

БАЗИ ДАНИХ, БАЗИ ЗНАТЬ ТА ІНЖЕНЕРІЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

УДК 004.91

Проскудина Г.Ю., Резниченко В.А.

Інститут програмних систем Національної академії наук України

КОНЦЕПТУАЛЬНА МОДЕЛЬ НАУЧНИХ ІНФОРМАЦІОННИХ РЕСУРСОВ НАН УКРАЇНИ

Дается краткая характеристика концептуальной информационной модели основных видов научных информационных ресурсов НАН Украины. Работа выполняется в рамках Программы информатизации НАН Украины.

Дается стисла характеристика концептуальної інформаційної моделі основних видів наукових інформаційних ресурсів НАН України. Робота виконується в рамках Програми інформатизації НАН України.

A brief description of the conceptual information model of the main types of scientific information resources of the NAS of Ukraine is proposed. The work is performed under the Program of Informatization of NAS of Ukraine.

Ключевые слова: концептуальное моделирование, информационный электронный ресурс, модель данных, научное информационное пространство.

Введение

В мире имеется достаточно большое количество информационных систем для работы с научными данными, наукоемкой информацией. Практически каждое научное учреждение представляет в электронной форме данные о публикациях сотрудников, о проводившихся или ведущихся научных исследованиях и проектах, о результатах исследований. Многие из учреждений имеют собственные информационные системы для наукоемкой информации, которые в каком-то виде ее хранят и предоставляют.

Однако при создании их описаний недостаточное внимание уделяется вопросам интероперабельности – слабо применяются соглашения по стандартизации электронного представления информационных ресурсов и соответствующие средства, призванные поддержать интеграцию информационных ресурсов, повышение полноты и точности поиска и т.п. Интересы пользователей не могут быть полностью обеспечены какой-то одной информационной системой, пусть даже весьма представительной. Как правило, интересы ученых выходят за рамки текущей исследовательской задачи и часто находятся на стыке научных областей. Практически невозможно в рамках одной организации собрать информацию, которая удовлетворила бы запросы всех ее сотрудников. Даже, если бы в какой-то момент это удалось, то в силу огромного динамизма научных исследований

невозможно было бы обеспечить приемлемую полноту и актуальность данных, представляемых такими замкнутыми системами.

В последнее время в мире предпринимаются усилия по полному и всестороннему описанию сферы единого информационного научного пространства, семантическому структурированию его контента и поддержке в явном виде воплощающих его связей между объектами. Что открывает новые возможности для научного творчества и существенно повышает информативность таких систем.

В данной статье предлагается подход по созданию единой информационной модели всех основных видов информационных ресурсов НАН Украины, который базируется на результатах современных исследований в этой области в мире.

Обзор исследований

В данном разделе приводится краткая информация о тех основных с точки зрения авторов статьи исследованиях и их результатах, которые имеют отношение к информационному моделированию научных информационных ресурсов и которые в той или иной мере нашли свое отражение в результатах проводимых авторами исследований.

ЕНИИП. В связи с темой данного отчета примечательна российская инициатива по организации Единого Научного Информационного Пространства

(ЕНИП), которая была призвана помочь научным коллективам сделать ряд шагов в направлении интеграции разнородных научных информационных и программных ресурсов отдельных научных учреждений, предоставлении пользователям более эффективных средств интеграции и поиска информации, научной коммуникации, сотрудничества и совместной работы. Под единым пространством здесь подразумевается не формирование централизованной системы, **не навязывание всем одних и тех же решений, а**

- **формулировка взаимосогласованного набора соглашений, правил и открытых стандартов;**
- подготовка совокупности макетов и типовых решений для реализации прикладных систем, инфраструктурных служб, поддерживающих разные уровни интероперабельности распределенных гетерогенных данных и приложений;
- создание целого ряда информационных систем общего назначения, следующих этим соглашениям, использующих эти реализации, допускающих модульную организацию, наращивание функциональных возможностей;
- применение этих результатов для решения соответствующих задач научных учреждений [1].

