Оригинальная статья / Original Article

DOI: 10.31857/S160578800018917-4

Метаданные лингвистических ресурсов: история и современное состояние

© 2022 г. А. Б. Антопольский

Доктор технических наук,
главный научный сотрудник Института научной информации
по общественным наукам РАН (ИНИОН РАН),
Россия, 117997, Москва, Нахимовский проспект, д. 51/21
ale5695@yandex.ru

Резюме. Описываются основные проекты метаданных для лингвистических (языковых) ресурсов, реализованные за последние 20 лет. В их числе инициатива IMDI, система метаданных OLAC, метамодель META-SHARE, Международный стандартный номер языковых ресурсов, оценочная карта языковых ресурсов, а также компонентная модель метаданных CLARIN. Излагается содержание стандарта ИСО на метаданные. Описываются проекты создания словарей, онтологий и лексических баз для метаданных языковых ресурсов.

Ключевые слова: метаданные, лингвистические ресурсы, языковые ресурсы, стандарты, словари, онтологии.

Для цитирования: *Антопольский А.Б.* Метаданные лингвистических ресурсов: история и современное состояние // Известия Российской академии наук. Серия литературы и языка. 2022. Т. 81. № 1. С. 21–36. DOI: 10.31857/S160578800018917-4

Metadata of Linguistic Resources: History and Current State

© 2022 Alexander B. Antopolsky

Doct. Sci. (Tech.),
Head Researcher at the Institute of Scientific Information
for Social Sciences of the RAS (INION RAN),
51-21 Nakhimovskiy Prospect, Moscow, 117997, Russia
ale5695@yandex.ru

Abstract. The main metadata projects for linguistic (language) resources developed over the past 20 years are described. These include the IMDI initiative, the OLAC metadata system. the META-SHARE meta-model, the International Standard Number of Language resources, the evaluation map of language resources, and the CLARIN component metadata model. The content of the ISO metadata standard is described. Projects for creating dictionaries, ontologies, and lexical databases for metadata of language resources are described.

Key words: metadata, linguistic resources, language resources, standards, dictionaries, ontologies.

For citation: Antopolsky, A.B. *Metadannye lingvisticheskih resursov: istoriya i sovremennoe sostoyanie* [Metadata of Linguistic Resources: History and Current State]. *Izvestiâ Rossijskoj akademii nauk. Seriâ literatury i âzyka* [Bulletin of the Russian Academy of Sciences: Studies in Literature and Language]. 2022, Vol. 81, No. 1, pp. 21–36. (In Russ.) DOI: 10.31857/S160578800018917-4

Введение

Лингвистические (языковые) ресурсы (ЛР)[1] – это наборы данных, представляющие примеры использования языка либо непосредственно, как в корпусах, либо в виде производных данных, как в лексиконах и онтологиях. ЛР используются в лингвистике и смежных областях, таких как язык жестов, антропология, компьютерная лингвистика, искусственный интеллект, фонетика, психология, распознавание речи, мультимодальные исследования и человеко-машинный интерфейс, дизайн. Лингвисты используют их для создания и проверки новых лингвистических гипотез; инженеры по распознаванию речи используют их для тестирования устройств распознавания речи и установки параметров распознавания.

Развитие Всемирной паутины с ее связанными веб-страницами открыло новые возможности для распространения и повторного использования ЛР, но и поставило новые задачи. Возникла потребность создать пространство связанных ЛР с информацией о них. Это пространство должно быть доступно через Интернет с соответствующими инструментами для просмотра и поиска. Иначе говоря, возникла необходимость в разработке системы метаданных для ЛР.

Но ЛР значительно различаются, поэтому возникает вопрос, как такое разнообразие приложений может быть представлено одной системой метаданных. Дискуссии по этому поводу ведутся достаточно давно, первый проект системы метаданных для ЛР появился в 2001 г. Позже появилось еще несколько проектов, как опирающихся на стандартные системы метаданных, прежде всего на Дублинское ядро метаданных, так и развивающих специальные системы метаданных для ЛР. Одна из систем метаданных получила статус стандарта ISO.

В настоящей статье предлагается обзор основных систем метаданных, применяемых в настоящее время для описания ЛР.

Все системы метаданных в той или иной степени опираются на словари (регистры, онтологии) лингвистических категорий, которые также активно разрабатывались международным сообществом лингвистов в последние годы. Эти проекты описаны в отдельном разделе статьи.

В список литературы включены как традиционные публикации, так и ссылки на информационные ресурсы и документацию, размещенные в Интернете.

Проект метаданных IMDI

Постановка задачи разработки специальной системы метаданных для ЛР принадлежит, по-видимому, рабочей группе EAGLES/ISLE[2], которая в 2001 г. предложила план разработки соответствующего стандарта [1]. Эта инициатива получила название Инициативы IMDI[3].

Рабочая группа IMDI разработала подробные предложения, в которых были учтены требования сообщества разработчиков и пользователей ЛР, существующий опыт разработки систем метаданных, в том числе Дублинского ядра (DC), RDF и других. Была определена сфера применения ЛР, среди которых разработчики выделили различные типы ЛР: текстовые корпуса, аннотированные корпуса, мультимедийные корпуса, лексиконы, типологические базы данных, грамматические данные, онтологии и другие.

На этой основе были определены структура метаописания, объем метаданных, элементы словаря метаданных, отображение элементов метаданных, в том числе повторное использование определений элементов метаданных из других сообществ.

Были сформулированы требования к инструментам. Нужны редакторы метаописаний, браузеры, которые понимают структуру связанных файлов метаописания и предоставляют графические изображения поддержки пользователя во время навигации, инструменты поиска, которые могут справиться со структурой файла метаописания и любыми элементами метаданных. Инструменты поиска должны эффективно использовать связи между метаописаниями.

Практически осуществимый сценарий внедрения стандарта должен включать такие темы:

- где хранить метаописания

[1] В данной статье, в отличие от англоязычных аналогов, термины лингвистические и языковые ресурсы не различаются, и для них используется общая аббревиатура ЛР.

[2] EAGLES – Консультативная группа экспертов по стандартам языковых технологий – Expert Advisory Group on Language Engineering Standards http://www.ilc.cnr.it/EAGLES/home.html

ISLE – Международные стандарты для языковых технологий – International Standard for Language Engineering https://www.mpi.nl/ISLE/

[3] IMDI – Инициатива метаданных ISLE – ISLE Metadata Initiative http://tla.mpi.nl/imdi-metadata/

- способы регистрации и привязки метаописаний
- способы построения просматриваемых иерархий
- способы контроля за связыванием новых описаний с существующим пространством
- требования к центрам, которые могли бы создать и поддерживать пространство метаданных ЛР

В результате деятельности рабочей группы EAGLES/ISLE появились проекты систем метаданных для лексиконов [2] и предложения по классификации и структуре словарей [3]. Наибольшее распространение метаданные IMDI получили применительно к мультимодальным ЛР. Также была разработана схема перехода от модели метаданных IMDI к стандарту метаданных OLAC [4]. Руководство пользователя для модели метаданных IMDI представлено по адресу [5]. Полный перечень документов, разработанных в рамках инициативы IMDI, доступен по адресу [6].

Метаданные OLAC[4]

Метаданные крупнейшего современного собрания ЛР – Консорциума открытых лингвистических архивов (OLAC) – определены в нормативном документе “Метаданные OLAC” [7].

