ИПМ им.М.В.Келдыша РАН

Абрау-2022 • Труды конференции

Н.Е. Каленов, А.Н. Сотников

О структуре онтологии Единого цифрового пространства научных знаний

Рекомендуемая форма библиографической ссылки

Каленов Н.Е., Сотников А.Н. О структуре онтологии Единого цифрового пространства научных знаний // Научный сервис в сети Интернет: труды XXIV Всероссийской научной конференции (19-22 сентября 2022 г., онлайн). — М.: ИПМ им. М.В.Келдыша, 2022. — С. 203-221.
https://doi.org/10.20948/abrau-2022-23
https://keldysh.ru/abrau/2022/theses/23.pdf

Видеозапись выступления

Презентация к докладу

О структуре онтологии Единого цифрового пространства научных знаний

Н.Е. Каленов [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). ЕЦПНЗ представляет собой цифровую информационную структуру, агрегирующую разнородную многоаспектную информацию, связанную с научными знаниями, и включающую совокупность подпространств, относящихся к различным тематическим научным направлениям. Источниками контента ЕЦПНЗ являются существующие информационные системы, такие как Большая российская энциклопедия, Единый каталог географических названий, Национальная электронная библиотека и др. Отличительной особенностью ЕЦПНЗ является (а) поддержка данных в структуре, соответствующей правилам semantic WEB, и (б) способность обрабатывать широкий спектр политематических запросов и предоставлять возможность навигации по разнородным ресурсам пространства, используя семантические связи между ними. Реальное проектирование ЕЦПНЗ должно начинаться с формирования онтологии пространства в целом и отдельных его подпространств. В данной работе предлагается иерархическая структуризация онтологии ЕЦПНЗ. Выделяются и определяются такие элементы как «подпространство», «класс объектов», «объект» «атрибуты объекта», три типа связей объектов и атрибутов (универсальные, квазиуниверсальные и специфические), вводится понятие «справочники» и «словари» ЕЦПНЗ. Справочники содержат информацию об атрибутах и их связях и используются в процессах формирования контента и реализации поисковой и навигационной логики; словари содержат конкретные значения атрибутов и связей. Предлагается формализация представлений этих элементов, позволяющая достаточно просто добавлять, по мере необходимости, новые атрибуты и связи между объектами.

Ключевые слова: цифровое пространство научных знаний, онтологии, структуризация, связанные данные, семантический WEB.

On the structure of the Common Digital Space of Scientific knowledge ontology

N.E. Kalenov[1], I.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 the research conducted by the authors in the field of creating a Common Digital Space of Scientific Knowledge (CDSSK). The CDSSK is a digital information structure aggregating heterogeneous polythematic information related to scientific knowledge and including a set of subspaces related to various thematic scientific areas. The sources of the CDSSK content are existing information systems, such as the Great Russian Encyclopedia, the Unified Catalog of Geographical Names, the National Electronic Library, etc. A distinctive features of the CDSSK are (a) data support in a structure that complies with the rules of the semantic WEB, and (b) the ability to process a wide range of polythematic queries and provide the ability to navigate through heterogeneous resources using semantic links between them. The real design of the CDSSK should begin with the formation of the ontology of space as a whole and its individual subspaces. In this paper, a hierarchical structuring of the CDSSK ontology is proposed. Such elements as "subspace", "class of objects", "object", "attributes of an object", three types of relations of objects and attributes (universal, quasi-universal and specific) are distinguished and defined. the concepts "reference books" and "dictionaries" of the CDSSK are introduced. Reference books contain information about attributes and their relationships and are used in the processes of content formation and implementation of search and navigation logic; dictionaries contain specific values of attributes and relationships. The formalization of these elements representations is proposed, which makes it quite simple to add, as necessary, new attributes and relationships between objects.

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

Введение

Единое цифровое пространства научных знаний (ЕЦПНЗ) - структурированная информационная среда, обеспечивающая поддержку процессов предоставления широкому кругу пользователей необходимой им информации в различных областях знаний. ЕЦПНЗ, в том числе, является информационной основой для решения задач искусственного интеллекта и строится на принципах semantic WEB. Цели создания, задачи и общие принципы построения ЕЦПНЗ приведены в [1-3]. Исследования в

области создания ЕЦПНЗ проводятся в МСЦ РАН, ФИЦ «Информатика и управление» РАН, ИНИОН РАН и в ряде других организаций.

Отличия ЕЦПНЗ как информационной системы от большинства существующих состоят в том, что оболочка ЕЦПНЗ должна обрабатывать широкий спектр запросов, не обязательно содержащих термины, в явном виде присутствующие в метаданных, относящихся к конкретным объектам ЕЦПНЗ. Например, на запрос «археологические находки Поволжья 19 века» должны быть выданы описания всех археологических объектов, найденных в Ярославской, Самарской областях, в Татарии и т.п. за период с 1801 по 1900 годы. При этом в информации об отдельном объекте может содержаться указание лишь на конкретное место его обнаружения, а заключение о том, что данное место относится к Поволжью, вытекает из автоматического анализа связей археологических объектов с объектами других категорий (в данном случае относящимися к географии и времени) Аналогично, но с добавлением анализа связей между публикациями, временными характеристиками и событиями, должны обрабатываться запросы типа «какие произведения поэтов Серебряного века были опубликованы вне России во время первой мировой войны» или «найти ныне живущих российских переводчиков литературы по физике с китайского языка».