В результате исследований и разработок по этой теме были разработаны схемы метаданных для описания всех основных научных информационных ресурсов, разработан протокол обмена этими метаданными, разработана система автоматизации ведения информационных ресурсов в институтах РАН, которая поддерживает упомянутые выше схемы метаданных и протокол обмена ими.

CERIF. Появление формальной концептуальной модели научных данных CERIF (Common European Research Information Format), в дословном переводе – Обеевропейский исследовательский информационный формат [2-4] стало важным результатом в технологии семантического структурирования контента академических научных сред, и научных электронных библиотек в частности. Модель CERIF предназначена для поддержки управления научно-исследовательской информацией, а также настройки и взаимодействия между исследовательскими информационными системами. Модель имеет следующие основные характеристики: (а) она построена с помощью сущностей или объектов с атрибутами, таких как публикация, проект, персона, организация, событие и т.д.; (б) она поддерживает связи многие-ко-многим и рекурсивные связи; с помощью сущностей-связей, которые также обладают

определенными свойствами, (с) она поддерживает многоязычность, (г) она является расширяемой, что обеспечивает гарантированную совместимость, по крайней мере, на уровне ядра, но и не исключает более широкую совместимость. Модель предназначена для использования при обмене данными в гетерогенных распределенных средах. Использование так называемого семантического слоя делает модель гибкой и масштабируемой для применения в разнородных средах.

CIDOC CRM. Онтология в области документации в сфере культурного наследия CIDOC CRM ("International Committee for Documentation" "Conceptual Reference Model") [6] представляет собой формальную онтологию, предназначенную для улучшения интеграции и обмена гетерогенной информацией по культурному наследию. Более конкретно, CIDOC CRM определяет семантику схем баз данных и структур документов, используемых в культурном наследии и музейной документации, в терминах формальной онтологии. Модель не определяет терминологию, появляющуюся в конкретных структурах данных, но предоставляет характерные отношения для ее использования.

Модель может служить как руководством для разработчиков информационных систем, так и общим языком для экспертов предметной области и специалистов по информационным технологиям. Она предназначена для охвата контекстной информации исторического, географического и теоретического характера об отдельных экспонатах и музейных коллекциях в целом.

Структурно CIDOC CRM состоит из иерархии классов и широкого набора свойств (бинарных отношений), связывающих классы между собой. Все классы и свойства модели можно разделить на три группы. Первая группа включает классы и отношения, охватывающие наиболее общие понятия окружающего мира: постоянные и временные сущности, отношения участия, зависимости, совпадения во времени. Вторая группа содержит понятия, частично поддерживающие функции управления: приобретение и учет единиц хранения, передача прав собственности на объекты культуры. К третьей группе относятся классы и свойства, используемые для внутренней организации самой онтологии: средства, необходимые для подключения внешних источников терминов, например, тезаурусов по отраслям культуры.

Модель CIDOC CRM – нужный и полезный стандарт, важным преимуществом которого является его формальный подход. Обобщая все многообразие музейных коллекций и задач, модель содержит широкий

набор универсальных понятий. Важную роль здесь играют временные сущности, так как они связывают объекты с временным диапазоном, местом и субъектами. Стандарт может служить основой для информационной составляющей концептуальной модели электронной библиотеки.

FRBR. Международной федерацией библиотечных ассоциаций и учреждений (International Federation of Library Associations and Institutions, IFLA) разработана модель "Функциональные требования к библиографическим записям" (Functional Requirements for Bibliographic Records, FRBR) как обобщенное представление библиографического универсума [7]. В настоящее время IFLA продолжает контролировать приложения модели FRBR и поддерживает ее использование и развитие.

FRBR включает описание концептуальной модели (сущности, их отношения и атрибуты), предлагает универсальные библиографические записи для всех типов материалов и пользовательских задач, связанных с библиографическими ресурсами, описанными в каталогах, библиографиях и других библиографических инструментах. Была также предложена модель **FRBRoo**-объектно-ориентированная FRBR [8].