Этот документ определяет формат метаданных, используемый OLAC для описания ЛР и предоставления связанных с ними услуг. OLAC использует формат XML для обмена метаданными ЛР в рамках Инициативы открытых архивов (OAI).

Набор метаданных OLAC основан на наборе метаданных Дублинского ядра (DC) и использует все пятнадцать элементов, определенных в этом стандарте. Чтобы обеспечить большую точность в описании ЛР, OLAC следует рекомендациям DC для квалификации элементов.

Цитируемый документ определяет только формальные (синтаксические) требования к описанию метаданных OLAC. Полный набор рекомендаций, уточнения значения элементов и схемы использования содержатся в Рекомендациях по использованию метаданных OLAC [8].

Квалификаторы, рекомендованные DC, применимы к широкому спектру ЛР. Однако для ЛР, которые не удовлетворяют этим общим стандартам, члены OLAC разработали специальные квалификаторы для сообщества, которые приняты в качестве рекомендуемой передовой практики для описания ЛР.

Формат метаданных. XML-реализация метаданных OLAC соответствует “Руководящим принципам реализации Дублинского ядра в XML” [9]. Схема метаданных OLAC включает в себя элементы из двух схем метаданных (простой и квалифицированной). Квалифицированный элемент может указывать уточнение (используя элемент, определенный в пространстве имен dcterms) или схему кодирования (используя схему, определенную в dcterms как значение атрибута xsi:type), или и то, и другое.

Документ “Метаданные OLAC” [7] содержит определение пространства имен, словарь рекомендуемых языковых идентификаторов, порядок использования расширений, включая внешние расширения, устанавливает порядок документирования расширений.

Одним из наиболее сложных вопросов метаописаний ЛР, является детализация ЛР. Кратко опишем подход OLAC.

При определении правильного уровня для единиц, описываемых как ЛР, нужно учитывать множество факторов. Уровень единицы измерения, подходящий для включения в агрегированный каталог, такой как OLAC, может отличаться (как правило, быть выше) от уровня, желательного для каталога конкретного учреждения, который, в свою очередь, обычно выше уровня, желательного для описания подробного содержания ресурса.

Хранилище метаданных должно рассматривать ресурсы с одним происхождением как составляющие единую единицу и поэтому должны быть описаны в рамках одной записи.

Для ресурса, опубликованного в той или иной форме, подходящей единицей описания для записи OLAC является единица самой публикации. Коллективная работа может требовать отдельных записей для отдельных документов, содержащихся в ней, которые должны быть связаны с записью для работы в целом через отношения isPartOf и hasPart.

В общем случае запись OLAC соответствует цитируемому источнику, и для опубликованных работ детализация не представляет особой проблемы. Трудности возникают для первичных исходных материалов (например, записей, транскрипций, аннотаций, заметок, наборов данных). Типичной практикой архивистов является

[4] OLAC – Консорциум открытых лингвистических архивов – the Open Language Archives Community http://olac.ldc.upenn.edu/

объединение таких материалов в коллекции, которые в свою очередь становятся первичными единицами архивного описания (т.е. результатом являются ресурсы, которые по правилам DC могут быть отнесены к типу коллекции).

Однако на уровне OLAC главным фактором при определении принадлежности материалов к единой коллекции является их происхождение.

Обычно собранные ресурсы имеют общее происхождение. Это может быть один исследователь или исследовательская группа. Это также может быть проект, который объединяет материалы из разрозненных источников для единой новой исследовательской цели, таким образом создается новая коллекция, основанная на вторичном использовании материалов. Общая история также имеет значение; тот факт, что набор ресурсов был перемещен или передан в другие руки или обработан в целом с момента его первоначального сбора, помогает установить его идентичность как единой единицы для архивного описания.

ЛР общего происхождения, которые составляют коллекцию, будут отличаться высокой степенью общности элементов метаданных, например, один и тот же исследователь, автор, предметный язык, приблизительные даты, охват, лингвистический тип.

Другие элементы метаданных, которые также могут быть важны для обнаружения ресурсов, но которые могут отличаться для элементов коллекции (например, формат или тип дискурса), могут быть повторены на уровне описания OLAC. В качестве альтернативы коллекция может быть разделена на субколлекции по жанру дискурса, говорящему или другому значимому признаку.

Коллекция, как правило, описывается более подробно и выделяются особенности, представляющие интерес. Дальнейшие характеристики отдельных элементов в коллекции (например, тема, дополнительные участники, специфика события, формат) документируются на более тонком уровне детализации при описании коллекции.

Метамодель META-SHARE

Платформа META-SHARE является сервисом Европейской ассоциации языковых ресурсов (ELRA)[5], предназначенным для обмена ЛР. В этой платформе реализованы разнообразные возможности для описания ЛР. Многие авторы считают модель META-SHARE наиболее качественной моделью метаданных ЛР. В связи с этим эту модель опишем более подробно, используя основные положения из *Руководства по применению системы метаданных META-SHARE* [10].

В этом документе представлена обновленная версия схемы метаданных 2.0, которая реализована в виде XML-схемы. Схема метаданных META-SHARE охватывает следующие типы ресурсов / носителей:

- корпуса (текстовые, аудио, видео, мультимодальные / мультимедийные корпуса, включая сенсомоторные ресурсы, n-граммные ресурсы)
- лексические / концептуальные ресурсы (например, компьютерные словари, лексика, онтологии, машиночитаемые словари, терминологические ресурсы, тезаурусы, мультимодальная / мультимедийная лексика, словари и т.д.)
- языковые описания (например, компьютерные грамматики)
- технологии (инструменты / сервисы), которые могут быть использованы для обработки информационных ресурсов.

Более подробное изложение теоретических принципов и общее введение в модель можно найти в работах [11]; [12].

Основы модели

В модели META-SHARE термин *метаданные* относится к описаниям ЛР, охватывающим как данные (текстовые, мультимодальные / мультимедийные и лексические данные, грамматики, языковые модели и т.д.), так и технологии (инструменты / услуги), используемые для их обработки.

Механизм, который был принят, – это компонентный механизм, в соответствии с которым семантически когерентные элементы группируются вместе, образуя компоненты. Элементы используются для кодирования конкретных описательных признаков ЛР. Чтобы обеспечить семантическую согласованность с другими связанными схемами и моделями, метамодель включает ссылки на концептуально одинаковые или аналогичные существующие элементы Дублинского ядра и Реестра категорий данных ИСО (ISO DCR)[6]; при необходимости новые элементы будут включаться в DCR ИСО.

[5] ELRA – Европейская ассоциация языковых ресурсов European Language Resources Association http://www.elra.info/en/

[6] Реестр категорий данных ISO DCR будет описан ниже.

Было введено понятие отношений для кодирования связей между ресурсами. Отношения могут быть между различными формами ЛР (например, первичные данные и аннотированные ЛР), различными ЛР (например, язык, ресурс и инструмент, который был использован для его создания, и т.д.) независимо от того, включены они в репозиторий META-SHARE или нет, а также между ЛР и дополнительными документами. Отношения представлены как элементы в текущей версии схемы.

