ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р ИСО 24619— 2013

МЕНЕДЖМЕНТ ЯЗЫКОВЫХ РЕСУРСОВ

Постоянная идентификация и устойчивый доступ

ISO 24619:2011
Language resource management —
Persistent identification and sustainable access
(IDT)

Издание официальное

Москва
Стандартинформ
2015

Предисловие

1 ПОДГОТОВЛЕН ЗАО «Проспект» на основе собственного аутентичного перевода на русский язык международного стандарта, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 55 «Терминология, элементы данных и документация в бизнес-процессах и электронной торговле»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 8 ноября 2013 г. № 1388-ст

4 Настоящий стандарт идентичен международному стандарту ИСО 24619:2011 «Менеджмент языковых ресурсов. Постоянная идентификация и устойчивый доступ» (ISO 24619:2011 «Language resource management — Persistent identification and sustainable access»).

При применении настоящего стандарта рекомендуется использовать вместо ссылочных международных стандартов соответствующие им национальные стандарты Российской Федерации, сведения о которых приведены в дополнительном приложении ДА

5 ВВЕДЕН ВПЕРВЫЕ

Правила применения настоящего стандарта установлены в ГОСТ Р 1.0—2012 (раздел 8). Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (gost.ru)

Содержание

Введение

Ссылки и цитаты являются важной частью документации и публикаций. Авторы обычно используют цитаты и ссылки как средство выражения своей признательности авторам других публикаций, служащих источником информации для их собственной работы, либо в поддержку собственной аргументации. Цитаты, как правило, содержат сведения, помогающие читателю оценить степень связи цитируемой публикации с рассматриваемой проблемой и однозначно идентифицировать эту публикацию. На основании приводимых в цитате сведений любой библиотекарь или соответствующее компетентное лицо могут легко найти нужный документ, воспользовавшись установленными для этого стандартными процедурами.

Существование документов, непосредственно доступных во всемирной сети, способствовало добавлению к ссылочной информации указаний о ее местонахождении (с помощью так называемого универсального идентификатора ресурсов URI [4]). Такая практика сделала возможным прямой доступ к ссылочным документам через веб-браузеры и другие аналогичные средства визуализации. Подобные механизмы уже были установлены в таких стандартах, как ИСО 690, хотя там основное внимание уделяется скорее простому поиску в библиографических каталогах публикаций, чем обеспечению их постоянной доступности. Все чаще и чаще справочники такого рода требуются для использования прикладными программно-техническими системами, равно как и людьми, которым необходим надежный доступ к запрашиваемым ресурсам. Трудности с получением доступа, возникающие при перераспределении ресурсов, привели к созданию инфраструктуры (framework) постоянных идентификаторов (PID) [23, 24]. В рамках современных подходов [18, 19, 24] к решению проблемы перераспределения ограниченных ресурсов организуются соответствующие координационные службы, которые преобразуют идентификатор ресурса в адрес его текущего местоположения. Такие координационные службы обладают одним важным дополнительным преимуществом, состоящим в ассоциировании с каждым идентификатором расширенных метаданных. В более сложных инфраструктурах, таких как система цифровых идентификаторов объекта DOI (Digital Object Identifier) [14], эта возможность используется для привлечения дополнительных услуг — например информации по авторским правам.

Практика использования наряду с отдельными ресурсами и наборами данных постоянных идентификаторов для целей цитирования и извлечения научной информации имеет более скромные результаты. Однако она располагает не меньшими ресурсами в том смысле, что позволяет читателям публикаций и пользователям ресурсов научных знаний получать прямой доступ к информации научных первоисточников, на которых основаны соответствующие ресурсы. При использовании ссылок для доступа к научным данным, включая языковые ресурсы, важную роль приобретает возможность доступа, в том числе и к отдельным компонентам (частям) ресурсов. В особенности это касается сферы языковых ресурсов, где один и тот же набор данных или коллекция ресурсов обычно имеют несколько уровней разбиения. Поэтому предметом настоящего стандарта, касающегося использования инфраструктур PID и требований к ним, является, в первую очередь, анализ механизмов эффективной организации идентификации и обеспечения доступа к компонентам требуемых ресурсов. Содержащиеся в стандарте конкретные рекомендации показывают, каким образом следует подходить к решению задачи разбиения применительно к использованию постоянных идентификаторов отдельных ресурсов и их коллекций.

Потребность в применении инфраструктур постоянных идентификаторов для определения ресурсов, имеющихся в наборах научных данных, возрастает еще и потому, что современные архивы и репозитории начали превращаться в сплетения сетей родственных комплексных ресурсов, которые могут быть распределены по множеству сетевых узлов. В такой ситуации обязательным условием становится наличие постоянных устойчивых связей. Например, в мультимедийной среде словарных ресурсов лексическая единица может быть связана с наглядными представлениями, которые необязательно физически присутствуют в данном словаре или даже могут запрашиваться по ссылке с другого сайта, принадлежащего совсем другой организации. Несмотря на это связь между данной лексической единицей и относящимся к ней изображением должна оставаться действующей даже в тех случаях, когда некоторые серверы или файлы могут со временем изменять свое местоположение. Появляющиеся сценарии электронного распространения научной информации, делающие возможным использование распределенных услуг по обработке распределенных ресурсов, целиком и полностью зависят от наличия удобного и прозрачного доступа к любым службам обработки данных независимо от их местонахождения и от конкретных организаций, эксплуатирующих соответствующие ресурсы. Это значит, что

механизм разрешения обращений к используемым ресурсам не должен обременяться никакими излишними ограничениями и зависимостями, которые влекли бы за собой создание нежизнеспособной или непредсказуемой сети услуг как в техническом, так и в организационном плане.

Выполнение требований доступности таких служб, как инфраструктуры PID для всего сообщества поставщиков языковых и технологических ресурсов, еще больше усложняется из-за необходимости предоставления системы разрешимых ссылок с постоянными идентификаторами без обременения поставщиков ресурсов какими-либо иными коммерческими условиями кроме четко сформулированных основополагающих требований технической поддержки и сопровождения соответствующих ресурсов в сети Интернет.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

МЕНЕДЖМЕНТ ЯЗЫКОВЫХ РЕСУРСОВ

Постоянная идентификация и устойчивый доступ

Language resource management. Persistent identification and sustainable access

Дата введения — 2015—01—01

1 Область применения

Настоящий стандарт устанавливает требования к инфраструктуре постоянных идентификаторов [persistent identifier (PID)] и их использованию в качестве средств запроса и цитирования языковых ресурсов в официальных документах, равно как и в рамках самих этих ресурсов. В этом плане примерами языковых ресурсов могут служить электронные словари, лингвистические терминологические ресурсы, лексикон систем машинного перевода, аннотированные мультимедийные и многоцелевые корпуса текстов, снабженные морфосинтаксической информацией и прочими атрибутами. Такие ресурсы создаются специалистами по прикладной лингвистике и информационным технологиям.

Настоящий стандарт охватывает также вопросы стойкости идентификаторов и разбиения запросов к ресурсам прежде всего в аспекте необходимости реализации постоянных ссылок на основе использования инфраструктуры PID и последующего предъявления жестких требований к любым применяемым для этого инфраструктурам PID.

Инфраструктуры PID позволяют ассоциировать метаданные общего характера с идентификатором, который может содержать также цитируемую информацию. Настоящий стандарт устанавливает минимальные требования, касающиеся эффективного использования постоянных идентификаторов применительно к языковым ресурсам, и содержит ссылки на несколько подходящих действующих стандартов и стандартов *de-facto*, таких как ИСО 690 [16], APA [3], MLA [9] для цитируемой информации, ИСО/МЭК 21000-17, IETF RFC 5147, Annotea [2], документ temporal-fragment [22], XPointer для синтаксиса идентификаторов разделов, PURL [23], ARK [18], Handle System [24] и DOI [14].

2 Нормативные ссылки

ИСО 12620:2009 Терминология, другие языковые ресурсы и ресурсы содержания. Спецификация категорий данных и ведение реестра категорий данных для языковых ресурсов (ISO 12620:2009, Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources)

ИСО/МЭК 21000-17:2006 Информационные технологии. Мультимедийная инфраструктура (MPEG-21). Часть 17. Идентификация фрагментов ресурсов MPEG (ISO/IEC 21000-17:2006, Information technology — Multimedia framework (MPEG-21) — Part 17: Fragment Identification of MPEG Resources)

Документ W3C 2003, *XPointer Framework*: [в режиме онлайн] Рекомендация W3C от 25 марта 2003 г. [по состоянию на 4 августа 2010 г.]. Доступна по адресу: http://www.w3.org/TR/xptr-framework/

Wilde, E. and Duerst, M. *URI Fragment Identifiers for the text/plain Media Type*, IETF RFC 5147, апрель 2008 г. [по состоянию на 22 декабря 2010 г.]. Ресурс доступен по адресу: http://www.rfc-editor.org/rfc5147.txt

Издание официальное

3 Термины и определения

В настоящем стандарте применены следующие термины с соответствующими определениями:

3.1 Ресурсы

3.1.1 **ресурс** (resource): Цифровой объект в сети Интернет, снабженный индивидуальным идентификатором, который может запрашиваться с помощью унифицированного идентификатора ресурса URI (3.2.2).

Примечание 1 — Определение заимствовано из документа IETF RFC 3986.

Примечание 2 — В контексте настоящего стандарта ресурсом может быть также языковой ресурс, имеющий оперативно доступное электронное представление.

Примечание 3 — Ресурс может иметь несколько разных представлений. В зависимости от **инфраструктуры PID** (3.2.5) способ распознавания конкретного представления может быть закодирован в идентификаторе (см. ARK, B.3) или определяться в процессе взаимодействия между **клиентом сети** (3.3.8), использующим развернутый PID для вызова **ресурса** (3.1.1), и **сервером ресурсов** (3.3.6).

3.1.2 **языковой ресурс** (language resource): Электронный ресурс, который предоставляет информацию по одному или нескольким языкам.

