УДК 013, 004.65

УНИФИЦИРОВАННОЕ ПРЕДСТАВЛЕНИЕ ОНТОЛОГИИ ЕДИНОГО ЦИФРОВОГО ПРОСТРАНСТВА НАУЧНЫХ ЗНАНИЙ

Н. Е. Каленов[1] [0000-0001-5269-0988], А. Н. Сотников[2] [0000-0002-0137-1255]

[1, 2]*Межведомственный суперкомпьютерный центр (МСЦ) РАН – филиал ФГУ ФНЦ Научно-исследовательский институт системных исследований (НИИСИ) РАН, Ленинский пр., 32а, г. Москва, 119334*

[1]nkalenov@jscc.ru, [2]asotnikov@jscc.ru

Аннотация

Единое цифровое пространство научных знаний (ЕЦПНЗ) представляет собой цифровую информационную среду, агрегирующую разнородную информацию, связанную с различными аспектами научных знаний. Одной из важных функций ЕЦПНЗ является предоставление информации для решения задач искусственного интеллекта, что обусловливает необходимость поддержки данных в структуре, соответствующей правилам Semantic Web. Особенностями ЕЦПНЗ являются, с одной стороны, политематичность и разнородность элементов контента, с другой – высокая динамика появления новых видов объектов и связей между ними, что обусловлено спецификой развития науки. При реализации ЕЦПНЗ должна быть обеспечена возможность навигации по разнородным ресурсам пространства с использованием семантических связей между ними. Возможности ЕЦПНЗ в значительной мере определяются структурой онтологии пространства, модель которой предложена в данной работе. В рамках модели проведена иерархическая структуризация онтологии ЕЦПНЗ; выделены и определены такие элементы, как «подпространство», «класс объектов», «объект», «атрибуты объекта», три типа попарных связей объектов и атрибутов (универсальные, квазиуниверсальные и специфические). Структура каждого типа элементов определяется «справочником» унифицированного вида; конкретные значения атрибутов и связей содержатся в словарях унифицированной структуры. Выделен класс объектов «Форматы», описывающих правила формирования атрибутов и значений связей. Пред-

ложена формализация представлений справочников и словарей ЕЦПНЗ. Предлагаемая модель позволяет достаточно просто добавлять в пространство, по мере необходимости, новые виды объектов, их попарных связей и атрибутов.

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

ВВЕДЕНИЕ

Единое цифровое пространство научных знаний (ЕЦПНЗ) представляет собой цифровую среду, агрегирующую разнородную информацию, связанную с различными аспектами научных знаний. ЕЦПНЗ должно обеспечить поддержку процессов предоставления широкому кругу пользователей необходимой им информации в различных областях науки. ЕЦПНЗ рассматривается как интегратор для научных целей государственных информационных систем (ИС) (таких как Большая Российская энциклопедия, Национальная электронная библиотека, Государственный каталог географических названий и пр.) с отраслевыми научными информационными ИС, электронными библиотеками (ЭБ), регистрами и т. п. В рамках ЕЦПНЗ необходимо объединить эти ресурсы на основе онтологического подхода и Semantic Web для решения широкого круга образовательных и научных задач, в том числе, ориентированных на применение методов искусственного интеллекта.

Отличительной особенностью ЕЦПНЗ являются политематичность и разнородность элементов контента с обеспечением возможности навигации по ресурсам пространства с использованием семантических связей между ними.

Программная 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 в Западной Сибири в 20 веке» должны быть выданы описания всех археологических объектов, найденных в Томской, Новосибирской областях, в Тобольске и т. д., за период с 1901 по 2000 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 Западной Сибири, вытекает из автоматического анализа связей между объектами пространства (в данном случае относящимися к географии и времени).

Цели создания, задачи и общие принципы построения ЕЦПНЗ приведены в [1–3].

Одним из первых шагов к практической реализации ЕЦПНЗ является разработка его онтологии – определение правил формирования его составляющих, включая наполнение контента разнородными, но связанными по единым правилам, данными. Общие подходы к формированию онтологии ЕЦПНЗ отражены в [4, 5].