Совокупность всех компонентов и элементов, описывающих конкретные типы и подтипы ЛР, представляет профиль этого типа. Очевидно, что некоторые компоненты включают информацию, общую для всех типов ресурсов (например, идентификация, контакты, лицензионная информация и т.д.), в то время как другие (например, компоненты, включающие информацию о содержании, аннотации и т.д.) различаются по типам.

Элементы относятся к двум основным уровням описания:

- начальный уровень, обеспечивающий базовые элементы для описания ресурса (минимальная схема)
- второй уровень с более высокой степенью детализации (максимальная схема), охватывающей все этапы производства и использования ЛР.

Онтологии META-SHARE

META-SHARE стремится предоставить пользователям не только каталог ЛР (данных и инструментов), но и информацию, которая может быть использована для улучшения их использования. Например, исследовательские работы, документирующие производство ресурса, а также используемые стандарты и методики.

В онтологии META-SHARE проводится различие между ЛР как таковыми и другими материалами, связанными с ресурсом (отчеты, инструкции и т.д.), лица / организации, участвующие в их создании и использовании (создатели, дистрибьюторы и т.д.), проекты, мероприятия и лицензии (для доступа к ЛР).

Основной интерес для META-SHARE представляют собственно ЛР. Остальные материалы – акторы, проекты, документы и т.д. – описываются, когда они связаны с конкретным ЛР. Например, библиография включает только документы, связанные с ЛР.

Таксономия ЛР

Основным элементом, используемым для классификации ЛР по типам, которые приводят к когерентным наборам описаний, является ResourceType со следующими значениями:

- корпус (включая письменные / текстовые, устные / речевые, мультимодальные / мультимедийные корпуса)
- лексический / концептуальный ресурс (включая терминологические ресурсы, списки слов, семантические словари, онтологии и т.д.)
- языковое описание (включая грамматики)
- инструмент / сервис (включая базовые средства обработки, приложения, веб-сервисы и т.д.).

Важное место в описании ЛР в контексте META-SHARE также занимает элемент MediaType, который определяет форму / физический носитель ресурса. Понятие медиа позволяет рассматривать ЛР как набор модулей, каждый из которых может быть описан через отличительный набор признаков. Предусмотрены следующие значения медиа:

- текст
- аудио
- изображение
- видео
- textNumerical
- textNgram

Ресурс может состоять из частей, принадлежащих к различным типам медиа: например, мультимодальный корпус включает в себя видеочасть (движущееся изображение), аудиочасть (диалоги) и текстовую часть (субтитры и/или транскрипцию диалогов); мультимедийный лексикон включает в себя текстовую часть, но может также включать в себя видео- и/или аудиочасть; ресурс языка жестов также является ресурсом с различными типами носителей (видео, изображение, текст).

Точно так же программные инструменты могут быть применены к ресурсам различных типов медиа: например, инструмент может использоваться как для видео-, так и для аудиофайлов. Таким образом, для каждой части ресурса создается соответствующий набор функций (компонентов и элементов), например, для устного корпуса и его транскрипций набор звуковых функций будет использоваться для звуковой части, а набор текстовых функций – для транскрибируемой части.

Основное содержание и структура модели

Ядром модели является компонент ResourceInfo, который содержит всю информацию, имеющую

отношение к описанию ЛР. Он включает в себя компоненты и элементы, которые объединяются вместе, чтобы обеспечить это описание. Различаются “административные” компоненты, общие для всех ЛР, и “содержательные”, характерные для конкретного типа ЛР.

Все компоненты типа ЛР расположены под resourceComponentTypecomponent. Аналогично, для каждого типа ЛР создаются компоненты, чтобы сгруппировать вместе наборы функций, относящихся к каждому типу ЛР. Элементы ResourceType и MediaType кодируют две оси классификации схемы, в то время как каждое из значений этих двух элементов связано с соответствующим компонентом. Набор компонентов ResourceType и MediaType включает в себя:

- corpusInfo, lexicalConceptualResourceInfo, languageDescriptionInfo, toolServiceInfo включают информацию, специфическую для каждого типа ЛР, и принимают значения corpus, lexical/conceptualResource, languageDescription and toolServicefor соответственно
- corpusTextInfo, corpusAudioInfo, corpusVideoInfo, lexicalConceptualResourceTextInfo, lexicalConceptualResourceVideoInfo предоставляют информацию в зависимости от типа носителя каждого типа ЛР и включают элемент MediaType со значениями text, audio, video и т.д. соответственно.

Набор из шести компонентов обладает “особым” статусом в том смысле, что они могут быть присоединены к различным компонентам, выполняющим различные роли, а именно PersonInfo, organizationInfo, communicationInfo, projectInfo, sizeInfo и DocumentInfo. Например, sizeInfo может использоваться либо для определения размера всего ресурса, либо в сочетании с другим компонентом для описания размера частей ресурса (например, для домена, языка и т.д.); PersonInfo используется для контактных лиц, создателей ресурсов, лицензиатов, аннотаторов корпуса и т.д.

Существенным является статус элемента; в метамодели принято 4 статуса:

- Обязательный
- Условно-зависимый (обязательный в определенных условиях)
- Рекомендованный
- Необязательный

Наконец, для других сущностей модели помимо ЛР были разработаны специальные элементы, позволяющие осуществлять их массовое кодирование независимо от ресурса, с которым они связаны. Например, используя такой элемент, поставщик ресурсов может загрузить соответствующие метаданные для всех персон за один раз, а затем, редактируя метаданные для ресурсов, создать соответствующие ссылки на сохраненных персон.

Структура представления и условные обозначения

В модели приводятся сначала “специальные” компоненты, за которыми следуют компоненты, общие для всех типов ЛР, а затем компоненты типа ресурсов в следующем порядке: корпусы, инструменты / сервисы, языковые описания и, наконец, лексические / концептуальные ресурсы.

Для каждого компонента предоставляется следующая информация:

- *определение*: краткое утверждение, объясняющее семантику компонента в META-SHARE;
- *тип*: обычно он принимает значение “component”; значение “special status component” используется для специальных элементов;
- *элементы*: набор элементов / компонентов, включенных в компонент, с гиперссылкой к объяснению для каждого элемента, далее информация предоставлена для ее статуса и повторяемости;
- *компонент*: используется вместо “элементов” для компонентов специального статуса.

Для элементов сопроводительная информация включает в себя:

- определение: краткое утверждение, объясняющее его семантику в контексте META-SHARE
- тип элемента
- пространство значений: там, где это возможно, используется ссылка на контролируемую лексику или на стандартизированные, управляемые словари
- значения: если используется контролируемый словарь, специфичный для META-SHARE, то набор значений перечисляется вместе с определениями, где это необходимо
- примеры: небольшой список возможных значений для целей иллюстрации, особенно в случае текстовых элементов

- DCLINK: имя соответствующего элемента схемы Дублинского ядра, предоставленное для целей сопоставления
- ISOcatLINK: имя соответствующего элемента DCR ISOcat
- комментарии: используется для заметок там, где это необходимо.

Международный стандартный номер языковых ресурсов (ISLRN) [13]

В рамках ELRA разработано еще несколько моделей метаописаний ЛР. Одна применяется в Международном стандартном номере языкового ресурса (ISLRN). Это универсальная схема идентификации ЛР, которая обеспечивает уникальный идентификатор с использованием стандартизированной номенклатуры.