Примечание — К языковым ресурсам относятся лексикографические, терминологические, морфосинтаксические, корпусно-ориентированные и семантические ресурсы, а также электронные ресурсы, используемые для изучения лингвистических явлений, подобных связным текстам и мультимедийным/многомодальным записям. Они создаются и применяются, в частности, лингвистами, специалистами по информационным технологиям, лексикографами и терминологами. Подобные информационные продукты состоят из множества коротких записей, скомпонованных в более крупные публикации, и зачастую являются авторитетными изданиями, как, например, стандартизованные терминологические справочники и глоссарии, выпускаемые органами по стандартизации, такими как ИСО, IETF, W3C и др.

3.1.3 **комбинированный ресурс** (complex resource): Ресурс (3.1.1), состоящий из множества компонентов, каждый из которых может быть запрошен отдельно.

Примечание — Комбинированный ресурс может быть интегрированным, если его составные части распределены по различным **репозиториям** (3.1.6).

3.1.4 **коллекция** (collection): Компоновка, образованная любым числом ресурсов (3.1.1) и запрашиваемая как единое целое.

3.1.5 **публикуемая коллекция** (published collection): Коллекция ресурсов определенного назначения, которая поддерживается как независимая сущность в рамках архива (3.1.7) либо репозитория (3.1.6) и для которой существует адекватная форма цитирования (3.1.16) информации.

3.1.6 **(цифровой) репозиторий** (digital repository; repository): Средство, обеспечивающее надежный доступ к управляемым электронным ресурсам (3.1.1).

3.1.7 **(цифровой) архив** (archive; digital archive): Репозиторий (3.1.6), предназначенный для надежного долговременного хранения специализированных данных.

Примечание — Часто данные цифровых архивов бывают доступны в интерактивном режиме, что требует использования надежных **постоянных идентификаторов** (3.2.4).

3.1.8 **реализация коллекции ресурсов, реализация** (resource collection incarnation; incarnation): Виртуальное воплощение не согласуемой иным способом разрозненной коллекции (3.1.4) определенного назначения, запрашиваемой по единственному идентификатору PID (3.2.4), который связан с идентификатором компонентов (3.2.7) в целях обеспечения раздельного доступа к ним.

Примечание — В библиографическом или предметном указателе может применяться единственный идентификатор PID в сочетании с расширениями для обеспечения доступа к различным компонентам совокупности **ресурсов** (3.1.1), используемой при подготовке монографии или разработке проекта, без фактического сосредоточения физических файлов в одном месте; то есть отдельные элементы коллекции остаются на своих исходных местах, но запрашиваются как части виртуального целого.

3.1.9 **версия** (version): Конкретная форма или вариант ресурса (3.1.1), отличающиеся от других реализаций ресурса хотя бы по одному аспекту или элементу информации.

Примечание — Версии часто идентифицируются по порядковому номеру (например, версия 1, версия 2 и т. д.), однако идентификация версий динамических ресурсов, подверженных частым изменениям, осуществляется зачастую с привязкой к отметке даты и времени.

3.1.10 **мгновенная копия** (snapshot): Мгновенная копия текущих характеристик ресурса (3.1.1), отображающая состояние ресурса или коллекции в определенный момент времени.

3.1.11 **абстрактный ресурс** (abstract resource): Запрашиваемый несетевой ресурс, идентифицируемый по URI (3.2.2); обычно это бывают такие концепты, как класс или свойство.

Примечание — На практике распространены, например, в рамках онтологий RDFS (RDF Schema) или OWL (язык сетевых онтологий) способы идентификации абстрактных ресурсов с помощью URI. В веб-архитектуре никогда не требуется вызов такого информационного ресурса. Если идентификатор абстрактного ресурса не отмечен как подлежащий **разыменованию** (3.4.1), что имеет место применительно к URI пространства имен языка XML, то для такого ресурса нет необходимости формировать идентификатор **PID** (3.2.4).

3.1.12 **раздел ресурса, часть** (resource part; part): Идентифицируемый доступный объект, встроенный в автономный ресурс (3.1.1) или в более крупную часть этого ресурса.

Примечание — Возможно встраивание одних разделов в другие. В динамичной сетевой среде разбиение на разделы подлежит изменению и интерпретации, вследствие чего требуется определенный уровень участия пользователя в принятии решений относительно обозначения и идентификации получаемых подобъектов.

3.1.13 **фрагмент** (fragment): Некоторая порция или подмножество первичного ресурса (3.1.1), какое-либо из представлений первичного ресурса или некоторый другой ресурс, определенный или описанный как компонент ресурса, определенного или описанного этими представлениями.

Примечание 1 — Определение заимствовано из документа IETF RFC 3986.

Примечание 2 — В контексте настоящего стандарта термин *фрагмент* используется только в смысле определения, данного в IETF RFC 3986, тогда как в сетевом контексте фрагмент извлекается **клиентским приложением** (3.3.5) из содержащего этот фрагмент ресурса.

3.1.14 **конечная (неделимая) часть** (terminal part): Часть (3.1.12) ресурса (3.1.1), которая не может быть разбита на более мелкие части.

3.1.15 **внутренняя часть** (internal part): Часть (3.1.12) ресурса (3.1.1), которая является вложением ресурса и одновременно разбита на более мелкие части.

3.1.16 **цитата** (citation): Информационный объект, содержащий сведения, которые переориентируют внимание читателя или пользователя с одного ресурса (3.1.1) на другой.

3.1.17 **ссылка** (reference): Цифровой объект, указывающий на данные, хранимые в каком-либо месте.

Примечание — Хотя термины **«цитата»** (3.1.16) и «ссылка» обычно используются почти как синонимы, для целей настоящего стандарта цитаты несут информацию для читателей и пользователей, тогда как ссылки указывают точное местоположение запрашиваемого **ресурса** (3.1.1). Ссылки могут быть машиночитаемыми и могут конфигурироваться как активизируемые при соответствии заданным критериям.

3.1.18 **уровень аннотирования** (annotation tier): Отдельный информационный уровень, содержащий комментарии, примечания, объяснения или иные типы внешних ремарок, которые могут присоединяться к ресурсу (3.1.1).

Примечание — Например, карты или изображения могут аннотироваться с привлечением дополнительной информации, а текстовые корпуса могут аннотироваться как прямо в тексте, так и автономно.

3.1.19 **автономное аннотирование** (standoff annotation): Режим аннотирования, при котором аннотации хранятся вне аннотируемого документа.

3.2 Идентификаторы

3.2.1 **(цифровой) идентификатор** (identifier; digital identifier): Последовательность символов, ассоциируемая с цифровыми/нецифровыми или абстрактными объектами, такими как книги, изображения, статьи, метаинформационные записи или события.

3.2.2 **унифицированный идентификатор ресурса** (URI; Uniform Resource Identifier): Символьная строка, используемая для идентификации или назначения имени некоторого ресурса (3.1.1) в соответствии с синтаксическими правилами, определенными документом IETF RFC 3986.

3.2.3 **схема именования URI** (URI naming scheme): Верхний уровень структуры присваивания имен URI.

Примечание 1 — Каждая такая схема устанавливает собственные синтаксические правила для **URI** (3.2.2).

Примечание 2 — Типичными примерами схем URI являются http, https, ftp, mailto и др.; все они регистрируются в Полномочном органе по цифровым интернет-адресам IANA (Internet Assigned Numbers Authority).

3.2.4 **постоянный идентификатор** (PID; persistent identifier): Уникальный идентификатор (3.2.1), который гарантирует постоянный доступ к цифровому объекту с помощью механизма доступа, не зависящего от физического расположения объекта и от его текущей принадлежности.

Примечание — «Уникальный» означает в данном случае, что PID не будет повторяться применительно к другим ресурсам. Однако один и тот же PID может по усмотрению поставщика ресурса вызывать разные формы представления или разные **реализации** (3.1.8) конкретного ресурса.

3.2.5 **инфраструктура PID** (PID framework): Схема определения строк идентификатора [схема PID (3.2.4)] для цифровых объектов, доступных через сеть Интернет, и механизм, позволяющий преобразовывать эти идентификаторы в текущие унифицированные индикаторы URI (3.1.1) запрашиваемых объектов.

Примечание 1 — Инфраструктура PID в контексте настоящего стандарта облегчает доступ к индивидуальным объектам, равно как к **разделам** (3.1.12) и **фрагментам** (3.1.13), содержащимся в таких объектах. Инфраструктура PID может целиком зависеть от существующих сетевых протоколов преобразования адресов или может закреплять схему взаимодействия преобразователей адресов на основе программ-посредников.

Примечание 2 — Инфраструктура PID в контексте настоящего стандарта позволяет также оперировать другой информацией, ассоциируемой с конкретным PID.

3.2.6 **активируемый идентификатор** (actionable identifier): URI (3.2.2), имеющий ресурсный идентификатор (3.2.1), который закодирован таким образом, что при вложении этого URI в веб-документ и последующем выборе с помощью указателя мыши браузер будет перенаправлен на запрашиваемый ресурс (3.1.1) и, возможно, на связанные с этим ресурсом дополнительные услуги.

Примечание 1 — Эта функциональная возможность приводит к тому, что URI указывает на подходящий **промежуточный преобразователь адресов** (3.3.7).

Примечание 2 — В некоторых **инфраструктурах PID** (3.2.5) **идентификаторы PID** (3.2.4) представляют собой URI и активируются автоматически.

3.2.7 **идентификатор раздела (ресурса)** (resource part identifier; part identifier): Строка символов, вызывающая раздел ресурса (3.1.12), который может идентифицироваться тем или иным способом в рамках данного типа ресурса (в рабочей среде это может быть время, в изображении — нужная область, в потоке данных — запись и т. п.).

Примечание — Идентификаторы разделов в контексте настоящего стандарта предназначаются для использования на стороне сервера — в противоположность клиентской стороне преобразования адресов, где для этого служат **идентификаторы фрагментов** (3.2.8).

3.2.8 **идентификатор фрагмента** (fragment identifier): Идентификатор (3.2.1), используемый для вызова раздела (3.1.12) ресурса (3.1.1) в сетевой среде.

Примечание 1 — Определение заимствовано из документа IETF RFC 3986.

Примечание 2 — Идентификатор фрагмента, как он определен в IETF RFC 3986, начинается со знака «решетки» («#») и завершается идентификатором **URI** (3.2.2). **Фрагменты** (3.1.13) в смысле RFC разрешаются и извлекаются из ресурса с помощью локального **клиентского приложения** (3.3.5).

Примечание 3 — Существует проектное предложение W3C по изменению указанного способа обработки фрагментов [27].