Если говорить о практической реализации ЕЦПНЗ, то первым шагом в этом направлении является представление пространства как многоуровневой системы, состоящей из совокупности подпространств (ПП), и описание правил формирования ее составляющих, включая наполнение контента разнородными, но связанными по единым правилам, данными. Набор подобных правил представляет собой онтологию ЕЦПНЗ, но прежде чем приступать к формированию собственно онтологии, необходимо определить ее будущую структуру, что мы и попытаемся сделать ниже.

Общие подходы к формированию онтологии ЕЦПНЗ отражены в [4, 5]. Онтологии должны содержать правила описания атрибутов объектов (метаданных, характеризующих объекты в свете задач, решаемых данным пространством знаний) и их связей. Причем эти правила должны обеспечивать возможность как автоматического импорта данных из различных информационных систем, так и ручного - с использованием развитого формально-логического контроля.

Построению онтологий и правилам их отражения в сети посвящено значительное количество исследований и публикаций. В рамках Simple Knowledge Organization System (SKOS) [6, 7] разработаны формальные правила отражения в цифровой среде связанных открытых данных (LOD) тезаурусов, свойств объектов и их связей с использованием правил OWL и RDF. В обзоре [8] приводятся примеры многочисленных реализаций онтологического подхода, разработанного в рамках SKOS, применительно

к различным областям человеческой деятельности (пищевая промышленность, музейное дело, география, социальные науки и т.д.) [9 - 12].

На сайте «онтологического форума» [13] ежедневно появляется информация о семинарах, симпозиумах, рабочих встречах и т.п., посвященных проблемам создания онтологий в различных сферах человеческой деятельности.

Хотя многочисленные реализации онтологий, представленные в интернет, построены по общим принципам, каждая из них строится независимо. И обеспечить на практике интеграцию ресурсов, построенных на основе этих онтологий, достаточно затруднительно. Примеров такой интеграции нам обнаружить не удалось.

ЕЦПНЗ, в отличие от других информационных систем, должно обеспечивать реальную интеграцию разнородных данных. Этого можно достичь, только используя унифицированную, четкую и в то же время достаточно простую технологию формирования онтологии пространства в целом и его отдельных подпространств. Вариант такой технологии, не противоречащей принципиальным подходам SKOS и OWL, но являющейся фактически их развитием в сторону упрощения модели, предлагается нами ниже.

Если вернуться к примерам запросов, приведенным выше, то необходимо отметить, что для обработки подобных запросов каждый объект, отражаемый в ЕЦПНЗ, должен иметь набор собственных атрибутов (свойств), позволяющих выделить его из множества объектов данной категории (класса), и формализованные связи с другими объектами как внутри данного класса, так и вне его. Соответственно, онтология ЕЦПНЗ должна обеспечивать такую возможность и строиться, исходя из задач, специфичных именно для научного пространства. В частности, учитывая непрерывное развитие науки, структура реализации онтологии должна обеспечивать возможность достаточно простого наращивания как перечня свойств (атрибутов) объектов, так и видов связей между ними.

В соответствии с предлагаемым подходом онтология ЕЦПНЗ строится на основе иерархической системы справочников и словарей. Справочники должны в структурированном виде описывать правила формирования атрибутов и связей объектов, включаемых в ЕЦПНЗ, начиная с верхнего уровня (описание правила формирования собственно пространства и отдельных тематических подпространств) и кончая правилами представления связей между атрибутами отдельных объектов. Справочники конкретного уровня должны содержать формализованное описание справочников более низкого уровня. Словари должны содержать конкретные значения атрибутов и связей, сформированные по правилам, описанным в справочниках.

1. Общие понятия

Элементами ЕЦПНЗ являются подпространства, классы объектов, объекты, атрибуты, связи, значения атрибутов, значения связей.

Каждый элемент ЕЦПНЗ имеет своё уникальное имя (URN) и характеризуется своими атрибутами и связями с другими элементами. Под атрибутами будем понимать характеристики, присущие конкретному объекту вне контекста связей с другими объектами; под связями – вид «взаимоотношений» между парами элементов ЕЦПНЗ. И атрибуты, и связи в обязательном порядке имеют наименования на естественном языке.

ЕЦПНЗ включает универсальное и тематические подпространства (ТПП).

Универсальное подпространство содержит информацию об объектах[1] мультидисциплинарного характера (персоны, события, единицы измерения и т.п.) и их связи внутри универсального подпространства и с объектами тематических подпространств.

Тематическое подпространство (например, подпространство «информатика», «космические исследования», «химия» и т.п.) содержит элементы, напрямую связанные с данным научным направлением, а также связи с элементами универсального и других тематических подпространств.