Построению онтологий и правилам их отражения в Сети посвящено значительное количество исследований и публикаций. В рамках Simple Knowledge Organization System (SKOS) [6–8] разработаны формальные правила отражения в цифровой среде связанных открытых данных (LOD), тезаурусов, свойств объектов и их связей с использованием правил OWL и RDF. 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.), отражены в [9–13].

На сайте «онтологического форума» [14] 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 интернете, 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 составляющих. Вариант такой технологии, не противоречащей принципиальным подходам SKOS и OWL, но являющейся фактически их развитием в сторону упрощения модели, предложен ниже.

1. Общие понятия

ЕЦПНЗ рассматривается как иерархическая структура, включающая подпространства, классы объектов, объекты, атрибуты объектов, значения атрибутов объектов. Наряду с этой структурой имеется структура попарных связей объектов и попарных связей значений атрибутов. Каждая связь, в свою очередь, имеет свое значение и может иметь атрибуты и их значения.

Все перечисленные составляющие назовем элементами ЕЦПНЗ. Каждый элемент имеет своё уникальное имя (URN).

Подпространство – это совокупность элементов ЕЦПНЗ, относящихся к определенному научному направлению; выделяется универсальное подпространство, содержащее информацию об объектах мультидисциплинарного характера (персоны, события, единицы измерения и т. п.).

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

Объект – совокупность структурированной многоаспектной информации о физической сущности (например, о конкретном человеке, конкретной книге, музейном предмете и т. п.), научном понятии (например, об уравнении Матье, Законе всемирного тяготения, корпусе текстов китайского языка и т. п.), событии или научном мероприятии и др. Объект как понятие может рассматриваться как аналог энциклопедического «слота», который также может являться объектом ЕЦПНЗ. Каждый объект характеризуется своими значениями атрибутов и связей с другими объектами.

Атрибуты – это характеристики, присущие элементу вне контекста связей с другими объектами. Атрибут – аналог понятия «имя поля данных», используемого при проектировании баз данных. Перечень атрибутов, присущих тому или иному объекту или связи, определяется, исходя из роли объекта в решении задач ЕЦПНЗ.

Значение атрибута – конкретное значение данной характеристики, присущее данному объекту или связи. В качестве значения атрибута могут выступать текст, число, дата, формула, изображение и т. д.

Класс – это совокупность объектов, относящихся к данному подпространству, имеющих заданный набор атрибутов. В универсальном подпространстве выделим класс «Форматы», объекты которого описывают правила формирования значений атрибутов и связей всех объектов.

Связи – это вид «взаимоотношений» между парами объектов или значений атрибутов. Понятие связей в ЕЦПНЗ существенно шире аналогичных понятий, принятых в SKOS и OWL.

Связи ЕЦПНЗ подразделяются на три группы, каждая из которых относится к одному из типов – универсальному, квазиуниверсальному или специфическому.

Связи могут быть простыми и составными. Простые связи содержат (в терминах триплетов RDF [15]) 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.

Универсальные связи 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 на соподчиненность подразделений организаций, на различные наименования организаций и публикаций, на различные 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 или ссылки на предметные рубрики классификационных систем.

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, у квазиуниверсальных связей, набор которых растет достаточно медленно, перечень специфических связей является достаточно динамичным, поскольку определяется развитием ЕЦПНЗ и возникающими перед ним задачами.

2. Справочники и словари

Для обеспечения процессов формирования контента ЕЦПНЗ, обработки запросов и навигации по ресурсам пространства необходимо иметь информацию о структуре элементов пространства и «взаимоотношениях» между ними. Эта информация содержится в соответствующих справочниках, которые формируются и дополняются администратором ЕЦПНЗ.

Справочники – структурированная информация, содержащая перечень и форматы представления элементов ЕЦПНЗ, связей между ними и их значениями. Структура справочников элементов ЕЦПНЗ определенного вида фиксирована и определяется элементами справочника CDSSK.

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.

Словари значений атрибутов и связей содержат их конкретные значения. Каждое значение является уникальным, относится к одному из словарей и имеет свое имя (URN).

Словари объектов в качестве элементов содержат перечень URN значений атрибутов и связей, относящихся к конкретному объекту.