SPAR. Уже много лет ведутся исследования в области анализа семантики связей между научными материалами. Системным обобщением этих результатов стало появление комплекса онтологий SPAR (Semantic Publishing and Referencing) [5], обеспечивающего достаточно детальную категоризацию отношений, которые могут возникать между научными материалами в электронном виде, и воплощающих их связей. SPAR включает:

онтологии семантической публикации:

– **FaBiO** (FRBR-aligned Bibliographic Ontology). Библиографическая онтология совместима с FRBR, предназначенная для записи и публикации в Semantic Web библиографических научных записей.

– **PRO** (Publication Roles Ontology). Онтология, характеризующая роли (например, автор, редактор, корректор, издатель, библиотекарь) агентов (людей, организаций и вычислительных средств) в процессе подготовки, публикации и использования информационных ресурсов. Позволяет записывать временную информацию (дата/время) исполнения роли.

– **PSO** (Publications Status Ontology). Онтология, характеризующая состояние документов и других библиографических сущностей на различных этапах процесса публикации (например, представление, отклонение, прием, редактирование, издание).

– **PWO** (Publications Workflow Ontology). Онтология, характеризующая основные этапы рабочего процесса, связанного с публикацией документа (например, находится на рассмотрении, редактирование, подготовка в нужном формате, верстка, создание дизайна страницы, публикация в Интернете).

и онтологии ссылок:

– **CiTO** (Citation Typing Ontology). Онтология для описания семантики ссылок, используемых в научно-исследовательских статьях и других академических работах, как на другие публикации, так и на информационные ресурсы веба, с последующей возможностью публикации этих описаний в семантическом вебе (например, приводится в качестве варианта решения, приводится в качестве источника данных, приводится в качестве рекомендованной литературы, содержит подтверждающие сведения).

– **BiRO** (Bibliographic Reference Ontology). Онтология для описания библиографических записей и библиографических ссылок для их последующего объединения в виде библиографических коллекций, например, библиотечных каталогов, или библиографических списков, например, в виде списка использованной литературы, авторов, таблиц, рисунков и т.д. в выпусках журналов и сборников, трудах конференций, книгах и другой научной литературы.

– **C4O** (Citation Counting and Context Characterization Ontology). Онтология, позволяющая характеризовать библиографические ссылки с точки зрения их количественных характеристик и контекста.

– **DoCO** (Document Components Ontology). Онтология, характеризующая составные части библиографического документа. Предоставляет структурированную лексику компонентов документа (например, введение, обсуждение, выражение признательности, список литературы, список рисунков, приложения).

Все эти онтологии имеют полностью специфицированные версии и доступны для использования.

Основные характеристики разрабатываемой модели

Разрабатываемая модель основана на европейской формальной концептуальной модели научных данных CERIF. Перечислим ее наиболее важные черты с некоторыми дополнениями, сделанными с учетом ее применения в информационной модели НАН Украины.

Базовые сущности. Основу или ядро модели составляют несколько базовых или главных сущностей (объектов) в данной предметной области – единого научного информационного пространства. Это – Организация, Персона, Проект, Публикация, Экспонат и Событие. Для каждой из этих базовых сущностей определяются наборы собственных атрибутов или свойств. Смысл этих базовых сущностей следующий:

• **Организация** – любое юридическое лицо, имеющее непосредственное отношение к НАНУ.

• **Персона** – любое физическое лицо, имеющее непосредственное отношение к НАНУ.

• **Проект** – любые официально зарегистрированные и выполняющиеся в НАНУ работы.

• **Публикация** – результат научно-производственной деятельности, представленный в бумажном или электронном виде.

• **Экспонат** – результат научно-производственной деятельности, представленный в виде образца, макета или другого физического представления.

• **Событие** – любое мероприятие, представляющее интерес с точки зрения информационного моделирования НАНУ.

