2: сек.

N_A_UNMU.3.3: м.

N_A_UNMU.4.1: Определение секунды в Википедии https://ru.wikipedia.org/wiki/Секунда

N_A_UNMU.4.2: Определение секунды в Большой российской энциклопедии https://bigenc.ru/physics/text/3546123

Элемент словаря объектов класса «Единицы измерения», относящийся к секунде, будет иметь вид:

UNMU.1: N_A_UNMU.1.1: N_A_UNMU.2.1; N_A_UNMU.3.1; N_A_UNMU.3.2; N_A_UNMU.4.1; N_A_UNMU.4.2.

Заключение

Использование предложенной модели онтологии ЕЦПНЗ позволяет унифицировать алгоритмы создания контента пространства, разработать типовой интерфейс добавления новых элементов вне зависимости от подпространства и конкретного вида данных, упростить и ускорить алгоритмы поиска и навигации по связанным данным. В настоящее время в МСЦ РАН ведутся исследования по развитию и конкретизации предложенной модели в части алгоритмизации формирования вложенных связей, а также моделирования формирования фрагментов ЕЦПНЗ на примере реальных данных, в том числе, составляющих контент электронной библиотеки «научное наследие России».

Работа выполняется в МСЦ РАН – филиале ФГУ ФНЦ НИИСИ РАН в рамках государственного задания № FNEF-2023-0014

Литература

1. Антопольский А.Б., Каленов Н.Е., Серебряков В.А., Сотников А.Н. О едином цифровом пространстве научных знаний // Вестник Российской академии наук, 2019. - Т. 89, - № 7. - С. 728-735. *DOI 10.31857/S0869-5873897728-735*
2. Савин Г.И. Единое цифровое пространство научных знаний: цели и задачи // Информационные ресурсы России, 2020. - № 5. - С. 3-5. *DOI: 10.51218/0204-3653-2020-5-3-5*
3. Каленов Н.Е., Сотников А.Н. Архитектура единого цифрового пространства научных знаний // Информационные ресурсы России, 2020. - № 5. - С. 5-8. *DOI: 10.51218/0204-3653-2020-5-5-8*
4. Абрамов А.Г., Гончар А.А., Евсеев А.В. Национальная исследовательская компьютерная сеть нового поколения как инфраструктурно-сервисная платформа Единого цифрового пространства научных знаний // Информационные ресурсы России, 2020. - № 5. - С. 43-46. *DOI: 10.51218/0204-3653-2020-5-43-46*
5. Соболевская И.Н. Об особенностях представления мультимедийных объектов в едином цифровом пространстве научных знаний // Информационные ресурсы России, 2020. - № 5. - С. 31-34. *DOI: 10.51218/0204-3653-2020-5-31-34*
6. Irina Sobolevskaya Some Aspects of 3D-objects Presentation in a Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings (CEUR-WS.org), 2021., Vol. 2990, P. 117-124. *DOI: 10.51218/1613-0073-2990-117-124*

7. Каленов Н.Е., Сотников А.Н. О структуре онтологии Единого цифрового пространства научных знаний // Научный сервис в сети Интернет: труды XXIV Всероссийской научной конференции. 2022. - С. 203-221. *DOI: 10.20948/abrau-2022-23*
8. Каленов Н.Е., Сотников А.Н. Унифицированное представление онтологии единого цифрового пространства научных знаний // Электронные библиотеки, 2023. - Т. 26, - № 1. - С. 80-103. *DOI: 10.26907/1562-5419-2023-26-1-80-103*.
9. Погорелко К.П. Динамика использования электронной библиотеки "Научное наследие России" // Информационное обеспечение науки: новые технологии: Сб. науч. тр., М., 2017. - С. 192-200.
10. Konstantin Pogorelko A New Version of the Software for the Information System “Scientific Heritage of Russia” // CEUR Workshop Proceedings (CEUR-WS.org), 2021, Vol. 2990. - P. 110-116. DOI: 10.51218/1613-0073-2990-110-116
11. Каленов Н.Е., Погорелко К.П., Сотников А.Н. О развитии электронной библиотеки “Научное наследие России” как составляющей Единого цифрового пространства научных знаний // Информационные процессы , 2022. - Т. 22, - № 3. - С. 155-166. *DOI: 10.53921/18195822_2022_22_3_155*

References

1. Antopol'skiy A.B., , Kalenov N.Ye., Serebryakov V.A., Sotnikov A.N., // O yedinom tsifrovom prostranstve nauchnykh znaniy // Vestnik Rossiyskoy akademii nauk, 2019. - T. 89, - № 7. - S. 728-735
2. Savin G.I. Edinoe cifrovoe prostranstvo nauchny`x znanij: celi i zadachi // Informacionny`e resursy` Rossii, 2020. - № 5. - S. 3-5. - https://doi.org/ 10.51218/0204-3653-2020-5-3-5.
3. Kalenov N.Ye., Sotnikov A.N. Arkhitektura yedinogo tsifrovogo prostranstva nauchnykh znaniy // Informatsionnyye resursy Rossii, 2020. - № 5. - S. 5-8. DOI: 10.51218/0204-3653-2020-5-5-8.
4. Abramov A.G., Gonchar A.A., Yevseyev A.V. Natsional'naya issledovatel'skaya komp'yuternaya set' novogo pokoleniya kak infrastrukturno-servisnaya platforma Yedinogo tsifrovogo prostranstva nauchnykh znaniy // Informatsionnyye resursy Rossii, 2020. - № 5. - S. 43-46
5. Sobolevskaya I.N. Ob osobennostyakh predstavleniya mul'timediynykh ob"yektov v yedinom tsifrovom prostranstve nauchnykh znaniy //

Informatsionnyye resursy Rossii, 2020. - № 5. - S. 31-34. DOI: 10.51218/0204-3653-2020-5-31-34
6. Irina Sobolevskaya Some Aspects of 3D-objects Presentation in a Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings (CEUR-WS.org), 2021., Vol. 2990, P. 117-124. *DOI: 10.51218/1613-0073-2990-117-124*
7. Kalenov N.Ye., Sotnikov A.N. O strukture ontologii Yedinogo tsifrovogo prostranstva nauchnykh znaniy // Nauchnyy servis v seti Internet: trudy XXIV Vserossiyskoy nauchnoy konferentsii. 2022. - S. 203-221. DOI: 10.20948/abrau-2022-23
8. Kalenov N.Ye., Sotnikov A.N. Unifitsirovannoye predstavleniye ontologii yedinogo tsifrovogo prostranstva nauchnykh znaniy // Elektronnyye biblioteki, 2023. - T. 26, - № 1. - S. 80-103. DOI: 10.26907/1562-5419-2023-26-1-80-103.
9. Pogorelko K.P. Dinamika ispol'zovaniya elektronnoy biblioteki "Nauchnoye naslediye Rossii" // Informatsionnoye obespecheniye nauki: novyye tekhnologii: Sb. nauch. tr., M., 2017. - S. 192-200.
10. Konstantin Pogorelko A New Version of the Software for the Information System “Scientific Heritage of Russia” // CEUR Workshop Proceedings (CEUR-WS.org), 2021, Vol. 2990. - P. 110-116. DOI: 10.51218/1613-0073-2990-110-116
11. Kalenov N.Ye., Pogorelko K.P., Sotnikov A.N. O razvitii elektronnoy biblioteki “Nauchnoye naslediye Rossii” kak sostavlyayushchey Yedinogo tsifrovogo prostranstva nauchnykh znaniy // Informatsionnyye protsessy , 2022. - T. 22, - № 3. - S. 155-166. DOI: 10.53921/18195822_2022_22_3_155