3.3 Ролевые функции, организации и службы

3.3.1 **институт архивирования** (archiving institution): Организация, ответственная за ведение электронного архива (3.1.7).

3.3.2 **поставщик ресурса** (resource provider): Организация, которая обеспечивает оперативную доступность ресурса (3.1.1).

Примечание — Ресурсом может быть и какая-либо служба.

3.3.3 **резольвер, преобразователь PID-идентификаторов** (resolver; PID resolver): Прикладная программа, преобразующая исходный идентификатор (3.2.1) в другой, более удобный идентификатор, который транслирует PID (3.2.4) ресурса в его URI (3.2.2) и таким образом указывает клиентскому приложению местоположение нужного ресурса (3.1.1).

3.3.4 **система разрешения идентификаторов** (resolution system): Система, предназначенная для поддержки представления неизменного идентификатора (3.2.4) сетевой службе в целях получения в ответ одной или нескольких порций текущей информации, относящейся к идентифицируемому объекту, например о местонахождении [URI (3.2.2)] объекта или метаданных.

Примечание — Сложная система разрешения идентификаторов может рассматриваться как **резольвер PID-идентификаторов** (3.3.3), но часто реализуется как совокупность различных преобразователей адресов или служб разрешения имен.

3.3.5 **клиентское приложение** (client application): Прикладная программа, которая получает доступ к дистанционному обслуживанию обычно в другой компьютерной системе.

3.3.6 **сервер выделения ресурсов** (resource server): Компьютер, который в конечном итоге предоставляет доступ к объекту, указанному в запросе конкретного клиентского приложения.

3.3.7 **промежуточный преобразователь адресов; преобразователь адресов HTTP** (resolver proxy; HTTP resolver proxy): Приложение, которое реализует службу поддержки использования PID (3.2.4), преобразованных в URI (3.4.3), для доступа к ресурсам или к иной информации, относящейся к PID, или к тому и другому одновременно.

3.3.8 **сетевой клиент** (web client): Клиентское приложение, способное осуществлять доступ к интернет-ресурсам с использованием протокола HTTP.

3.4 Действия

3.4.1 **разыменовать** (dereference): Осуществить доступ к запрошенному значению по ссылке (3.1.17).

Примечание — Использование **URI** (3.2.4) в контексте разыменования означает получение того представления ресурса, на которое указывает URI.

3.4.2 **преобразовать, разрешить** (resolve): Преобразовать идентификатор (3.2.1) в другое имя или в адрес, подходящий для осуществления доступа к ресурсу.

Примечание — Процесс преобразования (разрешения) адресов может потребовать нескольких шагов, вплоть до получения соответствующего адреса ресурса.

3.4.3 **преобразовать идентификатор в URI** (urlify an identifier): Закодировать идентификатор (3.2.1) как соответствующий URI (3.2.4).

Примечание — Например, это может быть сделано в целях создания **активируемого идентификатора** (3.2.6).

4 Общее описание

Постоянные идентификаторы (PID) могут существовать в любых видах электронных ресурсов, и настоящий стандарт не устанавливает для них никаких явных ограничений, однако следует иметь в виду, что тип ресурса, определенный PID, оказывает определенное влияние на требования к конкретным индивидуальным постоянным идентификаторам. Ресурсы могут принадлежать к одному из трех основных типов:

- независимые ресурсы, отображенные на рисунке 1;
- любой раздел подобного индивидуального ресурса, требующий в дальнейшем описания, и
- коллекция ресурсов, запрашиваемая как единое целое.

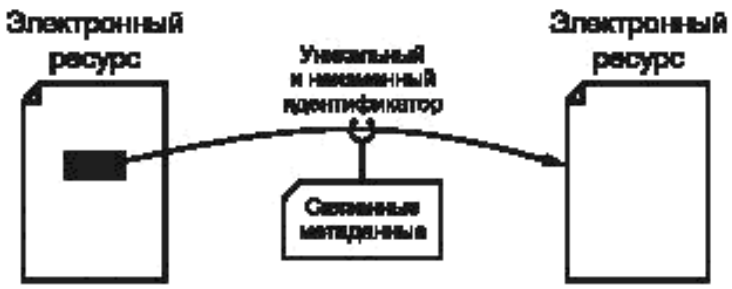

Рисунок 1 — Использование уникальных PID для указания пути от исходного ресурса к целевому

Настоящий стандарт распространяется на способы формирования машиночитаемых уникальных ссылок на электронные ресурсы. На рисунке 1 отображен постоянный идентификатор, включенный в исходный ресурс и указывающий на целевой (требуемый) ресурс. Этот PID может ассоциироваться с разнообразными метаданными.

В данном случае характер ресурса определен очень широко, и средства обращения к нему подлежат уточнению в соответствующем контексте. Например, изображение может быть независимым ресурсом, который снабжен своим собственным уникальным PID и может запрашиваться непосредственно, или оно может быть вложено в документ, где не имеет собственного идентификатора, и тогда оно является компонентом документа. Кроме того, ссылка может указывать на часть этого изображения. Отдельный ресурс в одной рабочей среде может рассматриваться как самостоятельный, а в другой — как раздел комбинированного ресурса. Внутренняя часть ресурса может считаться его конечной частью, но дальнейшая обработка такой части в динамической среде может привести к тому, что появится объект, который сам будет содержать доступные для запроса разделы нижележащего уровня. Настоящий стандарт предназначен для обеспечения поддержки всех таких ситуаций.

В случае комплексных языковых ресурсов некоторые из них должны снабжаться собственными индивидуальными неизменными идентификаторами. При этом другие ресурсы действуют как вложенные, состоящие из множества составляющих, и тогда содержащему их ресурсу должен присваиваться собственный PID, а его разделы могут запрашиваться посредством добавления к этому PID идентификаторов разделов. Настоящим стандартом устанавливаются руководящие принципы выбора надлежащего подхода применительно к любому заданному ресурсу.

В настоящем стандарте обобщаются опыт использования существующих стандартов и достижения сложившейся практики применения различных форматов идентификаторов разделов и фрагментов ресурсов; на этой основе предлагаются руководящие принципы принятия решений в тех ситуациях, когда существующие стандарты оказываются неадекватными или просто неприменимы. Более подробное обсуждение типов ресурсов, охватываемых настоящим стандартом, можно найти ниже, в приложении А.

Применительно к коллекциям языковых ресурсов далее выделяются два типа коллекций:

- Коллекции ресурсов, которые поддерживаются как публикуемые комплексные ресурсы в режиме, близком к статическому, таким образом, что сама коллекция как таковая определяется архивом или репозиторием как независимый объект и получает постоянный идентификатор соответствующего типа. Организация, которой принадлежит архив, отвечает за обеспечение связи между PID и коллекцией, представленной, например, как элемент метаданных в каталоге.

- Другой тип коллекции, которая изначально не рассматривалась создателями или архивными организациями как таковая, однако близка к достижению статуса комплексного ресурса в результате проведения научно-исследовательской или какой-то иной работы, в ходе которой требуется проверять достоверность информации, как, например, при подготовке монографии или выполнении научного проекта. Коллекции такого типа, хотя они и создаются авторами целенаправленно, могут не иметь никакой практической ценности вне контекста первоначальной работы, под которую создавались. Переход от научных документов к коллекции может стать громоздким уже при объеме коллекции в несколько сотен отдельных ресурсов. В результате этого появляется необходимость обращения к коллекциям такого типа с помощью постоянного идентификатора PID, который ассоциируется со всеми составляющими коллекцию ресурсами и надлежащими метаданными. Очевидно, что рассмотренный способ обращения к коллекции возможен лишь в том случае, если она реализована должным образом.

5 Требования к инфраструктуре и использованию постоянных идентификаторов

5.1 Общие сведения

Существующие стандарты и достижения сложившейся практики использования ссылок и цитат, главным образом, в сфере языковых ресурсов можно найти в приложении А, а в настоящем разделе сначала подробно рассматриваются требования к самой инфраструктуре PID, а затем формулируются требования к применению PID в качестве средства запроса и цитирования языковых ресурсов.

5.2 Требования к инфраструктуре постоянных идентификаторов

5.2.1 Общие положения

Инфраструктура PID должна поддерживать следующие функциональные возможности:

а) преобразование одиночного PID к множеству URI или служб;

b) установление связи с релевантными метаданными и обеспечение доступа к ним;

c) адекватная защита для предотвращения злоумышленного или случайного изменения отображений PID/URI и связей PID/метаданные;

d) адресация разделов ресурса (идентификаторы разделов или фрагментов, либо те и другие);

e) кодирование PID как URI для представления идентификаторов как активируемых в среде веб-документов без требования изменений на стороне клиента.

5.2.2 Обеспечение возможности дублирования ресурсов

В целях защиты данных и обеспечения высокоскоростного доступа принято предоставлять дубликаты либо образы ресурсов или резидентные копии на серверах разных ресурсов. Инфраструктура PID должна поддерживать такой вид дублирования, разрешая ассоциирование многих URI с единственным PID.

5.2.3 Доступ к метаданным ресурса

Помимо предоставления надежного URI для ресурса, инфраструктура PID используется также для высоконадежной защищенной привязки метаданных к ресурсу. Хотя настоящий стандарт не требует наличия доступных метаданных какого-либо конкретного типа, в нем все же устанавливается требование возможности преобразования PID к соответствующей записи в формате XML, на основе которой может строиться работа других служб.

5.2.4 Надежное защищенное администрирование

Инфраструктура PID должна обеспечивать адекватную защиту, с тем чтобы только владелец или блюститель ресурса имели возможность изменить отображение PID/URI или ассоциируемые с ним метаданные.

5.2.5 Идентификаторы разделов ресурсов

Невозможно предоставить PID для каждого идентифицируемого раздела ресурса или даже просто определить в глобальном масштабе все возможные варианты разделения ресурсов на части. Следовательно, инфраструктуры PID должны обеспечивать систему присваивания идентификаторов разделам или фрагментам в комбинации с PID-идентификатором ресурса. Поскольку цель здесь состоит в том, чтобы использовать для идентификации разделов ресурса единственную строку, синтаксис PID должен поддерживать операцию конкатенации PID и идентификатора раздела. Например, система разрешения PID должна допускать такое конфигурирование, чтобы комбинация идентификатора раздела PID преобразовывалась в идентификатор URI, который может правильно интерпретироваться сервером ресурса для выдачи запрошенного раздела ресурса [см. раздел А.2,а].