Каждое подпространство включает:

- Классы объектов и связей, объединенных по определенному признаку.
- Совокупность справочников, описывающих структуру элементов подпространства и их связей с другими подпространствами.
- Совокупность словарей значений атрибутов и связей; В качестве значения выступает конкретная «единица информации». Это может быть текст, формула, ссылка на WEB-ресурс, изображение, мультимедиа, 3D-модель.
- Совокупность словарей объектов, наполнение которых является конечным результатом формирования контента ЕЦПНЗ.

2. Объекты

Каждый объект имеет свое уникальное в рамках ЕЦПНЗ имя (URN) и характеризуется своими значениями атрибутов и связей с другими объектами.

---

[1] В роли «объекта» может выступать цифровая копия физической сущности (например, книги, музейного предмета, архивного документа и т.п.), структурированная информация о физической сущности (персона, географическое понятие, организация и т.п.), событии или научном факте.

3. Классы объектов

Класс объектов характеризуется названием и типом – принадлежностью к универсальному или тематическому подпространству. Класс объединяет объекты, имеющие одинаковые атрибуты и виды связей с другими объектами. Каждому классу объектов соответствуют свои справочники атрибутов и связей.

4. Атрибуты

Атрибуты объектов и связей представляют собой метаданные, отражающие свойства, присущие объектам или связям данного класса. Каждый атрибут имеет свое название и соответствующий ему словарь значений. Значения, в свою очередь, могут иметь свои атрибуты и значения. Атрибуты элементов используются в качестве ограничений области поиска и визуализации найденных элементов ЕЦПНЗ.

5. Связи

Связи ЕЦПНЗ подразделяются на классы, каждый из которых характеризуется одним из типов – универсальный, квазиуниверсальный или специфический.

Связи могут быть простыми и составными. Простые связи содержат (в терминах триплетов RDF [14]) указание на субъект, объект и (факультативно, в зависимости от конкретного вида связи) значение связи. Значения составных связей могут содержать «вложения» - иметь собственные атрибуты и их значения; количество вложений не ограничено и определяется справочником данной связи.

Связи универсального типа являются простыми и указывают лишь на факт отношений между элементами и не зависят от классов объектов, которые они связывают. Они могут связывать любые элементы одного или нескольких классов. К связям этого типа относятся:

- «эквивалентно»;
- «пересекается»;
- «содержит»;
- «содержится в» (является частью, входит в состав).

Этот вид связей широко употребляется в предметных тезаурусах и при установлении соответствия между элементами классификационных систем. В ЕЦПНЗ он дополнительно используется при указании соподчиненности подразделений организаций, различных наименований организаций, различного написания фамилий и имен персон и т.п.

Квазиуниверсальные связи связывают субъекты различных классов с объектами заданного класса, они могут быть простыми или составными. Перечень квазиуниверсальных связей может пополняться по мере развития

ЕЦПНЗ и добавления новых элементов. Примером квазиуниверсальных связей могут служить ссылки на статьи в энциклопедии.

Специфические связи устанавливаются между субъектами и объектами заданных классов; они могут быть простыми и составными. Количество и вид специфических связей определяются при формировании онтологий конкретных классов. В отличие от универсальных связей, которые имеют статичный характер, и квазиуниверсальных связей, набор которых растет достаточно медленно, перечень специфических связей является достаточно динамичным, поскольку определяется развитием ЕЦПНЗ и возникающими перед ним задачами.

Примеры специфических связей объектов универсальных классов:

Простая специфическая связь между объектами класса «единицы измерения», указывающая на количество единиц субъекта, содержащихся в единице объекта; значение связи представляет собой URN объекта класса «числовые значения».

Простая связь между публикацией и организацией, которая может принимать значения «автор», «издательство», «владелец», «спонсор», «владелец авторских прав» и т.п.

Составная связь между персоной и организацией, которая может принимать значения «сотрудник», «спонсор», «акционер» и т.д. Связь «сотрудник», в свою очередь, может принимать одно из значений должности, которая имеет атрибуты «дата начала работы» и «дата окончания работы». Эти атрибуты можно определить как эквивалентные, соответственно, связям «начало временного периода» и «конец временного периода».

6. Справочники

Справочники элементов ЕЦПНЗ предназначены для формального описания атрибутов объектов, связей между ними и их значениями. Справочники формируются и дополняются администратором ЕЦПНЗ; они используются при формировании контента ЕЦПНЗ и обработке запросов. Справочники содержат информацию о том, что, куда и в каком виде вводить, какой и где реализовать формально-логический контроль при вводе данных, а также как связывать элементы запроса и различные характеристики объектов, в том числе, не присутствующие в явном виде в их атрибутах. Каждый элемент ЕЦПНЗ описывается в соответствующем справочнике. Каждый справочник в обязательном порядке содержит информацию о словарях значений атрибутов и связей, которые в нем указаны.

7. Словари

Словари содержат конкретные значения атрибутов и связей. Они могут относиться к одному или нескольким классам. Словари наполняются в процессе формирования контента ЕЦПНЗ оператором ввода или программой пакетной загрузки данных.