Каждый объект в мире требует своего рода идентификации, чтобы быть правильно распознанным. Традиционные печатные материалы, такие как книги, например, обычно используют Международный стандартный номер книги (ISBN), контрольный номер Библиотеки Конгресса (LCCN). Многие повседневные продукты применяют Международный/Европейский артикул (EAN), который является универсальной системой штрих-кодирования. Для цифровых ресурсов применяется Цифровой идентификатор объекта (DOI). Существуют и другие идентификаторы в качестве уникальной схемы идентификации.

Идентификация – это важный шаг в сетевом мире, в котором стали применяться технологии человеческого языка: уникальные ЛР должны быть идентифицированы, а метакаталоги нуждаются в общем формате идентификации для правильного управления данными. Поэтому ЛР должны иметь идентичные схемы идентификации независимо от их представлений, типов и их физического местоположения (локального или в Интернете).

ISLRN не должен заменять местные и конкретные идентификаторы, он является не обязанностью, а передовой практикой. Например, ресурс, распределенный между несколькими центрами обработки данных, по-прежнему будет иметь “локальный” идентификатор центра обработки данных, но будет иметь уникальный ISLRN.

Поскольку основная цель схемы метаданных, используемой в ISLRN, – это идентификация ЛР, был выбран минимальный набор метаданных на основе системы метаданных OLAC.

Метаданные ISLRN

- Заглавие
- Полное официальное имя
- Имя, по которому ресурс упоминается в библиографии
- Тип ресурса
- Характер или жанр содержания ресурса Источник / URL
- Формат / Тип MIME[7]
- Формат файла (тип MIME) ресурса. Примеры: текст / xml, видео / mpeg
- Размер / Продолжительность
- Среда доступа
- Материальный или физический носитель ресурса
- Описание
- Версия
- Тип СМИ
- Тип характера или жанра содержимого ресурса
- Язык(и)
- Создатель ресурсов
- Распределитель
- Лицо или организация, ответственные за предоставление ресурса
- Правообладатель
- Правообладатель ЛР

Карта LRE [14]

Еще одна модель метаданных, которая разработана в рамках ELRA, это оценочная карта ЛР (LRE), разработанная для мониторинга создания ЛР в разнообразных проектах. Эта карта была впервые распространена на конференции LREC в 2010 г., имела большой успех и позже распространялась на многих других конференциях. В настоящее время при помощи LRE описано свыше 6 тыс. ЛР. Приведем содержание этой карты. В скобках указано количество ЛР в массиве LRE, о которых имеются соответствующие данные.

Оценка ЛР

- Оценочные данные (230)
- Инструменты оценки (71)

[7] MIME – Multipurpose Internet Mail Extensions – многоцелевые расширения интернет-почты.

- Оценочный пакет (25)
- Методология оценки / стандарты / руководящие принципы (15)

Ресурс-данные

- Корпус (2920)
- Лексикон (666)
- Онтология (162)
- Грамматика / Языковая модель (82)
- Терминология (66)
- Банки деревьев зависимостей (42)

Ресурс-руководство

- Представление – Аннотации Формализм / Руководство (62)
- Языковые ресурсы / Технологии Инфраструктура (20)
- Метаданные (10)

Ресурс-инструмент

- Таггер / Парсер (400)
- Инструмент аннотации (245)
- Корпус Инструмент (83)
- Распознаватель именованных объектов (60)
- Инструмент машинного перевода (51)
- Программный инструментарий (41)
- Токенайзер (35)
- Инструмент машинного обучения (32)
- Инструмент моделирования языков (29)
- Определитель многозначности слов (17)
- Распознаватель речи / Транскриптор (14)
- Обработка сигналов / Извлечение функций (14)
- Веб-сервис (9)
- Синтезатор преобразования текста в речь (9)
- Идентификатор языка (6)
- Распознаватель говорящего (4)
- Инструмент для анализа настроений (4)
- Просодический анализатор (3)
- Анализатор изображений (3)
- Инструмент разговорного общения (1)

Состояние производства ЛР

- Существующие и используемые (2587)
- Недавно созданные, в процессе (1408)
- Недавно созданные, законченные (1290)
- Существующие обновленные (487)
- Другое (354)
- Не применимо (17)

Доступность

- Свободно доступно (2772)
- Другое (1232)
- От владельца (1229)
- Из дата-центра (580)
- Нет в наличии (267)
- Не применимо (63)

Модальность (семиотический тип)

- Письменные (4355)
- Речь (430)
- Мультимодальные / мультимедиа (286)
- Не применимо (261)
- Речь / Письменность (186)
- Язык жестов (64)
- Независимо от модальности и другое (561)

Использование ресурсов

- Извлечение информации, поиск информации (608)
- Машинный перевод, перевод речи в речь (532)
- Разбор и тегирование (289)
- Языковое моделирование (282)
- Классификация документов, текстовая категоризация (201)
- Распознавание / понимание речи (185)
- Приобретение (181)
- Дискурс (178)
- Открытие / Представление Знаний (172)
- Смысл словосочетания (160)
- Распознавание / Генерирование эмоций (154)
- Оценка / Валидация (152)
- Создание / Аннотация Корпуса (148)
- Текстовое копирование (147)
- Распознавание именованного субъекта (144)
- Диалог (122)
- Подведение итогов (96)
- Ответы на вопросы (83)
- Морфологический анализ (80)

- Семантическая паутина (68)
- Веб-сервисы (67)
- Создание / аннотация лексиконов (63)
- Синтез речи (55)
- Генерирование речи на естественном языке (54)
- Машинное обучение (54)
- Текстовое вложение и перефразирование (53)
- Анализ мнений / анализ настроений (40)
- Распознавание / генерирование речи на жестовом языке (35)
- Идентификация языка (35)
- Идентификация личности (30)
- Анафора, Кореферентность (30)
- Семантическая ролевая маркировка (29)
- Обнаружение и отслеживание тем (25)
- Обработка мультимедийных документов (22)
- Голосовое управление (3)
- Другое (1566)

Тип языка

- Одноязычный (2507)
- Независимо от языка (2206)
- Многоязычный (961)
- Двуязычный (380)
- Трехъязычный (84)
- Не применимо (5)

Язык (Top-4)

- Английский (961)
- Немецкий (216)
- Французский (180)
- Испанский (130)

Инфраструктура компонентов метаданных (CMDI) CLARIN [15]

Ведущей европейской структурой по поддержке ЛР и языковых технологий является Общеевропейская исследовательская инфраструктура для языковых ресурсов и технологий (CLARIN)[8]. Очевидно, что CLARIN придает большое значение разработке системы метаданных.

CLARIN инициировала разработку Инфраструктуры компонентов метаданных (CMDI). Она обеспечивает основу для описания и повторного использования схем метаданных. Строительные блоки описания (“компоненты”, включающие определения полей) могут быть сгруппированы в готовый формат описания (“профиль”). Оба они хранятся и совместно используются другими пользователями в реестре компонентов для повторного использования. Каждая запись метаданных затем оформляется в виде XML-файла, включая ссылку на профиль, на котором она основана.

Подход CMDI сочетает архитектурную свободу при моделировании метаданных с мощными возможностями исследования и поиска в широком диапазоне ЛР.