Рисунок 2 — Обработка идентификаторов разделов резольвером PID (на примере реализации сервера дескрипторов)

Комбинированный ресурс «А» с составляющими x, y и z идентифицируется в рамках PID как «1839/А».

Резольвер PID преобразует идентификатор «1839/А» к URI вида http://oserver/objectA, который может распознаваться сервером как запрос на предоставление объекта «А». Раздел «z» ресурса «А» идентифицируется PID как «1839/A#z». Для простоты процедуры разыменования, выполняемой сервером ресурса, резольвер PID должен быть способен к преобразованию идентификатора «1839/A#z» в URI вида http://oserver/objectA?part=z или в какую-то похожую запросную строку, из которой серверу объекта А будет понятно, что требуется предоставить раздел «z».

5.2.6 Унифицированные PID-идентификаторы

Инфраструктура PID должна предоставлять реализацию промежуточного резольвера, который способен разрешать постоянные идентификаторы PID, закодированные как URI. Это позволяет сетевым клиентам преобразовывать такой идентификатор с использованием схемы HTTP, не прибегая к специальным браузерным вставкам или другим специализированным программным средствам. В некоторых инфраструктурах сами PID изначально представляют собой универсальные указатели ресурсов HTTP, и в таких случаях промежуточные резольверы адресов становятся излишними.

5.3 Использование постоянных идентификаторов

Ссылки на доступные в сети Интернет электронные языковые ресурсы должны сопровождаться постоянным идентификатором PID, который преобразуется либо непосредственно к URI запрашиваемого ресурса, либо к метаинформационной записи, характеризующей требуемый ресурс. Последний вариант может использоваться в том случае, если сам ресурс не может быть сделан непосредственно доступным или представляет собой коллекцию, применительно к которой метаинформационная запись должна содержать идентификаторы для составляющих ресурсов коллекции.

Если PID превращается в URI ресурса, то метаинформационная запись формата XML, совместимая с объявленной схемой, принадлежащей ресурсу, должна ассоциироваться с идентификатором и делаться доступной через систему преобразования. Цитируемая информация подлежит включению в метаинформационную запись. Никакие другие требования к записи метаданных, связанной с идентификатором ресурса, не определяются.

В веб-документах идентификатор (например, описатель), вложенный в ссылки, должен помимо соответствия синтаксической схеме PID еще и присутствовать в форме закодированного URI, с тем чтобы он активизировался в веб-браузере или в другом приложении для визуального просмотра документов. Например, таким приложением может быть реализация Handle System (HS):

1839/00-0000-0000-0000-4 -> http://hdl.handle.net/1839/00-0000-0000-0000-0000-4

Адрес резольвера http://hdl.handle.net указывает на промежуточный преобразователь HTTP, который выступает в роли приложения, способного получать HTTP-запросы в виде идентификаторов PID, преобразованных в URI, и использовать «реальный» резольвер PID для перенаправления клиента к запрошенному ресурсу. В примере с технологией Handle System промежуточный преобразователь HTTP является либо центральным интерфейсным модулем резольвера Handle System, либо интерфейсным модулем резольвера, доступ к которому гарантируется поставщиком ресурса.

Спецификации разделов ресурсов, аналогичные спецификациям страниц или разделов, подлежат присоединению всюду, где это возможно, к релевантному PID с помощью соответствующего разделительного знака — например, с использованием стандартного расширения HS:

1839/00-0000-0000-0000-4@time(100s,200s)

Это расширение представляет собой ссылку на сегмент аудиофайла (где идентификатор раздела — не стандартный).

Если совместимый идентификатор фрагмента существует для данного типа ресурса, такой элемент может быть добавлен к закодированному URI с целью создания составного активируемого идентификатора. Для технологии HS примером может быть следующий идентификатор:

http://handle.net/1839/00-0000-0000-0000-0000-4?urlappend=#ffp(track_ID=101)*mp(/~time('npt','50'))

В этом примере используется специальная функция «urlappend» промежуточного преобразования абстрактного идентификатора к разрешенному дескриптору.

5.4 Цитируемая информация и постоянные идентификаторы

Лучшие достижения сложившейся практики могут быть реализованы лишь при условии, что идентификаторы URI заменены идентификаторами PID, а спецификации разделов документа преобразованы в идентификаторы разделов или фрагментов. Ссылки из веб-документов должны также содержать PID, закодированный как URI, если синтаксис PID не согласуется со схемой URI, зарегистрированной IANA [13].

5.5 Обращение к разделам ресурсов

5.5.1 Общие замечания

В качестве идентификатора раздела ресурса может использоваться любой существующий идентификатор, определенный стандартом ИСО или IETF, если не оговорено иное. Применительно к ресурсам, для которых таких стандартов нет, разрешается использовать удобный для прочтения человеком

цитируемый текст; например, для интервала времени это может быть «от 10 до 120 с». Однако такой подход неприменим в случае клиентских программ.

При использовании идентификаторов разделов или фрагментов для извлечения или разыменования раздела ресурса важно четко понимать разницу между применением идентификаторов фрагментов, аналогичных определенным в стандарте IETF RFC 3986, и функциональных возможностей сервера соответствующего ресурса. Если, например, процесс разрешения идентификаторов PID возвращает URI, содержащий в себе идентификатор фрагмента со следующим URI (в котором фрагмент содержит спецификацию, отвечающую рекомендациям IETF RFC 5147 для носителя с открытым текстом):

http://myserver/myresource#line=10,20,

то такой URI приведет к выборке браузером целостного документа, а уже сам браузер выделит затем часть документа, содержащую строки с 10-й по 20-ю, и представит их пользователю. При малом размере документа такой ход операций приемлем; однако в случае необходимости представления фрагмента из файла, хранимого на носителе емкостью 2 Гб, возникает потребность в использовании специального сервера ресурса и такого URI, который имеет спецификацию разделов, определенную в Annodex [22]:

http://videoserver.com/videoA.anx?t=15.0/30.0

Такая запись будет инициировать передачу сервером ресурса только видеосегмента, записанного в интервале от 15-й до 30-й секунды.

Ожидается, что по мере обновления настоящего стандарта список применимых форматов идентификаторов разделов и фрагментов ресурсов будет расширяться.

5.5.2 Носители данных (по интервалам времени)

Стандарт ИСО/МЭК 21000-17, касающийся ресурсов формата MPEG-21, должен применяться к информационным носителям и временным интервалам. Для определения временных интервалов в запросах URI и фрагментах может использоваться синтаксис Annodex [22] применительно к любым используемым информационным носителям. Для других форматов идентификатор фрагментов ресурса будет зависеть от используемого формата.

5.5.3 Текстуальные ресурсы

К текстовым ресурсам формата XML должна применяться технология XPointer. Для документов с открытым текстом действует стандарт IETF RFC 5147. Для других форматов идентификатор разделов ресурса будет зависеть от используемого формата.

5.5.4 Реестры метаданных, терминология и онтологии

Согласно ИСО 12620:2009, спецификация каждой категории данных в рамках Реестра категорий данных DCR (Data Category Registry), который ведется в ИСО ТК 37, должна иметь собственный PID, формируемый как идентификатор целостного DCR. Эти постоянные идентификаторы конфигурируются как «слабые» URI [30], что является сложившейся практикой для Консорциума Всемирной паутины (World Wide Web Consortium).

Конкретные части графа RDF могут адресоваться с помощью одного из предложенных запросных языков RDF, однако какой-либо определенной спецификации на данный момент нет. Для других форматов идентификатор раздела ресурса будет зависеть от используемого формата.

5.6 Коллекции

Существующие «опубликованные» коллекции должны снабжаться соответствующим PID, который поддерживается организатором коллекции. При этом PID должен обращаться к представляющему коллекцию описанию, которое может быть записью каталога или отдельным описанием метаданных. В случае виртуальных коллекций PID должен обращаться к машиночитаемому описанию метаданных, которое обеспечивает доступ к релевантной информации и, в частности, к ресурсам, имеющим свои собственные PID.

6 Дополнительные требования

6.1 Разбиение идентификаторов

Применительно к разбиению в настоящем стандарте различаются идентификаторы разделов и идентификаторы фрагментов, как показано в 5.5. Идентификатор фрагмента определяется в IETF RFC 3986 как необязательный компонент ссылки URI. Согласно IETF RFC 3986, URI может снабжаться факультативным идентификатором фрагмента, в соответствии с которым он отделяется от

остальной части ссылки URI знаком решетки «#». Сам разделитель не считается частью идентификатора фрагмента.

URI с идентификатором фрагмента может использоваться приложением для идентификации конкретного ресурса и, как правило, последующего доступа к этому ресурсу, который является разделом первичного ресурса или вложен в него. Формат идентификатора фрагмента зависит от типа ресурса.

Интерпретация и разыменование идентификатора фрагмента является функцией веб-клиента и, следовательно, требует загрузки полного первичного ресурса, после чего клиентское приложение получает возможность извлечь запрошенный фрагмент. Кроме того, поскольку идентификатор фрагмента в интервале его извлечения не пересылается другим системам, некоторые «посреднические» модули сетевой архитектуры (как, например, программы-агенты) никак не взаимодействуют с идентификаторами фрагментов, и при переадресации HTTP фрагменты не принимаются в расчет. Исключение составляют идентификаторы фрагментов для RDF-документов, которые содержат ссылки не на их части, а на конкретный объект документа, описанный как объект с идентификатором фрагмента. Благодаря этому использование идентификаторов фрагментов в сочетании с URI позволяет клиентскому приложению выделять разделы ресурса на основе учета специфических сведений клиентского приложения.

В противоположность процедурам разрешения фрагментов использование такого идентификатора раздела, как, например, «z» в ссылке

http://myserver/myObjectService?part=z,

рассчитано на то, что удаленный сервер выделит только часть ресурса и перешлет ее клиентскому приложению[1)].

6.2 Рекомендации

Настоящий стандарт поддерживает различные уровни разбиения ресурсов. Для обеспечения надлежащей эффективности этого процесса и совместимости с другими схемами именования служат приведенные ниже рекомендации.