8. Формализация описаний элементов ЕЦПНЗ

Каждый элемент ЕЦПНЗ имеет свое уникальное имя URN, состоящее из имени справочника (или словаря), в который он входит, и порядкового номера элемента в справочнике (или в словаре), отделенного от имени точкой. В свою очередь, имя справочника может быть элементом другого справочника, поэтому URN элемента может содержать различное число точек-разделителей. Значение элемента отделяется от его URN двоеточием и пробелом. Значения элементов справочников отделяются точкой с запятой и пробелом.

Для описания структуры справочников отдельных элементов ЕЦПНЗ (подпространств, классов, атрибутов и связей разного рода) предлагается унифицированный подход, представленный в виде справочника верхнего уровня с именем CDSSK. Элемент CDSSK.1 описывает структуру справочников подпространств, элемент CDSSK.2 – справочников классов и т.д.

CDSSK.1:

Подпространство (SubSpace).

Справочник подпространств имеет имя SUBS; элемент справочника содержит три атрибута: – наименование; код типа подпространства; описание подпространства. Код типа подпространства (далее префикс – два символа), который принимает значение UN для универсального подпространства и обозначается другими символами для тематического. Код может быть представлен двумя цифрами - кодом тематики верхнего уровня ГРНТИ или двумя буквами, если ТПП относится к более узкой тематике или содержит междисциплинарную информацию. Например, подпространству «Информатика» может быть присвоен префикс 20, подпространству «Вычислительная техника» префикс HW (от англ. «hardware»).

Структура справочника подпространств.

Имя справочника подпространств SUBS

Атрибуты элементов справочника:

Наименование

Префикс ПП - 2 символа)

Описание

Примеры:

SUBS.1: Универсальное; UN; подпространство, включающее классы объектов, не связанные непосредственно с конкретной научной тематикой, в том числе универсальные справочные данные.

SUBS.2: Информатика; 20; подпространство включает объекты. относящиеся к научному направлению «информатика»

CDSSK.2.

Класс объектов (Class). Определены два типа классов объектов – универсальные и локальные. Последние принадлежат какому-либо тематическому подпространству. Элемент справочника содержит 6 атрибутов.

Структура справочника классов объектов.

Имя справочника классов: URN: Class.

Атрибуты:

Наименование

Тип (универсальный –UN, локальный -LC)

Префикс (UNxy для универсального и <ПР>xy для локального, где <ПР> префикс тематического подпространства - два символа; xy- 2 буквенно-цифровых символа)

URN словаря атрибутов

URN словаря связей.

Описание

Например:

Class_1: персоны; UN; UNPS; A_UNPS; C_UNPS; информация о персонах, в той или иной мере связанных с научными исследованиями.

Class.16: Форматы представления данных; UN; UNFT; A_UNFT; C_UNFT; форматы представления атрибутов объектов и связей.

CDSSK.3.

Класс универсальных и квазиуниверсальных связей. Элемент справочника содержит 6 атрибутов.

Структура справочника универсальных связей.

Имя справочника URN: COUN

Наименование

Необходимость справочника значений (да / нет)

URN словаря значений (или символ «-»)

URN элемента словаря значений формата данных, определяющего форму представления данной связи

URN связи, представляющей собой обратное значение данной (например, обратным значением связи «содержится в» является связь «содержит»)

Описание связи

Примеры:

COUN.1: Эквивалентность; нет; нет; URNf[2]; COUN1; используется для обозначения идентичных атрибутов или связей (разные написания фамилий и имен, переводные версии одной публикации, разные наименования одной организации синонимы терминов и т.п.)

COUN.2: Входит в состав; нет; нет; URNf COUN3; является частью: соподчиненность подразделений для организаций, статья по отношению к журналу, где она опубликована, город по отношению к стране, подчиненность терминов в тезаурусе и т.п.

COUN.3: Содержит; нет; нет; URNf; COUN2; сборник по отношению к его статьям, континент по отношению к расположенным на нем странам и т.п.

COUN.4: Пересекается; нет; нет; URNf; COUN.4; используется для обозначения частично совпадающих значений атрибутов - пересечение индексов классификаций, пустыня на территории нескольких стран, группы москвичей и школьников и т.п.

CDSSK.4:

Класс специфических связей.

Имя справочника (URN): COSP

Атрибуты:

A_COSP.1: Класс субъекта

A_COSP.2: Класс объекта

A_COSP.3: Наименование связи

A_COSP.4: URN справочника атрибутов связи

A_COSP.5: URN словаря значений связи

A_COSP.6: Формат представления связи (URN значения элемента словаря N_UNFT)

A_COSP.7: Количество подчиненных связей следующего уровня (0 - n)

Если не ноль, то:

A_COSP.8.1: Наименование подчиненной связи 1

A_COSP.8.2: URN словаря атрибутов подчиненной связи 1

A_COSP.8.3: URN словаря значений подчиненной связи 1[3]

A_COSP.8.4: Количество подчиненных связей следующего уровня (0 – n)

Если не 0, то определяется следующий блок подчиненных связей:

A_COSP.7.1.1 – A_COSP.7.1.4

И т.д.