Словари «стандартных» значений атрибутов (таких как перечень ученых степеней, должностей, рубрики ГРНТИ или УДК и пр.) наполняются при первона-

чальной инсталляции ЕЦПНЗ, остальные словари наполняются в процессе формирования контента ЕЦПНЗ оператором ввода или программой пакетной загрузки данных.

3. 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 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, основанный на формировании справочника верхнего уровня с именем CDSSK. Элемент CDSSK.1 описывает структуру справочников подпространств, элемент CDSSK.2 – справочников классов и т. д.

CDSSK.1: Структура справочника подпространств (ПП).

Справочник подпространств имеет имя SUBS; элемент справочника содержит три атрибута: наименование; код типа подпространства; описание подпространства. Код типа подпространства (далее – «префикс») состоит из двух символов; 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»).

Имя справочника подпространств SUBS

Элемент справочника содержит 3 составляющих:

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: Структура справочника класса объектов.

Класс объектов (Class). Определены два типа классов объектов – универсальные и локальные. Последние принадлежат какому-либо тематическому подпространству. Имя справочника классов: URN: Class.

Элемент справочника содержит 6 составляющих:

Наименование

Тип (универсальный – UN, локальный – LC)

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

URN словаря атрибутов

URN словаря связей.

Описание

Примеры:

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: Структура справочника атрибутов.

Имя (URN) справочника атрибутов формируется в форме A_префикс класса.

Элемент справочника содержит 5 составляющих:

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 (пояснительный текст).

Пример (фрагмент справочника):

A_UNPS.1: фамилия; UNFT.10 [URN объекта из класса «Форматы данных», сообщающий, что атрибут является обязательным текстовым]; N_A_UNPS.1; C_N_A_UNPS.1 [в этом словаре связей содержатся указания на эквивалентность разных написаний фамилий]; фамилия выбирается из словаря, при отсутствии она вводится и проверяется на эквивалентность с другими написаниями;

A_UNPS.2: имя; UNFT.3 [URN объекта из класса «Форматы данных», сообщающий, что атрибут является необязательным текстовым]; 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.4 [URN объекта класса «Форматы данных», сообщающий, что элемент представляется в формате «гггг[мм[дд]]», является обязательным, уникальным]; N_A_UNTC.2 [ссылка на элемент словаря временных характеристик]; ; ;

CDSSK.4: Структура справочника универсальных связей.

Имя (URN) справочника: REUN.

Элемент справочника содержит 3 составляющих:

Наименование

URN значения словаря формата данных, определяющего форму представления данной связи

Описание связи

Пример:

REUN.1: Эквивалентность; N_A_UNFT.2.6; используется для обозначения идентичных атрибутов или связей (разные написания фамилий и имен, разные наименования одной организации, синонимы терминов и т. п.)

CDSSK.5: Структура справочника квазиуниверсальных связей.

Имя (URN) справочника: RQUN.

Элемент справочника содержит 6 составляющих:

Наименование

Префикс класса, являющегося «объектом связи»

Необходимость справочника значений (Y / N)

URN словаря значений (если указано Y) или пустое поле

URN значения словаря формата данных, определяющего форму представления данной связи

Описание связи.

Пример:

RQUN.4: Местоположение; UNPC; Y; N_A_UNPC.1; N_ A_UNFT.2.6: указывается местоположение объекта в виде, присутствующем в словаре географических наименований.

CDSSK.6: Структура справочника специфических связей.

Имя (URN) справочника: RESP.

Элемент справочник имеет «шапку» из 7-ми составляющих, которая в случае составной связи дополняется блоками, содержащими по 4 составляющих, описывающими иерархию значений связи.

Составляющие элементов справочника:

1.Наименование связи

2. Класс субъекта

3. Класс объекта

4. URN справочника атрибутов связи

5. URN словаря значений связи

6. Формат представления связи (URN значения элемента N_UNFT)

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

Если не нуль, то добавляется блок связи второго уровня:

8. Наименование подчиненной связи 1

9. URN словаря атрибутов подчиненной связи 1

10. URN словаря значений подчиненной связи 1

11. Количество подчиненных связей следующего уровня (0 – n)

Если не 0, то определяется блок подчиненных связей третьего уровня, если 0, а в строке 7 n>1, определяется следующий блок связи второго уровня.