На сегодняшний день существует две поддерживаемые версии: CMDI 1.1 и CMDI 1.2. Они не взаимозаменяемы, но CMDI 1.1 метаданных может быть легко преобразован в CMDI 1.2.

Вместо единого формата метаданных CMDI предоставляет основу для создания и использования самостоятельных форматов метаданных. Она опирается на модульную модель так называемых компонентов метаданных, которые могут быть собраны вместе для улучшения повторного использования, взаимодействия и сотрудничества между разработчиками моделей метаданных. Относительно небольшой набор компонентов (общие метаданные, метаданные для текстовых ресурсов, метаданные для мультимедиа и метаданные о людях) может быть объединен в индивидуальные профили.

CMDI – это не просто еще один формат. Это гораздо больше: как метамодель он обеспечивает четко определенную структуру для определения и использования вашего собственного формата. Он также позволяет пользователю интегрировать существующие схемы (IMDI, OLAC) в качестве компонентов и, таким образом, обеспечивает совместимость с существующей базой.

Ни одна единая схема метаданных никогда не сможет удовлетворить все потребности разнородного сообщества исследователей гуманитарных и социальных наук: они варьируются от описания греческих текстов на вазах до анализа жестов в видеороликах YouTube и записи фонетических особенностей телефонных записей. Отсюда и необходимость гибкого решения.

Система метаданных CMDI представляет собой комплект документов, сервисов и инструментов, которые кратко описываются ниже.

[8] CLARIN – Общеевропейская исследовательская инфраструктура для языковых ресурсов и технологий – Common European Research Infrastructure for Language Resources and Technology https://www.clarin.eu/

Спецификация CMDI 1.2 **[16]**

Уже на подготовительном этапе, который начался в 2007 г., CLARIN нуждалась в гибкости в области метаданных, поскольку она сталкивалась со многими типами ресурсов, которые должны были быть точно описаны. Для версии 1.0 был создан инструментарий CMDI, состоящий из XML-схем и таблиц стилей XSLT для проверки и преобразования компонентов, профилей и записей. Эта версия использовалась на протяжении всего этапа строительства CLARIN.

Новая версия CMDI 1.2. добавляет функциональность, а также исправляет некоторые проблемы. Эти изменения освещены в документе CE-2014-0318. Переход от 1.1 к 1.2 поддерживается версией 1.2 инструментария CMDI. Описывается жизненный цикл метаданных, устанавливается содержание работ на каждом этапе. Спецификация CMDI содержит описание структуры файла CMDI, правила установлений отношений между ресурсами, язык описания.

Более подробная информация об изменениях в CMDI 1.2 может быть найдена на странице [17]. Общая информация на этой странице относится как к CMDI 1.1, так и к CMDI 1.2.

Примеры и наборы данных. Раздел **с**одержит примеры – тестовые, реальные, большие и малые наборы данных, в том числе полученные путем сбора метаданных

Руководство по передовой практике CMDI **[18].** Проект содержит общие рекомендации по моделированию и созданию метаданных CMDI CLARIN. Руководство также содержит набор описаний общих подходов и проблем.

Документ о проблемах детализации и моделирования CMDI **[19].** Документ формулирует некоторые условия для создания метаданных:

- CLARIN ориентирован на распределенное хранение ЛР в десятках центров, где большинство из них имеют собственный репозиторий
- ЛР доступны в Интернете
- Метаданные для ресурсов хранятся в формате CMDI, где каждый центр может использовать свой собственный профиль ресурсов
- Все метаданные собираются через OAI-PMH и впоследствии включаются в поисковые системы и порталы
- Создание метаданных полностью находится под контролем центра
- CMDI – основа для исследовательской инфраструктуры, которая должна быть реализована в краткосрочной перспективе, быть надежной, хорошо масштабироваться и быть пригодной для неподготовленных пользователей.

Специальный раздел документа посвящен соотношению метаданных и лингвистических аннотаций. Делается вывод, что хотя эти подходы к описанию ЛР в значительной степени пересекаются, метаданные отличаются большей стабильностью и универсальностью.

CMDI комплект первой помощи **[20].** Включает краткий перечень инструментов и методик:

- Инструменты для проектировщика метаданных
- Инструменты создателя и куратора метаданных
- Инструменты для провайдера метаданных

Реестр компонентов **[21]**. Этот сервис CMDI выполняет следующие функции:

- Регистрация и хранение компонентов / профилей.
- Просмотр зарегистрированных компонентов / профилей.
- Редактирование и создание компонентов / профилей.

Реестр понятий CLARIN **[22]**

Концептуальный (понятийный) реестр CLARIN (CCR) предлагает набор понятий, имеющих отношение к предметной области ЛР с их постоянными идентификаторами. Подробное описание реестра см. ниже, в разделе, посвященном словарям метаданных.

Инструментальные средства CMDI

Редактор метаданных COMEDI – это веб-редактор для метаданных. С помощью COMEDI можно интерактивно создавать новые записи метаданных CMDI или загружать и изменять существующие метаданные, экспортировать в виде XML-файла CMDI и передавать через OAI-PMH.

Coala – инструмент для преобразования различных наборов речевых данных в стандартизированные файлы CMDI.

CMDI Maker – это простое в использовании веб-приложение HTML5 для быстрого создания научных метаданных.

Arbil – это общий редактор метаданных, браузер и органайзер для IMDI, CMDI и аналогичных

форматов метаданных**.** Arbil можно использовать дистанционно, на любом этапе, частично или в целом.

Поисковые инструменты

Прежде всего, это Виртуальная языковая обсерватория [23] – основной поисковый сервис CLARIN. Другой инструмент для работы с ресурсами CLARIN – поисковая машина Института Меертенса [24], при помощи которой можно искать по наименованию коллекций или по схеме профиля ЛР.

Использование CMDI в различных проектах

В литературе описано несколько примеров использования CMDI в различных информационных системах. Так, в работе [25] описано формирование метаданных в формате CMDI для объектов, хранящихся в платформе Fedora. В этом варианте Fedora работала с MySQL и Tomcat. Это было реализовано под управлением Linux, но результат относится и к другим операционным системам.

В другой работе [26] описывается организация репозитория в Институте немецкого языка[9] (IDS). Репозиторий IDS использует Fedora в качестве базовой платформы. Проект описывает 4 этапа.

Выравнивание. Метаданные и данные выравниваются друг с другом. Часто метаданные представляются в виде вектора, разделенного запятыми, со ссылками на фактические данные. Этот шаг обеспечивает однозначное соответствие между данными и метаданными, что часто требует нормализации ссылок и идентификаторов.

Валидация / курирование. Форматы данных проверяются с помощью специальных валидаторов форматов, а для данных, которые недоступны в одном из рекомендуемых форматов, генерируются дополнительные представления (обычно на основе XML DocBook или TEI для письменных корпусов).

Извлечение метаданных. Дополнительные метаданные, такие как название или дата выпуска, извлекаются из данных, используя преимущества форматов данных на основе XML, созданных на предыдущем шаге.

Генерация CMDI. Метаданные преобразуются в подходящий профиль компонента CMDI. Часто этот шаг включает в себя спецификацию нового профиля, частично основанного на существующих компонентах CMDI.

Один из репозиториев CLARIN обеспечивает управление и поддержку многих форматов метаданных в репозиториях на DSpace [27], в том числе CMDI, причем предлагается сопоставительная таблица форматов CMDI и OAI_DC.