---

[2] Имя элемента словаря форматов, содержащего структуру применения данной связи

[3] Постулируем: формат представления вложенных связей определен элементом A_COSP.6

Если A_COSP.7 > 1, формируется следующий блок описания вложенной связи.

A_COSP.8.1: Наименование подчиненной связи 2

A_COSP.7.2: URN словаря атрибутов подчиненной связи 2

A_COSP.8.2: URN словаря значений подчиненной связи 2

И т.д.

CDSSK.5:

Структура справочника атрибутов:

URN справочника (формируется в форме A_префикс класса)

Наименование атрибута

Формат представления значений атрибута (URN соответствующего элемента справочника объектов класса «Форматы данных»).

URN словаря значений атрибута (формируется в форме N_URN атрибута)

URN справочника связей значений атрибута (формируется в форме C_N_URN атрибута)

Дополнительная информация

CDSSK.6:

Структура описания справочника значений атрибутов:

URN словаря (формируется в форме N_URN атрибута)

Значение, в соответствии с форматом, указанным в ссылке словаря атрибутов

CDSSK.7:

Структура словаря универсальных связей значений атрибута

URN словаря (формируется в форме C_URN атрибута)

Триплеты вида <URN1><URNc><URN2>, где URN1 и URN2 имена значений атрибута, а URNc – имя одной из 4-х универсальных связей, не требующих определения значений (см. выше). Среди связей обязательна ссылка на формат представления атрибута.

Элементы всех словарей формируются автоматически в процессе ввода данных в ЕЦПНЗ – либо программным путем (прикладная программа пакетного ввода данных обрабатывает справочники атрибутов и связей и записывает элементы в соответствующие справочники), либо как результат диалога с оператором ввода. Во втором случае оператору предлагаются (на основе программной обработки словарей) наименования атрибутов вводимого объекта и связей с другими объектами. По каждому атрибуту и связи оператор должен выбрать уже имеющиеся в ЕЦПНЗ их значения или ввести новые с указанием значений всех необходимых связей.

9. Примеры формального описания объектов и связей

В процессе проведения исследований мы выделили в универсальном подпространстве классы объектов, которые условно разбили на предметные и вспомогательные. К предметным отнесены 10 классов, в том числе, «Персоны», «Публикации», «Документы», «События», «Организации», «Политематические базы данных» и др. К вспомогательным - 11 классов, в том числе, «Форматы данных», «Тезаурусы (предметные онтологии)», «Местоположение (географические характеристики)». «Временные характеристики», «Единицы измерения» и др.

Для каждого класса объектов сформировано их формальное описание, предложен перечень атрибутов и виды попарных специфических связей.

Рассмотрим несколько примеров.

Описание класса «форматы представления данных»

Class.16: Форматы представления данных; UN; UNFT; A_UNFT; C_UNFT; форматы представления атрибутов объектов и связей.

Каждый элемент словаря форматов UNFT содержит 6 атрибутов, определяемых элементами справочника A_UNFT, структура которого определена справочником CDSSK.5:

A_UNFT.1: тип представления данных; ; N_A_UNFT.1; ;используется для формально-логического контроля вводимых данных;

A_UNFT.2: вид формата; ; N_A_UNFT.2; ; используется при обработке данных;

A_UNFT.3: обязательное (r) или факультативное (f) значение атрибута; ; N_A_UNFT.3; ;используется для формально-логического контроля вводимых данных;

A_UNFT.4: уникальное (u) или множественное (m) значение атрибута; ; N_A_UNFT.4; ; используется для формально-логического контроля вводимых данных;

A_UNFT.5: ограничения по кодировке; ; N_A_UNFT.5; ; используется при формировании контента;

A_UNFT.6: ссылка на подробное описание формата; ; N_A_UNFT; ; используется в качестве справочного материала;

Значения атрибутов выбираются из соответствующих словарей.

Словари значений атрибутов, за исключением N_A_UNFT.3 и N_A_UNFT.4, пополняются по мере необходимости. Примеры элементов словарей:

N_A_UNFT.1.1: текст
N_A_UNFT.1.2: изображение
N_A_UNFT.1.3: видео
N_A_UNFT.1.5: любое число
N_A_UNFT.1.6: целое число

N_A_UNFT.1.7: дата в формате гггг[.мм[.дд]]
N_A_UNFT.1.8: время в формате чч[.мм[.сек]]
N_A_UNFT.1.9: время в формате гггг.мм.дд. чч[.мм[.сек]]
N_A_UNFT.1.10: связи