Пример:

RESP.5: UNPS; UNPB; связь персоны с публикацией; N_A_UNFT.2.5; A_RESP.5; 0;

CDSSK.7: Структура словаря значений атрибутов объектов и связей.

URN словаря формируется в форме N_URN атрибута. Элемент словаря имеет одну составляющую – значение в соответствии с форматом, URN которого указан в справочнике атрибутов.

Примеры:

N_A_UNPS.1.1: Менделеев

N_A_UNPS.4.1: N_A_UNTC.2.1

N_A_UNTC.2.1: 1834.12.08

CDSSK.8: Структура словаря связей

Имя словаря совпадает с URN справочника связи, указанном в соответствующем справочнике CDSSK.

Примеры:

пусть

N_A_UNPS.1.1: Андреев

N_A_UNPS.1.2: Andreev

N_A_UNPS.1.3: Andreyev,

тогда

REUN.1.1: <N_A_UNPS.1.1>< N_A_UNPS.1.2>

REUN.1.2: <N_A_UNPS.1.1>< N_A_UNPS.1.3>

Если персона с URN=UNPS.r является редактором и автором перевода публикации с URN=UNPB.s, то этот факт будет представлен двумя элементами словаря значений N_RESP.5:

N_RESP.5.n: < UNPS.r >< UNPB.s >=<N_A_RESP.5.2>

N_RESP.5.n+1: < UNPS.r >< UNPB.s >=<N_A_RESP.5.4>,

где элементы словаря N_A_RESP.5 представлены в виде:

N_A_RESP.5.2: редактор

N_A_RESP.5.4: автор перевода.

CDSSK.9: Структура словарей объектов.

Имя словаря совпадает с URN класса, к которому относится данный объект.

Элемент словаря представляет собой перечень URN элементов словарей атрибутов и связей, относящихся к данному объекту.

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 словари), либо как результат диалога с оператором ввода. Во втором случае оператору предлагаются (на основе программной обработки справочников) наименования атрибутов вводимого объекта и связей с другими объектами. По каждым атрибуту и связи оператор должен выбрать уже имеющиеся в ЕЦПНЗ их значения или ввести новые с указанием значений всех необходимых связей.

4. Примеры формального описания объектов и связей

В настоящее время в универсальном подпространстве выделены классы объектов, которые условно разделены на две группы – предметные и вспомогательные. К предметным классам отнесены: «Персоны», «Публикации», «Квалификационные работы» «Документы», «Мультимедийные материалы», «Музейные предметы», «События», «Организации», «Политематические базы данных» «Награды». К вспомогательным: «Форматы данных», «Тезаурусы (предметные онтологии)», «Местоположение (географические характеристики)», «Временные характеристики», «Единицы измерения», «Научные направления», «Группы персон», «Числовые значения», «Языки», «Коллекции».

Для каждого класса объектов сформировано их формальное описание и предложен перечень атрибутов; для ряда предметных классов определены виды попарных специфических связей.

Рассмотрим несколько примеров.

4.1. 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 или структуре; ; 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 <URNc>:<URNO1><URNO2>, где URNc – URN конкретной связи. Примеры: фамилия «Петров» эквивалентна «Petrov»; статья входит в состав энциклопедии; организация включает подразделение и т. п.

N_A_UNFT.2.5: 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 (атрибут специфической связи «персона – «организация»1) 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

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

[блок 2.1] и т. д.]], где 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 ISBN. Если ISBN=N_1 N_2 … N_{10}, то $N_{10} = 11 - \left(S - 11*[S/11]\right)$, где $S = \sum_{i=1}^{9} i * N_i$, [S/11] – целая часть результата деления S на 11, если при вычислении N_{10} оказывается равным десяти, оно записывается римским числом X.