Стандартизация метаданных ЛР

На основе модели компонентов метаданных CMDI разработан стандарт *ISO 24622 Управление языковыми ресурсами – Инфраструктура компонентов метаданных (CMDI).* Стандарт имеет 2 части: ISO 24622-1: 2015 Часть 1: Модель компонентов метаданных [28], ISO 24622-2: 2019 Часть 2: Язык спецификации компонентов метаданных [29].

В стандарте указывается, что ландшафт метаданных для ЛР продолжает оставаться фрагментированным: для ЛР использовались широко распространенные схемы метаданных, например, OLAC – это адаптированная версия CMDI. Кроме того, существуют специально разработанные схемы метаданных для определенных типов ЛР (например, IMDI). В результате системы метаданных ЛР плохо совместимы.

Описание модели CMDI – это первая часть инфраструктуры, которая формирует полный пакет для создания схем метаданных. Полный стандарт инфраструктуры будет содержать также один или несколько языков спецификации компонентов метаданных и ряд рекомендуемых компонентов и профилей метаданных. Поскольку эта часть ISO 24622 определяет абстрактную модель, то для ее описания используется UML.

Эта часть ISO 24622 упрощает разработчикам моделей метаданных создание новых схем метаданных, которые, в свою очередь, могут использоваться либо для описания новых типов ресурсов, либо для включения в более подходящее описание ресурсов в конкретных обстоятельствах. Схема метаданных воплощается в записи метаданных.

Определение ресурса в этом контексте очень широкое. В этой части ISO 24622 используется прагматический взгляд: например, изображение может быть ресурсом само по себе, если оно связано с PID[10] и на него можно ссылаться, или оно может быть частью документа, в котором отсутствует собственная идентификация. Ресурс может находиться отдельно в одной среде и рассматриваться как часть коллекции в другой среде. Кроме того, описания ЛР могут различаться в разных контекстах. Эта часть ISO 24622 должна поддерживать все такие случаи, а модель

[9] Das Leibniz-Institut für Deutsche Sprache (IDS) https://www.ids-mannheim.de/

[10] PID – Постоянный идентификатор – Persistent identifier.

должна предоставлять описания на всех уровнях детализации.

В этой части ISO 24622 учитываются два типа коллекций:

а) Сложный ресурс мог быть изначально создан как коллекция, и, не считая управления версиями, он будет существовать как таковой в статической опубликованной форме. Его спецификация будет рассматриваться как независимый объект ответственным архивным учреждением, которое также предоставляет PID для такой коллекции. Архивное учреждение отвечает за поддержание метаданных, представляющих коллекцию.

б) Напротив, другой тип коллекции – это тот, который не планировался и не создавался как коллекция его создателями или архивом, но приобрел статус объединенного ресурса на основе исследований, которые должны быть проверены. Такие коллекции, хотя и специально созданы исследователем, могут не иметь никакого значения вне контекста исследования, для которого они были созданы. Ссылка из исследовательских документов на коллекцию также может стать сложной, если коллекция содержит сотни отдельных ресурсов. Поэтому необходимо фиксировать эти типы коллекций с помощью записи метаданных, которая связана со всеми составляющими ее ресурсами, но только как воплощение этой коллекции. Нет ответственной стороны за ведение этой записи метаданных. Маловероятно, что исследователь, создавший “виртуальную” коллекцию, имеет какой-либо способ постоянно поддерживать и курировать эту запись метаданных в долгосрочной перспективе. Цифровые архивы или издатели могут вести специальные реестры, в которых исследователи могут регистрировать такие виртуальные коллекции.

Оба типа коллекции идентифицируются с помощью PID, который относится к метаданным коллекции.

Сфера применения этой части ISO 24622 состоит в описании модели, которая обеспечивает гибкое построение интероперабельных схем метаданных для ЛР. Схемы метаданных, основанные на этой модели, могут использоваться для описания ресурсов на разных уровнях детализации.

Вторая часть стандарта ISO 24622-2: 2019 содержит представление языка спецификации метаданных компонентов.

В CMDI жизненный цикл метаданных начинается с потребности в моделировании метаданных для определенного типа ресурса. Разработчики моделей могут просматривать и искать в реестре компоненты и профили, которые соответствуют их требованиям. Компонент группирует вместе элементы метаданных, которые потенциально могут быть повторно использованы в другом контексте. Компоненты также могут группировать другие компоненты.

Существующие реестры компонентов могут уже содержать любое количество компонентов. Их можно повторно использовать в том виде, в каком они есть, или адаптировать путем изменения, добавления или удаления некоторых элементов и/или компонентов метаданных. Также могут быть созданы совершенно новые компоненты для моделирования уникальных аспектов рассматриваемых ресурсов. Все необходимые компоненты объединены в один профиль, соответствующий типу ресурсов.

Любой компонент, элемент и значение в таком профиле могут быть связаны с семантическим описанием – концептом, чтобы сделать их значение явным. Эти семантические описания могут храниться в семантическом реестре, например, Регистре понятий CLARIN (см. ниже). Создатели метаданных могут создавать записи для определенных ресурсов, которые соответствуют профилю типа ресурса, и эти записи могут быть предоставлены в локальные и глобальные каталоги.

Вторая часть стандарта, кроме вступления и определения терминов, включает следующие разделы:

- Условные обозначения и пространство имен XML
- Структура экземпляров CMDI
  - Общая структура
  - Основная структура
  - Элемент <Header> element
  - Элемент <Resources> element
  - Элемент <IsPartOfList> element
  - Компоненты CMD
- Язык спецификации компонентов CMDI (CCSL)[11]
  - Общая структура CCSL
  - Заголовки CCSL
  - Спецификации CMD
  - Определение элементов CMD
  - Определение атрибутов CMD

[11] CCSL – Язык спецификации компонентов CMDI – CMDI component specification language.

- Схемы значений для CMD элементов и атрибутов

- Метаданные компонентов (CMD)
  - Преобразование CCSL в определение схемы профиля CMD
  - Общие свойства определения схемы профиля CMD
  - Интерпретация спецификаций CMD в CCSL
  - Интерпретация определений элементов CMD в CCSL
  - Интерпретация определений атрибутов CMD в CCSL
  - Модель контента для элементов и атрибутов CMD в определении схемы

Словари метаданных и реестры категорий данных

Разработка и реализация систем метаданных для описания ЛР существенным образом зависит от качества (полноты, точности и однозначности) терминов, используемых в этих метаданных, как и в других языковых технологиях. Поэтому с начала деятельности по созданию метаданных была поставлена задача формирования реестра категорий лингвистических данных, т.е. словарей терминов (понятий), применяемых для метаописаний ЛР.

Соответствующий стандарт был впервые выпущен ISO TC37 как ISO 12620: 1999, который позже был признан устаревшим и появилось второе издание ISO 12620: 2009 [30]. Оно было также переработано и появился стандарт ISO 12620: 2019. Однако третье издание больше не предоставляет реестр терминов для языковых технологий, теперь оно ограничено терминологическими ресурсами, отсюда пересмотренное название “Управление терминологическими ресурсами – Спецификации категорий данных” [31].

Второе издание ISO 12620: 2009 было переведено на русский язык в качестве российского ГОСТ Р ИСО 12620-2012 [32].