- Если для какого-то типа ресурсов уже существует схема идентификации, как, например, номер ISBN для книги, этот уровень разбиения должен сохраняться, то есть не следует создавать новые идентификаторы PID без должного обоснования необходимости такого действия (например, для глав книги). Главы должны предпочтительно адресоваться с использованием идентификаторов разделов в сочетании с PID-идентификатором книги.

- Если ресурс ассоциируется со всем информационным содержимым (контентом) цифрового файла, то такому ресурсу, скорее всего, должен присваиваться индивидуальный PID.

- Если ресурс независим и существует вне более широкого контекста, то такому ресурсу должен присваиваться индивидуальный PID.

- Если какой-то ресурс должен цитироваться в отрыве от своего контейнера, то такому ресурсу должен присваиваться индивидуальный PID.

Эти рекомендации, однако, зависят от конкретных потребностей создателей ресурсов, которые сами определяют уровень разбиения, подходящий для их конкретной ресурсной среды.

[1)] Такое четкое разделение ролей между сетевым клиентом и сервером при разыменовании фрагментов может измениться. Рабочая группа по мультимедийным фрагментам разработала проект W3C, в котором предлагается, чтобы сетевые клиенты «договаривались» о транспортировке с сервера только раздела ресурса [27].

Приложение А
(справочное)

Автономные ресурсы, агрегированные ресурсы и разделы ресурсов

А.1 Общий обзор

А.1.1 Общие замечания

В науке и в промышленной сфере все острее ощущается потребность в непротиворечивом и постоянном обращении к электронным языковым ресурсам и их частям, а также к коллекциям таких ресурсов. При этом желательно не только иметь возможность извлечения и контроля нужной информации из научных публикаций, но и поддерживать различные типы перекрестных ссылок между языковыми ресурсами или их частями.

А.1.2 Ресурсы

Ресурсом называется всякая сущность, обладающая собственной индивидуальностью [7]. Несмотря на неопределенность такой характеристики, она важна для исследователей, которые стремятся идентифицировать лингвистически значимые единицы как связанные объекты репозитория для обеспечения удобства считывания их человеком и все чаще — машинным способом. Такие объекты должны участвовать в отдельных манипуляциях и использоваться в различных сценариях, а это означает, что они могут существовать автономно в более пространных контекстах. Зачастую подобный объект может идентифицироваться как «одиночный файл» в файловой системе, однако он может также извлекаться как включенный объект (например, как запись данных) из системы управления базой данных. Такие ресурсы могут относиться к различным типам или иметь разные форматы:

- оцифрованные видеозаписи интервью;
- аудиозапись песни;
- комплексная аннотация сеанса коммуникации;
- фотография, документирующая речевое событие;
- словарь для определенного языка;
- описание грамматики;
- запись траектории движения глаза при исследовании процесса чтения;
- описание метаданных ресурса или коллекции ресурсов;
- интегрированный документ, содержащий тексты и фотографии и др.

Объем такого ресурса оставляется на усмотрение его создателя. Некоторые ресурсы образуются группой аннотаций разного уровня, тогда как другие могут создаваться для каждого уровня аннотирования с учетом тех или иных научных и управленческих аспектов. Например, для многих ресурсов метаданные могут храниться в единственной крупной реляционной базе данных. В таких случаях «описание метаданных» существует только как результат запроса, и не вовлекается никакой идентифицируемый ресурс, поскольку запрашиваемым ресурсом является целостная база данных. Подобный сценарий говорит о необходимости поиска способов адресации разделов такого ресурса.

Когда уникальный и постоянный идентификаторы ассоциируются с ресурсами, требуется, чтобы сами ресурсы тоже оставались неизменными. Рассматриваемый репозиторий предоставляет ресурс с уникальным идентификатором для гарантии того, что при разрешении ссылки будет обслуживаться именно «исходный контент» (здесь не учитываются аспекты пользовательского интерфейса, который может со временем изменяться и выдавать ту же самую содержательную информацию в ином виде, хотя это может влиять на ее интерпретацию). Когда контент ресурса изменяется в результате какой-либо манипуляции, он образует новый ресурс, так как приобретает новую индивидуальность, которая отделена от предыдущей версии ресурса. Динамические базы данных, постоянно меняющие свою индивидуальность именно таким образом, считаются образованными из «снимков мгновенных состояний» (snapshots), которые и должны быть объектами ссылок. Любые зависимости между такими снимками могут выражаться соответствующими метаданными, например специфицируемыми связями, цитируемой информацией или тем и другим вместе.

А.1.3 Связки ресурсов

Иногда в репозиториях содержится указание на то, что некоторые из его ресурсов тесно связаны в соответствии с тем или иным формальным критерием. Примерами могут служить:

- сделанные одновременно аудио- и видеозаписи одинаковой длительности, которые могут включать в себя другие записи, созданные параллельно и снабженные касающимися их аннотациями; все эти ресурсы относятся к одному и тому же временному периоду и сделаны в одном месте, что тесно объединяет их, и пользователи часто хотят получать доступ к таким ресурсам как к единому целому;
- словарь, содержащий фотографии или другие мультимедийные ресурсы, которые расширяют словарные статьи; в этом случае может оказаться желательным обращение к тесно связанным объектам как к единому пакету;
- корпус текстов с соответствующими внешними аннотациями;
- корпус исторических текстов с соответствующими факсимиле.

Как правило, подобные связки должны конфигурироваться как воплощение их группового статуса, обычно представляемого совместным метаинформационным описанием, в котором расставлены указатели на все составляющие ресурсы.

А.1.4 Коллекции ресурсов

Коллекции — это произвольные группировки любого числа ресурсов с индивидуальными идентификаторами, подобранными в соответствии с конкретным критерием, но запрашиваемые тем не менее как единое целое. Примерами таких коллекций могут служить:

- опубликованный корпус речевых оборотов голландского языка (Dutch National Spoken Corpus) или британского варианта английского языка (British National Corpus) по состоянию на определенную дату;
- коллекция всех ресурсов, описывающих конкретный язык, хранимая в определенном репозитории;
- коллекция разрозненных ресурсов, используемая как основа для диссертационного исследования или написания монографии;
- коллекция ресурсов, объединенных с целью сохранения их как сжатого пакета[2)] в электронном архиве;
- коллекции разделов ресурса — например ссылок на ряд сегментов мультимедийного файла и т. п.

Связи между ресурсами коллекции могут быть произвольными. Коллекции имеют свой формальный статус, и их реализации должны быть доступны для запросов, что обычно обеспечивается с помощью метаинформационного описания. Связки ресурсов, рассмотренные в А.1.3, — это коллекции ресурсов, обладающих общими лингвистически значимыми параметрами.

А.1.5 Метаинформационные описания

Каждый ресурс и каждая коллекция ресурсов должны ассоциироваться с метаинформационным описанием, реализующим их существование. Это обеспечивает оптимизацию и долговременную защиту ресурсов; при этом различаются объекты как таковые и множество их копий, которые могут находиться в разных репозиториях. Метаинформационные описания не только служат воплощением ресурсов или коллекций, но еще и хранят дополнительные сведения о них. При необходимости обращения к самим метаинформационным описаниям они тоже должны трактоваться как объекты, наравне с другими ресурсами. Однако часто записи метаданных порождаются в динамическом режиме как результат выполнения запроса к комплексной базе данных, а это означает, что метаинформационное описание является не ресурсом, а разделом ресурса. Настоящий стандарт учитывает оба сценария развития событий, поскольку PID может указывать либо на ресурсы, либо на различные разделы ресурсов. Сервер ресурса не должен конфигурироваться на генерацию одной и той же информации в обоих указанных случаях.

В случае коллекций и связок ресурсов метаинформационное описание как таковое содержит ссылки на включенные в коллекцию ресурсы, и процесс разрешения этих ссылок может быть рекурсивным. Настоящий стандарт не нацелен на определение какого-то конкретного способа представления ресурсов в рамках метаинформационного описания. Однако он четко определяет механизм, который должен использоваться в таких описаниях для запроса ресурсов с помощью PID-идентификаторов.

Нет также и какой-то строго установленной процедуры контроля различных версий метаданных. В некоторых репозиториях исходный номер версии вообще не меняется; в других этот номер меняется при изменении дополнительной информации; бывает и так, что номер версии метаданных изменяется, когда меняется объект, к которому эти данные относятся. Настоящий стандарт не преследует цели установления какого-то единого подхода к управлению версиями метаинформационных описаний.

А.1.6 Аспекты разбиения ресурсов

Различия между разделами ресурса, самим ресурсом и конкретной коллекцией ресурсов порождают множество актуальных проблем, связанных с определением целесообразных уровней разбиения репозитория, которые должна поддерживать организация, ответственная за архивирование данных. Существует множество разных подходов к решению этих проблем, и все они могут отражать специфику конкретных потребностей и рабочей среды. Поэтому настоящий стандарт не определяет каких-либо жестких правил разбиения, а только дает некоторые руководящие принципы выделения уровней PID, присваиваемых коллекциям ресурсов, ресурсам и разделам ресурсов, с учетом конкретных целей обеспечения высокой производительности, совместимости, эффективности и управляемости.

В рамках многих проектов разработчикам приходится решать, какие из объектов подходят для присвоения им собственных индивидуальных PID-идентификаторов, а какие могут обрабатываться с использованием идентификатора раздела в сочетании с PID более крупного ресурса-контейнера. Здесь важны аспекты производительности и управляемости, и хотя инфраструктуры PID предназначаются для работы с множественными идентификаторами, каждый идентификатор все равно требует каких-то операций технической поддержки и обработки, вследствие чего достижениями сложившейся практики диктуется резервирование индивидуальных PID для использования в тех ситуациях, когда они действительно полезны. Существенную роль играют также соображения логической связности и совместимости, и потому объекты одного типа должны рассматриваться по возможности на одном и том же уровне разбиения.

[2)] Различные типы пакетов, которые могут использоваться для организации хранения собранных ресурсов как идентифицируемых объектов архива, обсуждаются в модели OAIS. При этом решение по контенту пакета принимается его создателем на основе какого-либо содержательного или управленческого критерия. Например, стандарт METS [20] обеспечивает именно такую идентификацию пакетов.