N_A_UNFT.2.11: алгоритмы контроля 13-значного номера ISBN. Если ISBN=N_1 N_2 ... N_{13}, то $N_{13} = 10 - \left(R - 10*[R/10]\right)$, где $R = \sum_{i=0}^{6} N_{2i+1} + 3\sum_{j=1}^{6} N_{2j}$, [R/10] – целая часть результата деления R на 10.

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 или структуре».

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.REm/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;

Любой текст, атрибут обязательный, значение уникальное

UNFT.2: N_A_UNFT.1.1; ; N_A_UNFT.3.1; N_A_UNFT.4.1;;;

Текст, только буквы, атрибут необязательный, значение множественное

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.5: N_A_UNFT.1.8; ;N_A_UNFT3.2; N_A_UNFT.4.1; ; ;

Любой текст, атрибут необязательный, значение уникальное

UNFT.6: N_A_UNFT.1.1; ; N_A_UNFT.3.2; N_A_UNFT.4.1;;;

4.2. 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.5; 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.5.; 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.2; N_A_UNPS.10; ; ;

A_UNPS.11: библиография персоны; UNFT.6; N_A_UNPS.11; ; ;

A_UNPS.12: библиография о персоне; UNFT.6; 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 RESP.5.

Справочник этой связи будет иметь вид:

RESP.5: UNPS; UNPB; связь персоны с публикацией; N_A_UNFT.2.5; A_RESP.5; 0;

Справочник атрибутов представляется в виде:

A_RESP.5.1: Роль персоны в создании публикации; UNFT.i; N_A_RESP.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_RESP.5.1: автор

N_A_RESP.5.2: редактор

N_A_RESP.5.3: составитель

N_A_RESP.5.4: автор перевода

N_A_RESP.5.5: художник

N_A_RESP.5.6: о нем

N_A_RESP.5.7: владелец авторских прав

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

N_RESP.5.n: < UNPS.r >< UNPB.s >=<N_A_RESP.5.2>

N_RESP.5.n+1: < UNPS.r >< UNPB.s >=<N_A_RESP.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_RESP.n.m).

UNPS.i: N_A_UNPS.1.a; N_A_UNPS.2.b;…;N_A_UNPS.12.k; N_RESP.i.j;…;N_RESP.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.

5. Заключение

Предложенная структура онтологии ЕЦПНЗ в настоящее время моделируется на примере развития электронной библиотеки «Научное наследие России» (ЭБ ННР) [16]. Библиотека поддерживает такие классы объектов, как «персоны»,

«публикации», «музейные объекты», «коллекции». Традиционный поисковый интерфейс позволял искать объекты определенного класса по заданным значениям их атрибутов с возможностью использования булевой логики. Реализованная в последней версии ЭБ опция «расширенный поиск» позволяет искать объекты не только по заданным значениям их атрибутов, но и по значению связей с объектами другого класса. Например, пользователь имеет возможность найти публикации, в которых персоны играли роль не только авторов, но и редакторов или переводчиков; найти музейные объекты, для которых персоны выступали в роли «автора сбора»; найти публикации, связанные с музейными объектами, и т. п.

В плане развития исследований предполагается расширить модельную базу путем постепенного добавления новых классов объектов и связей.

Работы выполняются в МСЦ РАН – филиале ФГУ ФНЦ НИИСИ РАН в рамках государственного задания по теме FNEF-2023-0014.

СПИСОК ЛИТЕРАТУРЫ

1. *Савин Г.И.* Единое цифровое пространство научных знаний: цели и задачи // Информационные ресурсы России. 2020. № 5. С. 3–5. https://doi.org/10.51218/0204-3653-2020-5-3-5

2. *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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.51218/0204-3653-2020-5-10-12

6. W3C 2009. SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009. URL: https://www.w3.org/TR/skos-reference/ (дата обращения: 10.01.2023).

7. SKOS Simple Knowledge Organization System. URL: http://www.w3.org/TR/skos-reference/#xl-Label (дата обращения: 10.01.2023).

8. Web Ontology Language (OWL). URL: https://www.w3.org/OWL/ (дата обращения: 10.01.2023).

9. *Marcia Lei Zeng & Philipp Mayr.* Knowledge Organization Systems (KOS) in the Semantic Web: a multi-dimensional review // International Journal on Digital Libraries. 2018. URL: https://arxiv.org/pdf/1801.04479.pdf/ (дата обращения: 10.01.2023).