Реестр категорий лингвистических данных под эгидой ISO TC37 был создан в 2008 г. в Институте психолингвистики Макса Планка (MPI) в Неймегене, Нидерланды под названием ISOcat. Оригинальный ISOcat был задуман как официальный онлайновый реестр информации о категориях данных для поддержки исследований и разработок в различных лингвистических дисциплинах. Со временем стало очевидно, что некоторым пользователям будет лучше служить концептуальная база данных, предназначенная для поиска данных в больших текстовых корпусах, что требует иной модели данных, чем ISOcat.

В результате отпала необходимость в регистрирующем органе, и хранилище ISOcat перестало быть проектом ISO. Был создан DatCatInfo (DCR) – репозиторий категорий данных, который заменяет ISOcat.

Следующим этапом развития регистра стало создание для пользователей CLARIN нового реестра CLARIN Concept Registry (CCR). Этот реестр поддерживает институт Meertens. В 2015 г. реестр компонентов был обновлен для использования CCR вместо ISOcat.

Еще несколько проектов по словарям метаданных ЛР (LIME, Lexinfo, ОПТЕЛ) кратко описаны ниже.

***ISOcat* [33].**

Как было отмечено выше, регистр ISOcat в настоящее время не поддерживается. Однако пока эта база данных была доступна, с ней проводились интересные исследования. Например, специалисты Тюбингенского университета разработали интерпретацию дерева зависимостей для регистра ISOcat.

Категории данных (т.е. дескрипторы) имеют определения на естественном языке и мало выраженных связей. С ростом реестра до многих сотен записей становится все более очевидным, что неформальные определения и их конструкция в виде глоссария затрудняют пользователям понимание, использование и управление содержимым реестра.

Разработчики взяли большое подмножество набора терминов ISOcat и восстановили из него древовидную структуру. Такой онтологический реинжиниринг дает пользователям представление о лингвистической терминологии, связанной с метаданными. Полученная иерархия доступна по адресу [34], а для академического и исследовательского использования приводится XML-файл. Каждая запись представлена только в RDF со своим мнемоническим идентификатором, именем, постоянным идентификатором и определением естественного языка. Никакой другой информации, в частности никакой структурной информации, не приводится.

***База данных DatCatInfo* [35].**

DatCatInfo – это репозиторий категорий данных (DCR), разработанный в соответствии с ISO 12620: 2019, который заменяет ISOcat.

Лидеры сообщества пользователей решили создать хранилище определений категорий данных, переименованное в DatCatInfo и поддерживаемое отраслевыми лингвистами и терминологами. DatCatInfo поддерживается LTAC Global/TerminOrgs, которая является связующим звеном с ISO TC37.

Для поиска в массиве данных DatCatInfo разработана база данных TERMWEB [36] с разнообразными поисковыми возможностями.

***Реестр понятий CLARIN* [22]**

Реестр понятий CLARIN (CCR)[12] образует основу семантического слоя совместимости CLARIN, особенно в контексте метаданных, т.е. компонентной инфраструктуры метаданных (CMDI). Реестр предлагает набор понятий с их постоянными идентификаторами, имеющими отношение к предметной области ЛР. CCR содержит 3163 понятия и определения. Пример:

writing systems *The visual representation of spoken language on paper or other media, and the issues involved in writing and creating a writing system. (source: CLARIN).*

системы письма *Визуальное представление устной речи на бумаге или других носителях, а также вопросы, связанные с написанием и созданием системы письма. (источник: CLARIN).*

Доступ к реестру понятий через фасетный браузер может получить любой пользователь в режиме *только для чтения*. Добавление новых понятий или изменение существующих могут быть осуществлены только национальными координаторами CCR.

Фасетный браузер для поиска в CCR [37] предоставляет возможность поиска по части термина или по термину целиком, по полям БД, а также с использованием нескольких фасетных фильтров с развитой классификацией.

***Словарь лингвистических метаданных (LIME)* [38]**

LIME[13] – это словарь для выражения лингвистических метаданных о ЛР и лингвистически обоснованных наборах данных. LIME и связанный с ним Java API опубликован в виде связанных данных.

Сегодня LIME является модулем метаданных пакета словарей OntoLex для обеспечения интерфейсов между онтологиями и лексиконами. Соответствующие метаданные включают список естественных языков, принятых для лексикализации набора данных, лексические модели, принятые для обеспечения лексикализации (например, rdfs:labels, SKOS или SKOS-XL labeling properties, или сам OntoLex), а также статистические данные о покрытии элементов набора данных лексическими записями для каждого данного языка.

Словарь LIME, который дополняет другие существующие словари метаданных, рекомендуется использовать для улучшения видимости онтологий и наборов данных, с тем чтобы улучшить их доступность и квалифицировать их лексическую характеристику.

***LexInfo* [39]**

LexInfo – это онтология, которая была разработана для предоставления категорий данных для модели Lemon с использованием языка разметки Lexical Markup Framework Версию 1.0. С тех пор она была обновлена в виде новой модели OntoLex-Lemon группы сообщества OntoLex. LexInfo теперь представлена на GitHub.

База данных лексики метаданных российских ЛР

К исследованиям лексики метаданных ЛР следует отнести также проект по созданию лексической и концептуальной основы для онтологии по лингвистике на основе лексики метаданных российских ЛР. Создана база данных, которая получила название Онтология поисковых терминов по лингвистике (ОПТЕЛ). Предполагается, что ОПТЕЛ также может служить для навигации и/или метапоиска в репозитории российских ЛР. Принципы отбора источников для ОПТЕЛ, структура БД, особенности отдельных словарей метаданных описаны в работе [40]. В настоящее время ОПТЕЛ реализована и размещена в Интернете [41].

Реализованная версия ОПТЕЛ включает 55 словарей лексики информационных языков и других метаданных, принадлежащих 28 ЛР разных типов. К ним относились тезаурусы, классификации, средства разметки Национального корпуса русского языка и многие другие источники. Всего в ОПТЕЛ представлено свыше 430 тыс. уникальных лексических единиц, объем каждого словаря указан в работе [42]. Там же представлены данные распределения пересечений лексики ОПТЕЛ по словарям и другие сведения.

Заключение

Представленный обзор освещает основные проекты и стандарты, связанные с метаданными ЛР за последние годы. Как легко видеть, работа

[12] CCR – Реестр понятий CLARIN – CLARIN Concept Registry.

[13] LIME – лингвистические метаданные – LInguistic MEtadata.