Для некоторых типов ресурсов существуют действующие и предлагаемые стандарты, указывающие способ адресации разделов интернет-ресурсов, основанный на использовании идентификатора фрагмента, присоединяемого к URI. В разделе 6 показано, что при большом объеме ресурса этот подход не эффективен; тем не менее само наличие такой функциональной возможности обеспечивает мощный механизм адресации, при наличии которого в документах могут поддерживаться активируемые идентификаторы.

А.2 Общие практические рекомендации по запрашиванию интернет-ресурсов

В архитектуре сети Интернет унифицированный идентификатор ресурса URI (Uniform Resource Identifier) определяется как последовательность символов, используемая для обращения к сетевым ресурсам. Унифицированные идентификаторы ресурсов разделяются на множество подклассов или схем: многие схемы URI определяют механизм доступа к ресурсам и потому имеют второе (неофициальное) название «унифицированные указатели ресурсов» [Uniform Resource Locators (URL)], поскольку они указывают местонахождение ресурса в сети [5]. Использование URI — наиболее распространенный способ вызова ресурсов, хотя бы потому, что при вложении в веб-документ URI становится непосредственно активируемым объектом. Синтаксическая структура URI зависит от конкретной используемой схемы (http, ftp, file, gopher и др.). Самая популярная схема HTTP, определяемая стандартом IETF RFC 2616 [8], поддерживает следующие функциональные возможности, важные для настоящего стандарта:

a) возможность добавления параметров в URI в рамках запроса для передачи дополнительной информации удаленному веб-серверу. Например, в URI вида http://server/service?part=12 параметры интерпретируются удаленным сервером ресурса;

b) возможность добавления идентификатора фрагмента с целью определения запрашиваемого раздела ресурса. Например, в URI вида http://server/document1#part1 идентификатор фрагмента не пересылается на удаленный сервер, а используется клиентским приложением для выделения требуемого фрагмента документа из полностью передаваемого ресурса.

Довольно часто данные о размещении ресурса в сети и о локальном пути доступа к нему вкладываются в URI ресурсов[3)], что порождает целый ряд проблем при изменении местонахождения ресурса. «Подмешивание» в идентификатор семантики указателя временного местоположения опасно. Например, защищенный на какой-то момент времени ресурс

http://myhost.ourdomain/protected/R1.wav

может на следующий день стать незащищенным, а со временем может измениться даже семантическая структура имени ресурса.

Схема URN URI, определенная в стандарте IETF RFC 2141 [21], наоборот, предоставляет имена для ресурсов вместо их адресации. Унифицированное имя ресурса URN (Uniform Resource Name) несет в себе идентификатор пространства имен из списка, который ведет IANA [13]; этот идентификатор обеспечивает бесконфликтную интеграцию данных из других схем именования (например, ISBN, ISSN) как подмножеств URN. Схемы URN должны предоставлять глобально уникальный постоянный идентификатор, используемый для идентификации ресурсов и доступа к ним.

URN имеет следующую синтаксическую структуру:

urn:<идентификатор пространства имен (NID)>:<конкретная строка данных пространства имен (NSS)>

Строка данных пространства имен может принимать любую форму, определяемую администрацией службы имен, при условии, что имя будет уникальным в этом пространстве и согласно IETF RFC 2141 не будет ограничено малым числом разрешенных символов. Хотя идентификатор URN очень хорошо подходит для именования ресурсов, для него нет никакого общепринятого механизма разрешения имени к адресу местоположения ресурса, что стало поводом к появлению различных систем постоянных идентификаторов, рассматриваемых ниже, в приложении В.

Консорциум W3C, осознавая трудности вложения адресной информации в идентификаторы URI, опубликовал результаты изысканий Группы по технической архитектуре [Technical Architecture Group (TAG)] в этой области [25, 30]. Полученные этой группой результаты содержат такие важные ссылки, как документ Бернерса—Ли (Berners—Lee) по идентификаторам URI, и показывают, что конструирование неизменных URI возможно и должно быть предпочтительным по отношению к другим техническим решениям и, в частности, к инфраструктурным PID.

А.3 Системы разрешения постоянных идентификаторов

Постоянные идентификаторы, ассоциируемые со службой разрешения ссылок, предназначены для решения общей проблемы разорванных связей, которая возникает в тех случаях, когда сетевой ресурс переносится в другое место или вообще удаляется. Для работы с неизменными идентификаторами было предложено много разных подходов, нацеленных как на предоставление непротиворечивых схем именования оперативно доступных ресурсов, так и на создание службы разрешения ссылок для переадресации пользователей к текущему местоположению запрошенного ресурса с помощью неизменного идентификатора.

[3)] Это может показаться неудачным практическим результатом, так как большинство http-серверов допускает манипулирование унифицированными идентификаторами ресурсов независимо от конкретного местоположения ресурса. Однако в действительности бывает очень трудно поддерживать эту независимость без применения механизмов переадресации с аналогичными функциями администрирования, как, например, в системе PURL, которая рассматривается в разделе В.1.

Свойство неизменности PID-идентификатора означает, что он никогда не может быть «переназначен» ни для какого иного ресурса и не должен изменяться ни при каком перемещении данного ресурса или изменении протокола доступа к нему. Известно множество формальных схем присвоения идентификаторов и имен; эти схемы уже обсуждались выше в контексте именования электронных ресурсов (например, URI, URN, handles, DOI, ARK, ISBN, ISSN, SICI, BICI, PIIs). Однако применительно к упрощению доступа к сетевым ресурсам в распределенной системе лишь очень немногие из существующих схем могут быть действительно эффективными, если они официально не зарегистрированы в качестве схем именования URI и не поддерживаются системой разрешения идентификаторов или не встроены в другую схему присвоения имен, которая оснащена соответствующей системой разрешения.

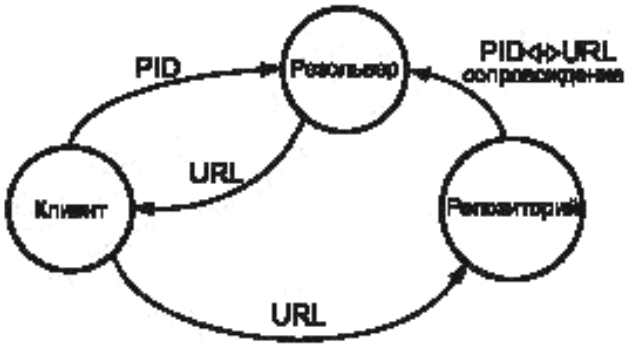

Рисунок А.1 — Схема разрешения PID-идентификатора

Без системы разрешения запросы, в которых используется идентификатор, не могут быть направлены нужному серверу для последующего извлечения запрошенного ресурса или его правомерного «заменителя» в виде метаданных (см. рисунок А.1). Краткие описания большинства используемых систем разрешения PID-идентификаторов можно найти в приложении В.

А.4 Адресация разделов ресурса

А.4.1 Общие замечания

Ниже описываются сложившаяся практика и существующие стандарты адресации разделов ресурсов. Некоторые стандарты и практические методы используются в действующих соглашениях о цитировании и в рамках сетевой архитектуры. Основной предметной областью настоящего стандарта являются интерпретируемые машинным способом идентификаторы, которые однозначно указывают требуемую часть или сегмент более крупного ресурса. Такие идентификаторы способны переключать удаленные серверы ресурсов на выдачу некоторого представления запрошенной части ресурса клиентскому приложению или «заставлять» его самостоятельно выделять из ресурса нужный раздел. Применительно ко многим типам ресурсов, для которых нет действующих стандартов или практических рекомендаций, возможно внесение соответствующих поправок в уже существующие стандарты и рекомендации.

А.4.2 Мультимедийные средства (в порядке появления)

В настоящее время существуют следующие практические рекомендации и стандарты, применимые к мультимедийным средствам:

ИСО/МЭК 21000-17 Информационные технологии. Основы мультимедиа (MPEG-21). Часть 17. Идентификация фрагментов ресурсов MPEG определяет нормативную синтаксическую структуру идентификаторов фрагментов ресурсов URI, которые подлежат использованию при адресации разделов любого интернет-ресурса, относящегося к одному из следующих типов сред: audio/mpeg, video/mpeg, video/mp4, audio/mp4, application/mp4, video/MPEG4-visual, application/mp21.

Цифровой формат Annodex, разработанный австралийским государственным объединением научных и прикладных исследований [Commonwealth Scientific and Industrial Research Organisation (CSIRO)], включает в себя еще и спецификацию временно́й адресации с использованием специального синтаксиса URI [22]. Этот синтаксис позволяет обращаться к сегментам аудио- и видеофайлов. Синтаксическая структура такого URI подлежит интерпретации удаленным сервером, который выделяет запрошенный сегмент из мультимедийного файла и пересылает его клиенту. Например, URI, имеющий вид

http://example.com/video.anx?t=15.2/18.7,

будет вызывать для передачи сегмент записи *video.anx*, начинающийся в точке 15,2 секунды и заканчивающийся в точке 18,7 секунды.

А.4.3 Текстовые ресурсы

В случае текстовых ресурсов конкретные разделы могут адресоваться идентификаторами разделов до тех пор, пока они специфицируются поставщиком ресурса. Кроме того, если ресурс текстовых корпусов состоит из пре-

жде опубликованных текстов и по ним доступны соответствующие метаданные, то такие тексты и их части при необходимости тоже будут адресоваться с использованием обычных соглашений по косвенному цитированию.

К числу практических рекомендаций по адресации фрагментов текстовых файлов относятся следующие материалы:

- по использованию байтовых или символьных сдвигов для указания позиций или «пропусков» в тексте имеется релевантная технология: Типстера [26], GATE [10] и всевозможные производные методы от GATE;

- IETF RFC 5147, URI Fragment Identifiers for the text/plain Media Type (Идентификаторы URI для текстовых и открытых мультимедийных фрагментов), основывающиеся на использовании строковых или символьных сдвигов либо тех и других одновременно;

- для документов XML, HTML, XPointer имеется, например, технология Annotea [2], которая представляет собой проект, поддерживаемый консорциумом W3C и предназначенный для расширения сотрудничества в области электронных документов на основе использования тегов, закладок и прочих атрибутов аннотирования.