N_A_UNFT.2.1: TEX
N_A_UNFT2.2: PDF
N_A_UNFT.2.3: таблицы Excel, csv
N_A_UNFT.2.4: простая связь первого типа между объектами, атрибутами или значениями O1 и O2, она описывается «простым триплетом» вида <URNO1><URNc><URNO2>, где URNc – URN конкретной связи. Пример: наименование языка эквивалентно его коду; фамилия «Петров» эквивалентна «Petrov»; статья входит в состав энциклопедии
N_A_UNFT.2.5: простая связь второго типа, указывающая на субъект, объект, URN связи и URN значения связи. Формат представления связи вид: <URN субъекта><URN объекта><URNc>=<URN элемента словаря значений соответствующего атрибута связи>. Пример: персона P1 является сотрудником организации O1 (атрибут специфической связи «персона – «организация»[4]) в должности инженера (значение атрибута).
N_A_UNFT.2.6: составная связь третьего типа - «многоуровневый триплет»– случай, когда у значения атрибута связи имеются свои атрибуты с соответствующими значениями, у значений имеются атрибуты, каждый из которых, в свою очередь, имеет свое значение; Формат представления связи имеет вид: <URN субъекта><URN объекта> <URNc> = <URN элемента словаря значений соответствующего атрибута связи> <URN атрибута элемента словаря значений> = <URN значения атрибута>. Пример: персона P1 является сотрудником организации O1, работает в должности инженера с такой-то даты
<URN P1> <URN O1><URNc>=<URN значения «сотрудник»><URN атрибута значения «должность>=<URN значения «инженер>»><URN атрибута «начало работы»>=<URN значения даты>.
N_A_UNFT.2.9: Составная связь четвертого типа – «древовидный триплет», используется в случаях, когда у одного значения атрибута связи может быть несколько атрибутов со своими значениями, у каждого из которых могут быть свои атрибуты со своими значениями и т.д. Формат представления связи имеет вид: <URN субъекта><URN объекта> <URNc> = <URN элемента словаря значений соответствующего атрибута связи> [[блок 1 [блок 1.1 [блок 1.1.1.] [блок 1.1.2]] блок 2 [блок 2.1] и т.д.]], где

[4] Другими атрибутами могут быть «спонсор», «учредитель», акционер и т.п.

блок представляет собой структуру <URN атрибута значения связи i-того уровня>=<URN одного из значений этого атрибута>

N_A_UNFT.2.10: алгоритмы контроля 10-значного номера и 13-значного номеров ISBN

N_A_UNFT.2.11: алгоритм контроля номера ISSN

Третий и четвертый атрибуты справочника форматов принимают одно из двух значений:

N_A_UNFT.3.1: r

N_A_UNFT.3.2: f

N_A_UNFT.4.1: u

N_A_UNFT.4.2: m

Пример словаря значений атрибута «ограничения по кодировке».

N_A_UNFT.5.1: JPG

N_A_UNFT.5.2: MP4

N_A_UNFT.5.3: UniCode UTF-8

N_A_UNFT.5.4: арабские цифры

Словарь значений атрибута «ссылка на подробное описание формата»:

N_A_UNFT.6.1: https://habr.com/ru/post/454944/

N_A_UNFT.6.2: https://open-file.ru/types/mp4

N_A_UNFT.6.3: https://ru.wikipedia.org/wiki/Юникод

N_A_UNFT.6.4: https://ru.wikipedia.org/wiki/Коды_языков

Примеры конкретных элементов справочника форматов, используемые при описаниях структур других справочников:

Текст, только буквы, в кодировке UniCode UTF-8, атрибут обязательный, значение уникальное

UNFT.1: N_A_UNFT.1.1; ; N_A_UNFT.3.1; N_A_UNFT.4.1; N_A_UNFT.5.3; N_A_UNFT.6.3;

Целое число, атрибут обязательный, значение повторяющееся

UNFT.2: N_A_UNFT.1.7; ; N_A_UNFT.3.1; N_A_UNFT.4.2; ; ;

Текст, только буквы, атрибут необязательный, значение уникальное

UNFT.3: N_A_UNFT.1.2; ;N_A_UNFT.3.2; N_A_UNFT.4.1; ; ;

Формат описания связей типа <URN субъекта> <URN связи> <URN объекта>

UNFT.4: N_A_UNFT.1.10; N_A_UNFT.2.4; ; ; ; ;

Дата в формате гггг[.мм[.дд]], атрибут необязательный, значение уникальное

UNFT.11: N_A_UNFT.1.8; ;N_A_UNFT3.2; N_A_UNFT.4.1; ; ;

Любой текст, атрибут необязательный, значение уникальное
UNFT.14: N_A_UNFT.1.1; ; N_A_UNFT.3.2; N_A_UNFT.4.1;;;
Любой текст, атрибут обязательный, значение уникальное
UNFT.18: N_A_UNFT.1.1; ; N_A_UNFT.3.1; N_A_UNFT.4.1;;;

Описание класса «персоны».

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

Справочник персон будет иметь имя UNPS, а конкретные объекты, входящие в этот класс, будет иметь URN=UNPS.k.