Связи. Все базовые сущности связаны друг с другом и сами с собой связями типа многие-ко-многим (см. рис. 1). Рассматриваются только бинарные связи, то есть в связи всегда участвует пара сущностей. Например, связь Организации с Персоной может указывать принадлежность лиц организациям. Между парами базовых сущностей может быть множество связей. Например, между Организацией с Персоной могут быть установлены такие связи, как «работает в», «курирует», «контролирует» и др.

Итак, экземпляры базовых сущностей могут быть связаны между собой как внутри одной сущности (рекурсивные связи, на рисунке показаны петли), так и со всеми остальными сущностями отношением типа многие-ко-многим. Кроме того, связи сами по себе могут обладать своими свойствами. В связи с этим вводятся, так называемые, **переходные сущности** или **сущности-связи**, которые преобразуют бинарную связь на сущность, связывающую исходные сущности бинарной связи. На рис. 2. Приведены переходные сущности для всех бинарных связей базовой сущности Организация. Так, например, переходная сущность Организация-Персона может обладать такими свойствами, как дата поступления на работу, занимаемая должность, оклад, занимаемое помещение и т.д. Обратим внимание, что перечисленные свойства не являются свойствами ни

организации, ни человека, а именно пары организация-человек.

Базовые сущности представляют научные субъекты (персоны и организации) и их основную деятельность (проекты, публикации, экспонаты, события), а связи устанавливают различного рода ассоциации между ними: сотрудничество одних ученых с другими (Персона-Персона), их участие в проектах (Персона-Проект), публикация статей (Персона-Публикация), проведение мероприятий типа семинаров и конференций (Организация-Событие). Проекты включают людей (Проект-Персона) и организации (Проект-Организация). Научные публикации публикуются организацией (Организация-Публикация) и относятся к проектам (Проект-Публикация) и т.д.

Свойства. Любая сущность (базовая или переходная) может обладать свойствами. Свойства присущи только сущностям. Свойства должны быть однозначными, то есть конкретный экземпляр той или иной сущности может принимать только одно значение. Тем не менее, во многих случаях те или иные свойства являются по сути многозначными, например наличие в статьях многих названий или аннотаций (например, на различных языках) или ключевых слов. В связи с этим допускается, что такие многозначные свойства представляются в виде самостоятельных **сущностей-свойств**.

Классификация. При разработке любой более-менее крупной информационной системы всегда возникает вопрос о классификаторах, которые позволяют использовать множество классификационных схем для описания тех или иных сущностей. Например, при описании конкретных индивидов из класса Персона могут потребоваться классификаторы должностей, званий, степеней, квалификаций, наград и т.д. Обычно в разрабатываемую информационную модель включаются конкретные классификаторы и осуществляется жесткая привязка классификаторов к тем или иным классам сущностей. Предлагаемая нами модель не требует предварительного определения множества необходимых классификаторов и установления жесткой взаимосвязи между сущностями и классификаторами. Исходя из предположения, что между сущностями и классификаторами в общем случае существует связь многие-ко-многим и что эта связь может обладать своими свойствами, вводятся дополнительные переходные сущности Сущность-Классификатор, которые связывают любую сущность модели (базовую, переходную, сущность-свойство и сущность второго уровня) с классификаторами. Таким образом, обеспечивается определенная гибкость в

выборе и использовании классификационных схемы, нет жесткой привязки сущностей к классификаторам. На рис. 3 приведена классификационная схема для базовых сущностей. На этой схеме то, что представлено понятием Классификация по сути является множеством переходных сущностей-классификаторов для любой пары Сущность-Классификатор.

Итак, для классификации любой сущности может использоваться множество классификаторов. Так, например, по классификации, предложенной в проекте CERIF, сущность Публикация может быть следующих типов: Книга, Монография, Диссертация, Учебное издание, Журнальная статья, Труды конференции и т.д. [3]. По классификации, предложенной ЕНИИ [1], все публикации на первоначальном абстрактном уровне классифицируются на издания сводного, монографического и аналитического уровня, а затем конкретизируются как, например, серийное издание, многотомное издание, труды института, периодическое издание и т.д. Предлагаемая модель допускает включение обоих классификаторов и любых других классификаторов для сущности Публикация, как и для любых других сущностей.