А.4.4 Источники знаний

К источникам знаний относятся словари, терминологии, реестры основных понятий, онтологии и т. п. Эта предметная область отличается динамичными разработками, в том числе в части форматов представления информации. В задачу настоящего раздела не входит исчерпывающее описание современного состояния разработок; его основная цель — показать лишь некоторые возникающие проблемы. В настоящем стандарте не ставится также задача предложить какие-то конкретные решения, поскольку выбор способов доступа к тем или иным разделам имеющихся ресурсов, как и гарантирование живучести используемых механизмов адресации — это сфера ответственности конкретных профессиональных сообществ, которые могут остановить свой выбор, например, на методах, проверенных практикой. Таким методом может быть использование постоянных унифицированных указателей ресурсов (PURL), как в стандарте «Дублинское ядро».

Что касается словарей, то в этой области ИСО 24613:2008 [15] предлагается метамодель для представления информации в лексических базах данных. Необходимо, чтобы эта информация была доступной на разных уровнях для внешних ресурсов или инструментальных средств, которые обеспечивают адресацию отдельных лексических единиц, смысловых значений или даже более мелких информационных объектов, таких как хранимые в словаре морфологические части слов. Когда словарь представляется в формате XML, для описания ссылок на него может использоваться расширяемая спецификация XPointer. Сам словарь может тоже содержать ссылки на внешние ресурсы, такие как звук или видео, для иллюстрации произношения или примеров использования лексических единиц. Такие ссылки, как правило, относятся не только к ресурсу в целом, но и к его частям, таким как видеофрагменты. В случае обращения к части ресурса его идентификатор должен быть дополнен специализированным указателем раздела ресурса, предназначенным для обеспечения доступа к нужному фрагменту информации. Наконец, лексиконы LMF, как и все другие модели, охватываемые семейством стандартов технического комитета ИСО/ТК 37, обеспечивают ссылки на элементы Реестра категорий данных DCR (Data Category Registry), определенного ИСО 12620.

Такие инструментальные средства, как LEXUS [17] (где реализована схема лексической разметки), работают в среде мультимедийных словарей, в которых ресурсы различных типов оказываются тесно переплетенными друг с другом. В таких случаях необходимо иметь механизм непротиворечивой адресации для гарантии совместимости ресурсов.

Применительно к терминологиям и реестрам основных понятий, таким как ISO DCR, должна существовать возможность обращения к отдельным статьям. Базовый уровень совместимости может быть обеспечен путем выборки статей, которые считаются идентичными, и объявления их таковыми.

Если отдельный ресурс может адресоваться с использованием некоторого унифицированного механизма, такого как PID-идентификатор, то идентификаторы разделов ресурса в значительной степени зависят от типа запрашиваемого ресурса. Для некоторых типов ресурсов появляются стандарты, в которых рассматривается эта проблема, однако существуют типы, для которых ситуация остается в значительной мере неясной, что иллюстрируется следующим примером.

Пример — Применительно к базам знаний, которые содержат в дополнение к определениям концептов еще и отношения между ними, для представления информации все чаще используется формат RDF. В таких случаях для обращения к графу или подграфу должен применяться идентификатор раздела ресурса. Для этой цели обсуждались различные технические решения, в том числе стандартизованные запросы, сформулированные на языке запросов RDF [12].

А.5 Адресация коллекций ресурсов

Проблема адресации коллекций ресурсов, относящихся к классу «публикуемых», хорошо известна, однако до сих пор нет ясности в отношении сложившихся практических методов, а используемые механизмы доступа не всегда пригодны для машинной реализации и удобной интерпретации. Обычно пользователи работают с публикациями текстовых лингвистических корпусов так, как описано в документе Пенсильванского университета, касающемся корпуса американского варианта разговорного английского языка:

John W. Du Bois, Santa Barbara Corpus of Spoken American English. Parts 1, 2 and 3, Linguistic Data Consortium, University of Pennsylvania, Philadelphia (2000),

либо обращаются непосредственно к оригиналам, представленным в той или иной форме, как, например, коллекция Калифорнийского университета:

Santa Barbara Corpus of Spoken American English. Part I. 2000. Collected by University of California, Santa Barbara Center of the Study of Discourse, directed by John W. Du Bois.

Подобные ссылки осуществимы лишь в тех случаях, когда четко определена ответственность за процесс создания корпусов в части авторства участвующих творческих организаций, архивных учреждений и т. п. Однако, даже при наличии всех этих атрибутов, желательно иметь прямой доступ к машиночитаемому PID-идентификатору, который мог бы использоваться далее, например для автоматического извлечения из формализованного метаинформационного описания сведений о размерах корпуса или для получения какой-либо другой информации.

Не существует сложившейся практики и для виртуальных коллекций, которые создаются исследователями в процессе выполнения конкретных проектов и могут содержать богатый набор информационных ресурсов, порождаемых различными творческими коллективами многочисленных организаций и учреждений; при этом, однако, необходимо документирование выполняемой работы в такой форме, чтобы другие специалисты имели возможность проверить правильность утверждений авторов. Причина отсутствия такого документирования кроется в том, что лишь в редких случаях (как в репозиториях, основанных на стандарте метаданных IMDI для описания мультимедийных ресурсов [28]) исследователи могут создавать метаинформационные описания виртуальных коллекций, играющие роль их реализаций, например для инициирования операций поиска.

А.6 Цитируемая информация

А.6.1 Общие положения

Необходимо понимать различия между адресуемыми текстами разного характера. Применительно к некоторым контекстам, таким как «картинки» в словаре, запрашивающую сторону обычно не интересует никакая информация об изображении в дополнение к тем сведениям, которые уже были найдены в тексте словарной статьи. Инициатора запроса будет интересовать только активация ссылки и мгновенное получение самого изображения. В других случаях, когда полученная по ссылке информация предназначается для чтения человеком, более подходящим, может оказаться представление пользователю цитируемой информации в полном объеме, как она выглядела бы в традиционной распечатке. В случае виртуальных коллекций пользователю может понадобиться увидеть более подробные сведения о ресурсах, включенных в коллекцию, и, следовательно, он может запросить метаинформационное описание. В первом и последнем случаях ссылки представляют собой PID-идентификаторы с присоединенным, при необходимости, указателем фрагмента. Второй из рассмотренных выше случаев требует более глубокого рассмотрения цитируемой информации.

В научных работах цитирование символизирует признание важности работ предшественников и выражение доверия к ним. Читатели могут при этом проверить свои утверждения, сравнив их с результатами цитируемых работ. То же самое можно сказать и о первичных научных данных, на базе которых авторы строят свои исследования и желают выразить свою признательность создателю первоисточника или составителю коллекции. Поэтому при цитировании электронного ресурса в каком-либо документе запрос такого ресурса или ссылка на него сопровождается выдачей цитируемой информации: текста (метаинформации) для прочтения человеком, указывающего держателей ресурса (создателя, составителя коллекции, издателя и др.).

Существует множество стандартов и практических рекомендаций, касающихся определения цитируемой информации, и перечислять их здесь нецелесообразно, но некоторые широко используемые документы приведены в А.6.2—А.6.6.

А.6.2 ИСО 690:2010

В ИСО 690:2010 [16] представлен обзор многочисленных типов электронных документов и определены элементы данных, подлежащие включению в библиографические ссылки. Это важный источник информации в части сложившейся практики цитирования. ИСО 690 в основном касается примеров использования идентификаторов DOI, но допускает и применение PID-идентификаторов. В отличие от ИСО 690, который не выдвигает никаких требований к инфраструктуре PID, настоящий стандарт устанавливает такие требования. Представленная в нем инфраструктура PID позволяет, например, осуществлять более эффективный контроль над дополнительной информацией (метаданными) с такими идентификаторами и обеспечивает возможность использования других бизнес-моделей, отличных от DOI. Все рекомендации, содержащиеся в ИСО 690, могут быть реализованы при условии, что допускаются к использованию PID-идентификаторы из любых инфраструктур, которые удовлетворяют требованиям настоящего стандарта.

Для оперативно доступных ресурсов важно присутствие следующих элементов:

- «Готовность и способ доступа» (например, «Доступен по адресу: URL» или «Доступен также в формате PDF по адресу: URL»);
- «Стандартный номер» (например, «ISBN 0-7710-1932-7» или «ISSN 1045-1064»);
- «Нумерация и пагинация» (например, «Инвентарный номер 01209277», «страницы 5—21» или «строки 30—40»).

Значением элемента «Готовность и способ доступа» обычно является URL.

Для использования рекомендаций ИСО 690 в контексте настоящего стандарта потребуется замена URL на сам PID при одновременном четком определении типа схемы разрешения идентификаторов или использование идентификатора, закодированного как URI. Однако до тех пор, пока синтаксис PID не будет официально утвержден консорциумом W3C как инструмент, пригодный для идентификации веб-ресурсов, наряду с ним должен присутствовать и вариант PID, закодированный как URI. Например, применительно к Handle System должны существовать следующие ссылки:

доступен по адресу: hdl:4263537/4069, http://hdl.handle.net/4263537/4086

или

доступен по адресу: http://hdl.handle.net/4263537/4086.

В качестве элемента «Стандартный номер» возможно использование самого PID или (при наличии подходящего пространства имен URN) составного специфицированного URN наподобие записи: urn:doi:10.1392/BC1.0. (Предполагается наличие запроса пространства имен этого URN).

Элементы «Нумерация и пагинация» могут принимать значения идентификатора раздела ресурса или идентификатора фрагмента.

А.6.3 Руководство APA по стилистическому оформлению

Американская психологическая ассоциация (АПА) определяет многие реквизиты рукописных документов и официальной документации, равно как и организацию процедур цитирования и указания ссылок. Документ этой ассоциации под названием «APA Style Guide to Electronic References» [13] («Руководство АПА по стилю оформления ссылок на интернет-ресурсы») определяет два соответствующих элемента для оперативно доступных ресурсов: «получен» и «откуда». Назначение первого элемента состоит в указании даты извлечения материала с целью использования, а второго — в определении источника (URL) документа, возможно с предваряющим описанием URL.

Пример

Роджерс, Б. (2078). Путешествие быстрее света: чему мы научились в первые двадцать лет. Получено 24 августа 2079 г., с веб-сайта Института по исследованиям Марса при Марсианском университете, http://www.eg.spacecentraltoday.mars/university/dept.html

Руководство АПА по стилю оформления ссылок явным образом устанавливает, что при наличии неизменного идентификатора, например такого как DOI, в качестве источника вместо URL должен использоваться PID-идентификатор.