по ним идет весьма активно. В то же время, к сожалению, роль российских специалистов в этом процессе невелика. Разработан, точнее, переведен, единственный российский ГОСТ, но и тот, насколько известно автору, в практике ЛР не используется. Тем не менее, использование современных развитых моделей метаданных необходимо, чтобы результаты российских разработок ЛР стали частью мирового лингвистического пространства. Поэтому российские разработчики ЛР должны знать и по возможности использовать мировой опыт разработки и применения систем метаданных.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. A Proposal for a Meta Description Standard for Language Resources https://www.mpi.nl/ISLE/documents/papers/white_paper_11.pdf
2. Metadata Elements for Lexicon Descriptions https://www.mpi.nl/ISLE/documents/draft/ISLE_Lexicon_1.0.pdf
3. IMDI Team, (August 2001), Vocabulary Taxonomy and Structure, Version 1.1, MPI Nijmegen
4. Mapping IMDI Session Descriptions with OLAC Draft Proposal Version 1.0 August, 2001 IMDI Technical Report Max-Planck-Institute for Psycholinguistics NL, Nijmegen
5. Arbil for editing and managing IMDI metadata. Version 2.6. https://www.mpi.nl/corpus/html/arbil-imdi/index.html
6. IMDI Documents https://www.mpi.nl/ISLE/documents/docs_frame.html
7. OLAC Metadata http://olac.ldc.upenn.edu/OLAC/metadata.html
8. OLAC Metadata Usage Guidelines http://olac.ldc.upenn.edu/NOTE/usage.html
9. Dublin Core XML https://dcxml.readthedocs.io/en/latest/
10. Documentation and User Manual of the META-SHARE Metadata Model http://www.meta-net.eu/public_documents/t4me/META-NET-D7.2.4-Final.pdf
11. Gavrilidou, M., Labropoulou, P., Piperidis, S., Speranza, M., Monachini, M., Arranz, V., Francopoulo, G. META-NET Deliverable D7.2.1 – Specification of Me_tadata-Based Descriptions for Language Resources and Technologies, 2011, http://t4me.dfki.de/intranet/document_repository/deliverables/wp07-infrastructure-functional-and-technical-specification/meta-net-d7.2.1-final.pdf/view
12. Technologies for the Multilingual European Information Society. Specification of metadata-based descriptions for language resources and technologies. Penny Labropoulou, Maria Gavrilidou, Elina Desipri, Stelios, Piperidis (R.C. Athena. ILSP), Francesca Frontini, Monica Monachini (ILC. CNR), Victoria Arranz (ELDA), Gil Francopoulo (LIMSI). Final Report, 2012 http://www.meta-net.eu/public_documents/t4me/META-NET-D7.2.2-Final.pdf
13. International Standard Language Resource Number http://www.islrn.org/
14. LRE map http://www.elra.info/en/catalogues/lre-map/
15. Component Metadata https://www.clarin.eu/content/component-metadata
16. CMDI 1.2 specification Version 1 Date 2016-10-20 https://office.clarin.eu/v/CE-2016-0880-CMDI_12_specification.pdf
17. CMDI 1.2 https://www.clarin.eu/cmdi1.2
18. CMDI Best Practices Guide https://www.clarin.eu/content/cmdi-best-practices-guide
19. AP3-007-CMDI_and_granularity.pdf https://www.clarin.eu/media/1790
20. CMDI-first-aid-kit.pdf https://www.clarin.eu/sites/default/files/CMDI-first-aid-kit.pdf
21. Component Registry Documentation. Component Registry, Browser and Editor Reference Manual https://www.clarin.eu/content/component-registry-documentation
22. CLARIN Concept Registry https://www.clarin.eu/ccr
23. Virtual Language Observatory (VLO) https://www.clarin.eu/content/virtual-language-observatory-vlo
24. Поисковые сервисы и инструменты Института Меертенса [Search Services and Tools of the Mertens Institute] https://www.meertens.knaw.nl/cmdi/search/#q=*%3A* (In Russ.)
25. Fedora_OAI_Konfiguration_v3.pdf https://www.clarin-d.net/images/leipzig/Fedora_OAI_Konfiguration_v3.pdf
26. IDS Repository Architecture and Ingest Pipelines http://repos.ids-mannheim.de/reposdescription.html
27. Linguistic Data and NLP Tools. About metadata https://lindat.mff.cuni.cz/repository/xmlui/page/ metadata
28. ISO 24622-1:2015 Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model https://www.iso.org/ru/standard/37336.html
29. ISO 24622-2:2019 Language resource management – Component metadata infrasctructure (CMDI) – Part 2: Component metadata specification language https://www.iso.org/obp/ui/#iso:std:iso:24622:-2:ed-1:v1:en
30. ISO 12620:2009 Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources https://www.iso.org/standard/37243.html

31. ISO 12620:2019 Management of terminology resources – Data category specifications https://www.iso.org/standard/69550.html

32. ГОСТ Р ИСО 12620-2012 Терминология, другие языковые ресурсы и ресурсы содержания. Спецификация категорий данных и ведение реестра категорий данных для языковых ресурсов http://docs.cntd.ru/document/1200104401 [*GOST R ISO 12620-2012 Terminologiya, drugie yazykovye resursy i resursy soderzhaniya. Specifikaciya kategorij dannyh i vedenie reestra kategorij dannyh dlya yazykovyh resursov* [GOST R ISO 12620-2012 Terminology, Other Language Resources and Content Resources. Specification of Data Categories and Maintaining a Register of Data Categories for Language Resources] http://docs.cntd.ru/document/1200104401 (In Russ.)].

33. The Center for Sustainability of Linguistic Data (NaLiDa) http://www.sfs.uni-tuebingen.de/nalida/en/

34. Rational Reconstruction for TDG Metadata http://www.sfs.uni-tuebingen.de/nalida/images/isocat/isocat_hierarchy.html

35. Data Category Repository (DCR) http://datcatinfo.net/

36. TERMWEB https://datcatinfo.termweb.se/termweb/app

37. CLARIN Concept Registry Browser https://concepts.clarin.eu/ccr/browser/

38. Linguistic Metadata (LIME) vocabulary https://lod-cloud.net/dataset/lime

39. About the ontology. What is LexInfo? https://lexinfo.net/

40. *Антопольский А.Б., Савчук С.О., Тамеев А.А.* О разработке онтологии поисковых терминов по лингвистике // Информационные ресурсы России. 2020. № 4. С. 2–7. [Antopolsky, A.B., Savchuk, S.O., Tameev, A.A. *O razrabotke ontologii poiskovyh terminov po lingvistike* [On the Development of an Ontology of Search Terms in Linguistics] *Informacionnye resursy Rossii* [Information Resources of Russia]. 2020, No. 4, pp. 2–7. (In Russ.)].

41. Онтология поисковых терминов по лингвистике http://db.inion.ru/optel/ [*Ontologiya poiskovyh terminov po lingvistike* [Ontology of Search Terms in Linguistics] http://db.inion.ru/optel/ (In Russ.)].

42. *Антопольский А.Б., Максимов Н.В., Тамеев А.А.* Экспериментальная база данных источников для создания онтологии по лингвистике // Информационные ресурсы России. 2021. № 3. С. 24–30. DOI: 10.46920/0204-3653_2021_03181_24 [Antopolsky, A.B., Maksimov, N.V., Tameev, A.A. *Eksperimentalnaya baza dannyh istochnikov dlya sozdaniya ontologii po lingvistike* [Experimental Database of Sources for Creating an Ontology on Linguistics]. *Informacionnye resursy Rossii* [Information Resources of Russia]. 2021, No. 3, pp. 24–30. DOI: 10.46920/0204-3653_2021_03181_24 (In Russ.)].

Дата поступления материала в редакцию: 15 июня 2021 г.
Статья поступила после рецензирования и доработки: 30 октября 2021 г.
Статья принята к публикации: 02 ноября 2021 г.
Дата публикации: 28 февраля 2022 г.

Received by Editor on June 15, 2021
Revised on October 30, 2021
Accepted on November 02, 2021
Date of publication: February 28, 2022