Сущности второго уровня.

Определяется набор так называемых сущностей второго уровня – Патент, Продукт, Оборудование, Финансирование, Метрики и т.д. По аналогии с базовыми сущностями, для них также определяются наборы собственных свойств, многозначных сущностей-свойств и переходных сущностей. Любая из базовых сущностей может иметь связь с какой-либо из сущностей второго уровня, для них также создаются свои переходные сущности (например, Публикация-Метрика, Публикация-Финансирование).

Примерами сущностей второго уровня являются следующие:

Патент – охранный документ, удостоверяющий исключительное право, авторство и приоритет изобретения, полезной модели либо промышленного образца.

Продукт – результаты научной деятельности: интернет-системы, базы данных, экспериментальные данные, математические модели, программные системы, технологии и т.д.;

Финансирование – выделение денежных средств или финансовых ресурсов на выполнение научных работ;

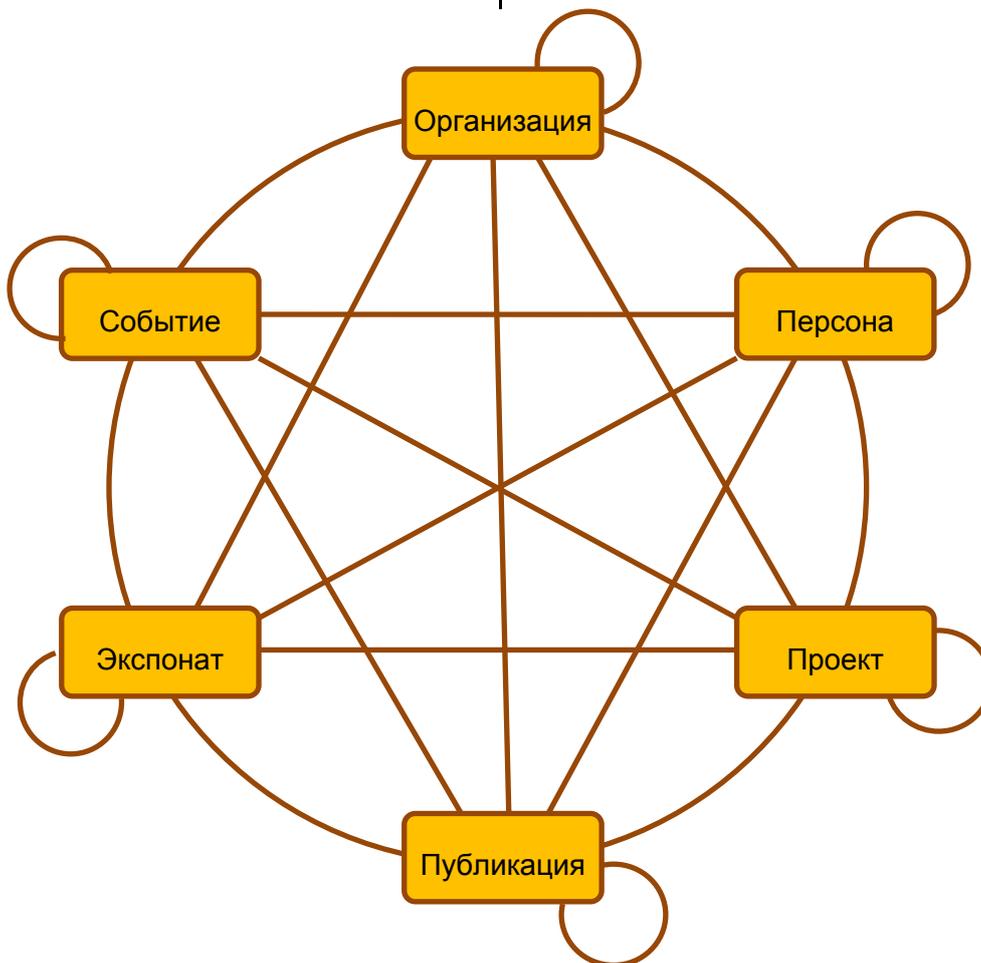


Рис. 1. Базовые сущности и их взаимосвязь



Рис. 2. Переходные сущности для базовой сущности Организация



Рис. 3. Классификационная схема для базовых сущностей

Цитирование – включение в научную работу фрагмента текста другой работы с указанием точных выходных данных позволяющих найти работу и место, из которого был взят фрагмент текста;

Метрики – мера, позволяющая получить численное значение некоторого свойства какого-либо объекта.