10. *Pattuelli M. Cristina, Alexandra Provo, and Hilary Thorsen 2*015. Ontology building for Linked Open Data: A pragmatic perspective. Journal of Library Metadata. 2015. Vol. 15, No. 3-4. P. 265–294.

11. *Volkan Çağdaş and Erik Stubkjær*. A SKOS vocabulary for Linked Land Administration: Cadastre and Land Administration Thesaurus. Land Use Policy. 2015. Vol. 49. P. 668–679.

12. *Zapilko Benjamin, Johann Schaible, Philipp Mayr, and Brigitte Mathiak.* TheSoz: A SKOS representation of the Thesaurus for the Social Sciences. Semantic Web Journal (SWJ). 2013. Vol. 4, No. 3. P. 257–263.

13. *Zeng Marcia Lei*. Create micro thesauri and other datasets from the Getty LOD vocabularies. In MW17: Museums and the Web Conference, April 19–22, 2017 Cleveland, Ohio, USA. URL: http://www.getty.edu/research/tools/vocabularies/zeng_microthesauri_getty_lod.pdf (дата обращения: 10.01.2023)

14. Ontolog-Forum. URL: https://groups.google.REm/forum/#!forum/ gettyvocablod (дата обращения: 10.01.2023).

15. Resource Description Framework (RDF): Concepts and Abstract Syntax. URL: https://clck.ru/gwVBC (дата обращения: 10.01.2023).

16. Электронная библиотека «Научное наследие России». URL: http://heritage1.jscc.ru/ (дата обращения: 10.01.2023).

UNIFIED REPRESENTATION OF THE COMMON DIGITAL SPACE OF SCIENTIFIC KNOWLEDGE ONTOLOGY

N. Kalenov[1] [0000-0001-5269-0988], A. Sotnikov[2] [0000-0002-0137-1255]

[1, 2]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"

[1]nkalenov@jscc.ru, [2]asotnikov@jscc.ru

Abstract

The Common Digital Space of Scientific Knowledge (CDSSK) is a digital information environment aggregating heterogeneous information related to various aspects of scientific knowledge. One of the important functions of the CDSSK is to provide information for solving artificial intelligence problems, which makes it necessary to support data in a structure that complies with the rules of the semantic WEB. The features of the CDSSK are, on the one hand, the polythematics and heterogeneity of content elements, on the other hand, the high dynamics of the emergence of new types of objects and connections between them, which is due to the specifics of the development of science. At the same time, it should be possible to navigate through heterogeneous space resources using semantic links between them. The possibilities of the CDSSK are largely determined by the structure of the ontology of space, the model of which is proposed in this paper. Within the framework of the model, the hierarchical structuring of the CDSSK ontology is carried out; such elements as "subspace", "class of objects", "object", "attributes of an object", three types of pairwise relations of objects or attributes (universal, quasi-universal and specific) are distinguished and defined. The structure of each elements type is determined by a "reference book" of a unified type; specific values of attributes and relationships are contained in dictionaries of a unified structure. A class of "Formats" objects describing the rules for the formation of attributes and values of relationships is allocated. The formalization of CDSSK reference books and dictionaries representations is proposed. The proposed model allows you to simply add new types of objects, of their pairwise relationships and attributes to the space, as needed.

Keywords*: digital space of scientific knowledge, ontologies, structuring, related data, data attributes, semantic WEB.*

REFERENCES

1. Savin G.I. Yedinoye tsifrovoye prostranstvo nauchnykh znaniy: tseli i zadachi // Informatsionnyye resursy Rossii. 2020. № 5. S. 3–5. https://doi.org/0.51218/0204-3653-2020-5-3-5

2. *Antopol'skiy A.B. i dr.* Printsipy postroyeniya i struktura Edinogo tsifrovogo prostranstva nauchnykh znaniy // Nauchno-tekhnicheskaya informatsiya. ser. 1. 2020. № 4. S. 9–17. https://doi.org/10.36535/0548-0019-2020-04-2