Объект класса «персоны» в ЕЦПНЗ идентифицируется значениями атрибутов, перечисленных в справочнике A_UNPS, структура которого описана в справочнике CDSSK.5. По мере необходимости он может дополняться новыми элементами, что не нарушит существовавшую до этого структуру. Значения атрибутов содержатся в словарях, указанных в соответствующих элементах справочника. Пример элементов справочника атрибутов объектов класса «Персоны»:

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

A_UNPS.2: имя; UNFT.1; N_A_UNPS.2; C_N_A_UNPS.2; имя выбирается из словаря, при отсутствии оно вводится и проверяется на эквивалентность с другими написаниями;

A_UNPS.3: отчество; UNFT.3; N_A_UNPS.3; ; отчество выбирается из словаря, при отсутствии оно вводится;

A_UNPS.4: дата рождения; UNFT.k [URN объекта класса «Форматы», сообщающий, что элемент представляется в формате «дд.мм.гггг», является уникальным]; N_A_UNTC.2 [URN словаря значений соответствующего атрибута объектов класса «временные характеристики»]; ; ;

A_UNPS.5: место рождения; UNFT.3; UNGC [URN словаря объектов класса «местонахождение»]; ; ;

A_UNPS.6: дата смерти; UNFT.11.; N_A_UNTC.2; ; ;

A_UNPS.7: место смерти; UNFT.3; UNGC; ; ;

A_UNPS.8: квалификация (ученая степень); UNFT.3; N_A_UNPS.8

A_UNPS.9: ученое звание; UNFT.3; N_A_UNPS.9; ; ;

A_UNPS.10: биография; UNFT.18; N_A_UNPS.10; ; ;

A_UNPS.11: библиография персоны; UNFT.14; N_A_UNPS.11; ; ;

A_UNPS.12: библиография о персоне; UNFT.14; N_A_UNPS.12; ; ;

Элементы словарей N_A_UNPS.8 и N_A_UNPS.9 заполняются на административном уровне на основе существующих градаций ученых степеней и званий. Словарь местонахождений UNGC может быть также заполнен данными из имеющихся географических информационных

систем и дополняться по мере необходимости. Остальные словари заполняются данными, относящимися к конкретным персонам, по мере наполнения ЕЦПНЗ.

Дополнительные характеристики персон описываются как связи с другими классами объектов. В частности, идентификаторы авторов в российских и международных системах представляются как связи с объектами класса «политематические базы данных». Рассмотрим, в качестве примера, структуру связи персоны с публикацией.

Связь персоны с публикацией является простой связью второго типа, описываемой форматом N_A_UNFT.2.5. Она может принимать несколько значений (персона может быть автором и художником издания, одним из авторов и редакторов и т.п.). Обозначим эту связь как COSP.5.

Справочник этой связи будет иметь вид:

COSP.5: UNPS; UNPB; связь персоны с публикацией; N_A_UNFT.2.5; A_COSP.5; 0;

Справочник атрибутов представляется в виде:

A_COSP.5.1: Роль персоны в создании публикации; UNFT.i; N_A_COSP.5; ; ;

Второй элемент (UNFT.i) указывает, что значение атрибута содержит только буквы, является обязательным, и одной персоне может соответствовать несколько его значений.

Словарь возможных значений атрибута (дополняется по мере необходимости):

N_A_COSP.5.1: автор
N_A_COSP.5.2: редактор
N_A_COSP.5.3: составитель
N_A_COSP.5.4: автор перевода
N_A_COSP.5.5: художник
N_A_COSP.5.6: о нем
N_A_COSP.5.7: владелец авторских прав

Пример конкретного значения - персона с URN=UNPS.r является редактором и автором перевода публикации с URN=UNPB.s:

N_COSP.5.n: < UNPS.r >< UNPB.s ><COSP.5>=<N_A_COSP.5.2>
N_COSP.5.n+1: < UNPS.r >< UNPB.s ><COSP.5>=<N_A_COSP.5.4>.

Итоговое представление данных о конкретной персоне и ее связях с другими объектами универсального и тематических подпространств представляется в виде строки словаря, содержащей последовательность URN значений словарей атрибутов (N_A_UNPS.i.j), последовательность URN значений связей между персонами и другими объектами (N_COSP.n.m).

UNPS.i: N_A_UNPS.1.a; N_A_UNPS.2.b;…;N_A_UNPS.12.k; N_COSP.i.j;…;N_COSP.q.z

Аналогично представляются и другие объекты. Совокупность словарей объектов и словарей значений связей представляет собой замкнутую систему, внутри которой, используя мнемонику формирования справочников разного уровня, можно реализовать многоаспектный поиск данных и навигацию между разнородными элементами.

10. Заключение

В настоящее время в МСЦ РАН ведутся исследования по развитию и конкретизации предложенной модели. Работа выполняется в рамках государственного задания по теме FNEF-2022-0014 и при поддержке РФФИ (проект 20-07-00773).