Оснащение – совокупность технических средств или приспособлений, которыми оснащено что-либо.

Оборудование, в частности, лабораторное оборудование – различные инструменты и снаряжение, используемое учёными, работающими в лаборатории для выполнения экспериментов или осуществления измерений. Лабораторное оборудование подразделяется на общелабораторное, измерительное, специализированное, испытательное и аналитическое.

Сервис – совокупность средств для обслуживания пользователей, процесс оказания услуг.

Награда – что-либо, выдаваемое в качестве поощрения за заслуги, отличия, в знак благодарности за что-либо. К наградам относят: почётные знаки, призы, премии,

почётные грамоты, ценные подарки и почётные звания.

Опыт и Навыки – квалификация, степень профессиональной подготовленности к выполнению определенного вида работы.

Множество сущностей второго уровня не фиксируется, оно может быть специфическим для каждой из базовых сущностей.

Сущности второго уровня могут быть классифицированы (или типизированы) в соответствии с определенными классификациями, как и базовые сущности. Так, например, Финансирование можно классифицировать как [3]: Программа финансирования, Запрос на финансирование, Тендер, Спонсорская поддержка.

Сущности-связи между базовой и сущностью второго уровня формируются также как и для пары базовых. Ее двойное название всегда указывает на то, между чем и чем создана такая связь, например сущность-связь Персона-Финансирование, подразумевает, что создана связь между базовой сущностью Персона и сущностью второго уровня Финансирование.