3. *Kalenov N.Ye., Sotnikov A.N.* Arkhitektura shirokogo rasprostraneniya nauchnykh znaniy // Informatsionnyye resursy Rossii. 2020. № 5. S. 5–8. https://doi.org/10.51218/0204-3653-2020-5-5-8

4. *Atayeva O.M., Kalenov N.Ye., Serebryakov V.A.* Ontologicheskiy podkhod k opisaniyu obshchedostupnykh nauchnykh prostranstv // Elektronnyye biblioteki. 2021. T. 24, № 1. S. 3–19. https://doi.org/10.26907/1562-5419-2021-24-1-3-19

5. *Kalenov N.Ye., Serebryakov V.A.* Ob ontologii Yedinogo otkrytogo prostranstva nauchnykh znaniy // Informatsionnyye resursy Rossii. 2020. № 5. S. 10–12. https://doi.org/10.51218/0204-3653-2020-5-10-12.eller

6. W3C 2009. SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009. https://www.w3.org/TR/skos-reference/ (accessed 10.01.2023).

7. SKOS Simple Knowledge Organization System. URL: http://www.w3.org/TR/skos-reference/#xl-Label (accessed: 10.01.2023).

8. Web Ontology Language (OWL). URL: https://www.w3.org/OWL/ (accessed: 10.01.2023).

9. *Marcia Lei Zeng & Philipp Mayr.* Knowledge Organization Systems (KOS) in the Semantic Web: a multi-dimensional review // International Journal on Digital Libraries. 2018. URL: https://arxiv.org/pdf/1801.04479.pdf/ (accessed: 10.01.2023).

10. *Pattuelli M. Cristina, Alexandra Provo, and Hilary Thorsen 2*015. Ontology building for Linked Open Data: A pragmatic perspective. Journal of Library Metadata. 2015. Vol. 15, No. 3-4. P. 265–294.

11. *Volkan Çağdaş and Erik Stubkjær*. A SKOS vocabulary for Linked Land Administration: Cadastre and Land Administration Thesaurus. Land Use Policy. 2015. Vol. 49. P. 668–679.

12. *Zapilko Benjamin, Johann Schaible, Philipp Mayr, and Brigitte Mathiak.* TheSoz: A SKOS representation of the Thesaurus for the Social Sciences. Semantic Web Journal (SWJ). 2013. Vol. 4, No. 3. P. 257–263.

13. *Zeng Marcia Lei*. Create micro thesauri and other datasets from the Getty LOD vocabularies. In MW17: Museums and the Web Conference, April 19–22, 2017 Cleveland, Ohio, USA.
URL: http://www.getty.edu/research/tools/vocabularies/zeng_microthesauri_getty_lod.pdf (accessed: 10.01.2023)

14. Ontolog-Forum. URL: https://groups.google.REm/forum/#!forum/gettyvocablod (accessed: 10.01.2023).

15. Resource Description Framework (RDF): Concepts and Abstract Syntax. URL: https://clck.ru/gwVBC (accessed: 10.01.2023).

16. Elektronnaya biblioteka “Nauchnoe nasledie Rossii”.
URL: http://heritage1.jscc.ru/ (accessed: 10.01.2023).

СВЕДЕНИЯ ОБ АВТОРАХ

КАЛЕНОВ Николай Евгеньевич – главный научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», доктор технических наук, профессор.

Nikolay Evgenievich KALENOV – Chief Researcher of the Joint SuperComputer Center of the Russian Academy of Sciences – Branch of the Federal State Institution «Scientific Research Institute for System Analysis of the Russian Academy of Sciences», Doctor of Technical Sciences, Professor.

email: nekalenov@mail.ru;

ORCID: 0000-0001-5269-0988

СОТНИКОВ Александр Николаевич – заместитель директора Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», доктор физико-математических наук, профессор.

Alexander Nikolaevch SOTNIKOV – Deputy Director of the Joint SuperComputer Center of the Russian Academy of Sciences – Branch of the Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”, Doctor of Sciences (Math), Professor.

email: asotnikov@jscc.ru;

ORCID: 0000-0002-0137-1255

Материал поступил в редакцию 10 января 2023 года