Литература

1. 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*
2. Антопольский А.Б., Босов А.В., Савин Г.И., Сотников А.Н., Цветкова В.А., Каленов Н.Е., Серебряков В.А., Ефременко Д.В. Принципы построения и структура единого цифрового пространства научных знаний (ЕЦПНЗ) // Научно-техническая информация. Сер. 1, 2020. - № 4. - С. 9-17. *DOI: 10.36535/0548-0019-2020-04-2*.
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. Атаева О.М., Каленов Н.Е., Серебряков В.А. Онтологический подход к описанию единого цифрового пространства научных знаний // Электронные библиотеки, 2021. - Т. 24, - № 1. - С. 3-19. *DOI: 10.26907/1562-5419-2021-24-1-3-19*
5. Каленов Н.Е., Серебряков В.А. Об онтологии Единого цифрового пространства научных знаний // Информационные ресурсы России, 2020. - № 5. - С. 10-12. *DOI: 10.51218/0204-3653-2020-5-10-12* .
6. W3C 2009. SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009. Available at <https://www.w3.org/TR/skos-reference/>.
7. SKOS Simple Knowledge Organization System Reference https://webarchive.library.unt.edu/web/20170125143526/http://www.w3.org/TR/skos-reference/#xl-Label.
8. Marcia Lei Zeng & Philipp Mayr. Knowledge Organization Systems (KOS) in the Semantic Web: a multi-dimensional review // International Journal on

Digital ibraries. 2018. https://doi.org/10.1007/s00799-018-0241-2. Полный текст https://arxiv.org/pdf/1801.04479.pdf/
9. Pattuelli, M. Cristina, Alexandra Provo, and Hilary Thorsen 2015. Ontology building for Linked Open Data: A pragmatic perspective. Journal of Library Metadata, 15(3-4), 265-294.
10. Volkan, Çağdaş and Erik Stubkjær 2015. A SKOS vocabulary for Linked Land Administration: Cadastre and Land Administration Thesaurus. Land Use Policy, 49 (2015), 668-679.
11. Zapilko, Benjamin, Johann Schaible, Philipp Mayr, and Brigitte Mathiak. 2013. TheSoz: A SKOS representation of the Thesaurus for the Social Sciences. Semantic Web Journal (SWJ), 4(3), 257–63.
12. Zeng, Marcia Lei 2017. Create microthesauri and other datasets from the Getty LOD vocabularies. In MW17: Museums and the Web Conference, April 19-22, 2017 Cleveland, Ohio, USA <Available at http://www.getty.edu/research/tools/vocabularies/zeng_microthesauri_getty_lod.pdf>.
13. Ontolog-Forum https://groups.google.com/forum/#!forum/gettyvocablod.
14. Resource Description Framework (RDF): Concepts and Abstract Syntax. https://clck.ru/gwVBC

References

1. 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.
2. Antopol'skiy A.B., Bosov A.V., Savin G.I., Sotnikov A.N., Tsvetkova V.A., Kalenov N.Ye., Serebryakov V.A., Yefremenko D.V. Printsipy postroyeniya i struktura yedinogo tsifrovogo prostranstva nauchnykh znaniy (YETSPNZ) // Nauchno-tekhnicheskaya informatsiya. Ser. 1, 2020. - № 4. - S. 9-17. DOI: 10.36535/0548-0019-2020-04-2.
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. Atayeva O.M., Kalenov N.Ye., Serebryakov V.A. Ontologicheskiy podkhod k opisaniyu yedinogo tsifrovogo prostranstva nauchnykh znaniy // Elektronnyye biblioteki, 2021. - T. 24, - № 1. - S. 3-19. DOI: 10.26907/1562-5419-2021-24-1-3-19
5. Kalenov N.Ye., Serebryakov V.A. Ob ontologii Yedinogo tsifrovogo prostranstva nauchnykh znaniy // Informatsionnyye resursy Rossii, 2020. - № 5. - S. 10-12. DOI: 10.51218/0204-3653-2020-5-10-12 .
6. W3C 2009. SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009. Available at <https://www.w3.org/TR/skos-reference/>.

7. SKOS Simple Knowledge Organization System Reference https://webarchive.library.unt.edu/web/20170125143526/http://www.w3.org/TR/skos-reference/#xl-Label.
8. Marcia Lei Zeng & Philipp Mayr. Knowledge Organization Systems (KOS) in the Semantic Web: a multi-dimensional review // International Journal on Digital ibraries. 2018. https://doi.org/10.1007/s00799-018-0241-2. Полный текст https://arxiv.org/pdf/1801.04479.pdf/
9. Pattuelli, M. Cristina, Alexandra Provo, and Hilary Thorsen 2015. Ontology building for Linked Open Data: A pragmatic perspective. Journal of Library Metadata, 15(3-4), 265-294.
10. Volkan, Çağdaş and Erik Stubkjær 2015. A SKOS vocabulary for Linked Land Administration: Cadastre and Land Administration Thesaurus. Land Use Policy, 49 (2015), 668-679.
11. Zapilko, Benjamin, Johann Schaible, Philipp Mayr, and Brigitte Mathiak. 2013. TheSoz: A SKOS representation of the Thesaurus for the Social Sciences. Semantic Web Journal (SWJ), 4(3), 257–63.
12. Zeng, Marcia Lei 2017. Create microthesauri and other datasets from the Getty LOD vocabularies. In MW17: Museums and the Web Conference, April 19-22, 2017 Cleveland, Ohio, USA <Available at http://www.getty.edu/research/tools/vocabularies/zeng_microthesauri_getty_lod.pdf>.
13. Ontolog-Forum https://groups.google.com/forum/#!forum/gettyvocablod.
14. Resource Description Framework (RDF): Concepts and Abstract Syntax. https://clck.ru/gwVBC