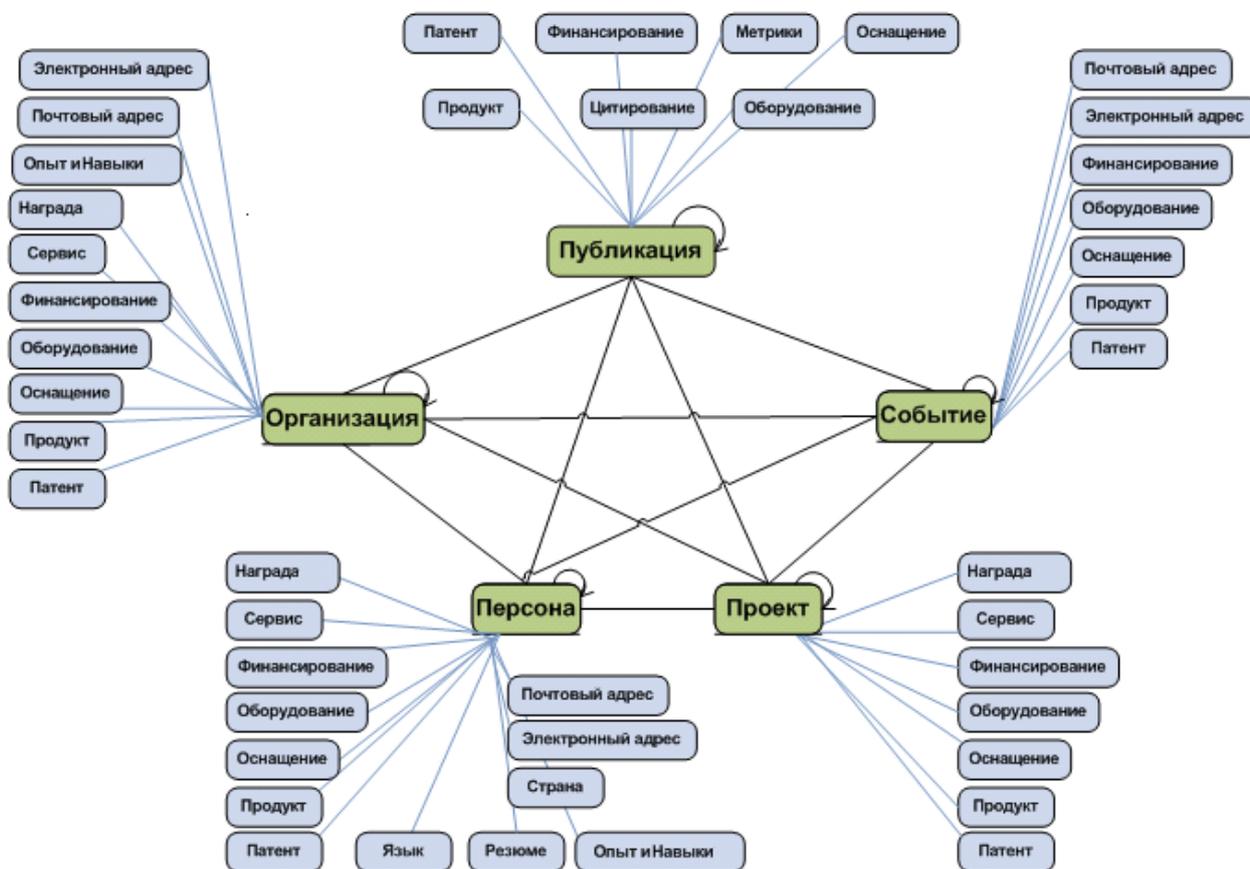


Рис. 4. Связь базовых сущностей с сущностями второго уровня

Пример описания базовой сущности Проект

В рамках данной модели были разработаны концептуальные информационные схемы для всех базовых сущностей. В качестве примера рассмотрим сущность Проект.

Проект – это научная деятельность, имеющая начало и конец во времени, направленная на достижение заранее определённого результата/цели, создание определённого, уникального продукта или услуги, при заданных ограничениях по ресурсам и срокам, а также требованиям к качеству и допустимому уровню риска.

На рис. 5 представлена концептуальная схема сущности Проект.

Зеленым цветом показана сама сущность Проект, со своими основными атрибутами: ИД проекта, дата начала и дата окончания, акроним, URI, а также многими другими, которые описывают именно эту сущность и являются однозначными.

Желтым цветом выделены еще три ее описательных атрибута, ставшие в представляемой модели отдельными сущностями, поскольку они многозначные. Атрибуты, требующие многозначного представления: Заголовок, Резюме и Ключевые слова.

Для того чтобы сущность Проект могла быть классифицирована по-разному с помощью различных классификационных схем, в модель введена связующая сущность Проект-Классификация (она изображена розовым цветом). Одной из таких классификационных схем могут использоваться, например, типы проектов, определенных в Седьмой Рамочной программе Европейского Сообщества (FP7), а именно):



Рис. 5. Концептуальная информационная модель сущности Проект

– Совместный проект – поддержка научно-исследовательских проектов с участниками из разных стран, направленных на разработку новых знаний, новых технологий, новых продуктов.

– Сети передового опыта – поддержка совместной программы мероприятий, осуществляемых в ряде научно-исследовательских организаций по интеграции их деятельности в данной области, проведенных исследовательскими группами в рамках долгосрочного сотрудничества. Координация и поддержка действий – поддержка деятельности, направленной на координацию и поддержку научно-исследовательской деятельности и политик (сети, биржи, транснациональный доступ к научно-исследовательской инфраструктуре, исследованиям, конференциям и т.д.).

– Индивидуальные проекты – поддержка проектов, осуществляемых отдельными национальными или транснациональными исследовательскими группами, поддержка новых областей исследований.

– Поддержка обучения и карьерного роста исследователей. Помощь исследователям в выработке профессиональных навыков и повышении мобильности и компетентности в ходе роста карьеры.

– Исследования в интересах конкретных групп. Поддержка научно-исследовательских проектов, где основная часть исследований осуществляется университетами, исследовательскими центрами и другими юридическими лицами, в интересах конкретных групп.

Данная классификация, как и любые другие, естественным образом включаются в модель без какой-либо ее перестройки.

Серым цветом показана группа связующих сущностей с базовыми: Проект-Проект, Персона-Проект, Организация-Проект, Публикация-Проект, Событие-Проект, Экспонат-Событие. Типы каждой из этих связующих сущностей задаются атрибутами класса и классификационной схемы. Например, атрибут класс может иметь значение: Основатель проекта, Координатор проекта, проекта. Причем, такая классификация типов может быть применима, как для связующей сущности Организация-Проект, так и для связующей сущности Персона-Проект. Для связующей сущности Проект-Проект согласно [3] определена такая семантика: Часть, Основан на, Приемник, Совместный и т.д..

Сведения об авторах:



Проскудина Галина Юрьевна – научный сотрудник Института программных систем НАН Украины. Научные интересы: электронные библиотеки, концептуальное информационное моделирование.

e-mail: proskydina@mail.ru



Резниченко Валерий Анатольевич – к.ф.-м.н, с.н.с., ведущий научный сотрудник Института программных систем НАН Украины. Научные интересы: информационные системы и базы данных, электронные библиотеки, семантический веб.

e-mail: vreznichenko_47@mail.ru

И, наконец, группа связующих сущностей второго уровня показана на рис. 5 голубым цветом. К ним относятся сущности связи, имеющие очевидный смысл.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Бездушный А.А. Интеграция метаданных Единого Научного Информационного Пространства РАН / А.А. Бездушный, А.Н.Бездушный, В.А.Серебряков, В.И.Филиппов // Вычислительный центр РАН. – М.: 2006. – 238с. – Библ. 61 наим. – рус.
2. CERIF 2008 - 1.2 Full Data Model (FDM). Introduction and Specification [Электронный ресурс] – Режим доступа: http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/CERIF2008_1.2_FDM.pdf
3. CERIF 2008 - 1.2 Semantics, euroCRIS. http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/Release_1.2/CERIF2008_1.2_Semantics.pdf
4. CERIF 2008 - 1.2 XML Data Exchange Format Specification [Электронный ресурс] – Режим доступа: http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/Release_1.2/CERIF2008_1.2_XML.pdf
5. David Shotton. Introduction the Semantic Publishing and Referencing (SPAR) Ontologies. October 14, 2010. [Электронный ресурс] – Режим доступа: <http://opencitations.wordpress.com/2010/10/14/introducing-the-semantic-publishing-and-referencing-spar-ontologies/>
6. Crofts N., Doerr M., Gill T., Stead S., Stiff M. (editors), Definition of the CIDOC Conceptual Reference Model, January 2008. Version 4.2.4.
7. Functional Requirements for Bibliographic Records, Final Report / IFLA Study Group on the Functional Requirements for Bibliographic Records. – München: K.G. Saur, 1998. (UBCIM Publications, New Series; v. 19). [Электронный ресурс] – Режим доступа: <http://archive.ifla.org/VII/s13/frbr/frbr.htm>
8. Doerr M., Leboeuf P. Modelling intellectual processes: The FRBR-CRM harmonization // In Conf. Proc. of ICOM-CIDOC Annual Meeting. Gothenburg, Sweden. – 2006. – P. 10 – 14.
9. SWAN (Semantic Web Applications in Neuromedicine) - Scientific Discourse Relationships Ontology Specification. [Электронный ресурс] – Режим доступа: <http://swan.mindinformatics.org/spec/1.2/discourserelationships.html>
10. David Shotton. CiTO, the Citation Typing Ontology. J. of Biomedical Semantics 2010, 1(Suppl 1): S6. [Электронный ресурс] – Режим доступа: <http://www.jbiomedsem.com/content/1/S1/S6>
11. David Shotton, Silvio Peroni. DoCO, the Document Components Ontology. 17/02/2011. [Электронный ресурс] – Режим доступа: <http://purl.org/spar/doco/>