А.6.4 Справочник MLA

Справочник Американской ассоциации современного языка (MLA) [9] является руководством по научному стилю, в котором представлены руководящие принципы оформления статей и документации по исследованиям в области гуманитарных наук.

В седьмом издании этого справочника использование URL применительно к оперативно доступным ресурсам характеризуется как необязательное. Однако если URL все же требуется или ресурс не может быть найден иным способом, то URL может добавляться к ссылке после указания автора, заголовка, издателя и даты запроса.

Пример — Запрос конкретной веб-страницы:

Библиотека Корнельского университета. «Введение в научные исследования». Библиотека Корнельского университета. Корнельский университет, 2009, Web. 19 июня 2009 <http://www.library.cornell.edu/resrch/intro>.

Процедура замены URL постоянным идентификатором совместима со стилем MLA.

А.6.5 Проекты STD-DOI и DataCite

Проект под названием Publication and Citation of Scientific Primary Data (STD-DOI) [6] («Публикация и цитирование научных первоисточников») был основан Немецким научным фондом (German Science Foundation). Целью этого проекта было обеспечение цитируемости первичной информации научных исследований подобно публикациям. В системе STD-DOI совокупность научных данных снабжается атрибутом в форме фамилий исследователей в качестве авторов, как это принято для научных работ, цитируемых в обычной научной литературе. В таком представлении первичные научные данные не раскрываются по своей сути как в научной публикации, однако могут обрести свою индивидуальность.

Пример

Kamm, H; Machon, L; Donner, S (2004): Gas Chromatography (KTB Field Lab), GFZ Potsdam. doi:10.1594/GFZ/ICDP/KTB/ktb-geoch-gaschr-p

На смену проекту STD-DOI в настоящее время пришел проект DataCite (http://www.datacite.org) консорциума преимущественно технических библиотек и информационных центров, который ставит перед собой те же цели и задачи, что и STD-DOI.

А.6.6 Проект стандарта цитирования количественных данных научных исследований

В рамках недавно предложенного проекта стандарта цитирования количественной информации данных в области социальных наук [1] выдвигается требование, чтобы в цитируемых материалах использовался постоянный идентификатор. При этом предлагается указывать в атрибутах цитирования не менее шести обязательных (мета-

информационных) компонентов. Три первые составляющие — это традиционные данные по автору, дате и заголовку; за ними должны следовать постоянный идентификатор, унифицированный числовой код, предназначенный для проверки целостности первичных данных, и так называемый «служебный мост», который позволяет преобразовывать PID, закодированный как URI, в активируемый идентификатор.

Пример

Micah Altman; Karin MacDonald; Michael P. McDonald, 2005, «Replication data for: From Crayons to Computers: The Evolution of Computer Use in Redistricting», hdl:1902.1/AMXGCNKCLU UNF:3:J0PkMygLPflyT1E/8xO/EA== = http://id.thedata.org/hdl%3A1902.1%2FAMXGCNKCLU

Приложение В
(справочное)

Реализации системы постоянных идентификаторов

В.1 Постоянный URL (PURL)

Постоянный URL или PURL (Persistent URL) [23] представляет собой унифицированный указатель информационного ресурса (то есть локальный идентификатор ресурса URI), который не описывает напрямую местоположение запрашиваемого ресурса, а вместо этого указывает лишь его промежуточное (более устойчивое) местонахождение, обращение к которому инициирует переадресацию идентификатора к текущему местоположению нужного ресурса. Это стандартизованная процедура переадресации по протоколу HTTP, благодаря которой не требуется принятия новых протоколов передачи или внесения изменений в клиентское программное обеспечение (ПО).

Указатели PURL были разработаны Центром оперативно доступных компьютеризованных библиотек [Online Computer Library Centre (OCLC)] в середине 1990-х годов, главным образом для того, чтобы уменьшить бремя эксплуатационных затрат по поддержанию URL, содержащихся в записях каталога, которые относятся к интернет-ресурсам. Центр OCLC стал активным участником рабочих групп IETF по URN, и его специалисты получили полную информацию о том, как далеки были эти группы от общего согласия по вопросу стандартизации постоянных идентификаторов. Поэтому разработанные Центром постоянные указатели PURL были приняты в качестве промежуточного решения для «закрытия бреши» в схеме присваивания постоянных имен ресурсам сети Интернет.

Инфраструктура PURL позволяет лишь ассоциировать одно конкретное местоположение ресурса с определенным идентификатором и не предусматривает никакой дополнительной информации, относящейся к метаданным. Резольвер и управляющее ПО доступны для широкого использования.

В.2 Система дескрипторов Handle System (HS)

Handle System (HS) [24] — это распределенная система присваивания постоянных имен, предназначенная для прикладных систем электронных библиотек. Система была разработана Корпорацией по национальным инициативным проектам [Corporation for National Research Initiatives (CNRI)] и начала свою деятельность с проекта в области представления научно-технических отчетов Networked Computer Science Technical Reports Library (NCSTRL), учрежденного Управлением перспективных исследований министерства обороны США [Defense Advanced Research Projects Agency (DARPA). Одной из задач этого проекта стала разработка принципов построения распределенной электронной библиотечной системы с открытой архитектурой.

Система HS широко известна в мире электронных библиотек. Она предоставляет спецификацию синтаксической структуры PID-идентификаторов и реализацию схемы их разрешения. Синтаксис идентификаторов PID и дескрипторов Handle System чрезвычайно проста:

<префикс>/<суффикс> (например: 15.12345/abcd6789)

Служба высшего уровня Handle System присваивает по запросам префиксы учреждениям и организациям и потому всегда в состоянии разрешить любой присвоенный дескриптор; а это означает, что глобальный реестр дескрипторов (Global Handle Registry) позволит идентифицировать префиксы, префиксы буду указывать на локальные службы дескрипторов, а эти службы будут знать, каким образом следует интерпретировать суффиксы. Синтаксическая структура суффиксов оставляется на усмотрение владельца локальной системы обработки идентификаторов до тех пор, пока она будет соответствовать спецификациям URI. Ниже приведены примеры, взятые с реальных веб-страниц DOI:

10.1000/123456, 10.1000/ISBN1-900512-44-0
10.2345/S1384107697000225
10.4567/0361-9230(1997)42:<OaEoSR>2.0.TX;2-B

Handle System — это нечто большее, чем простая система именования; она поддерживается системой разрешения ссылок, которая образована распределенной системой глобальных, локальных и кэширующих серверов. Глобальный реестр дескрипторов, поддерживаемый CNRI, регистрирует службы именования верхнего уровня с целью обеспечения как уникальности имен, так и маршрутизации запросов для разрешения дескрипторов. Эта процедура уникальна для служб дескрипторов лишь тем, что обеспечивает координацию именующих инстанций, которые целиком и полностью управляются как дескрипторы. Дескриптор службы именования — дескриптор специального вида, который предоставляет информацию, подлежащую использованию клиентами для вызова и последующего использования локальной службы дескрипторов применительно к дескрипторам, поддерживаемым конкретной службой именования. Локальные службы дескрипторов контролируются организациями. В их задачи входят разрешение адресованных им запросов и возврат текущих адресов (одного или нескольких) или иной информации, относящейся к ис-

комому ресурсу. Следовательно, они хранят дескрипторы, которые предоставляют информацию о ресурсах, зарегистрированных соответствующей службой именования. Служба локальных дескрипторов сама может состоять из нескольких серверов. Наконец, кэширующие серверы, ассоциируемые с локальными серверами, позволяют разрешать часто используемые дескрипторы без обращений к глобальному реестру.

Хотя система HS не определяет пространство имен URN, она может использоваться для реализации функций разрешения ссылок в схеме URN, как это делается в рамках DOI [14].

Программные средства HS общедоступны и могут быть скачаны на сайте CNRI Handle. В свободном доступе на этом сайте находятся ПО локальных служб, клиентское ПО, а также простые инструментальные средства управления, кэширующий сервер дескрипторов, средства создания и администрирования дескрипторов и служб именования, и прокси-сервер, обеспечивающий веб-клиентам возможность разрешения дескрипторов. Для предоставления браузерам возможности разрешения дескрипторов без использования прокси-сервера корпорация CNRI разработала специальную программную вставку резольвера дескрипторов (Handle Resolver), которая доступна для скачивания.

Система HS допускает ассоциирование дескриптора с множественными записями типа URL (особенно URI в кодировке UTF8), которые определяют местоположение объекта, идентифицируемого дескриптором. HS поддерживает также пользовательские типы данных, которые могут применяться для ассоциирования метаданных с ресурсом. Хотя в принципе в HS может использоваться протокол HTTP, например с применением промежуточных резольверов, внутренне эта система никак от него не зависит, так как определяет свой собственный протокол.

В настоящее время HS не поддерживает прямое использование идентификаторов разделов ресурса или фрагментов данных, и возможно только кодирование идентификатора раздела как части дескриптора, преобразованного в URI, что, конечно же, сразу делает систему HS зависящей от протокола HTTP.

В.3 Ключ архивного ресурса (ARK)

Одной из самых последних предложенных схем идентификации является схема ключа архивного ресурса ARK (Archival Resource Key), которая разработана Джоном Кюнце (John Kunze) для Национальной медицинской библиотеки США [National Library of Medicine (NLM)]. Схема ARK пока еще имеет статус проектируемого интернет-ресурса, последняя версия которого была опубликована в июле 2008 года [19]. Сейчас она проходит тестирование и реализуется Калифорнийской электронной библиотекой [California Digital Library (CDL)] применительно к коллекциям, находящимся в ее владении. Ключ архивного ресурса (ARK) — это технологическая схема, призванная упростить процедуры присваивания постоянных имен ресурсам и извлечения информационных объектов; она разрабатывается специально для удовлетворения потребностей организаций, связанных с хранением и обслуживанием архивных объектов в цифровой форме.

Система ARK обладает целым рядом важных функциональных возможностей:

- Идентификатор ARK ассоциируется с тремя службами:
  - предоставления ссылки на ресурс;
  - предоставления ссылки на метаданные о ресурсе и
  - предоставления ссылки на соглашение с поставщиком ресурса о постоянстве идентификатора.
- Схема именования не предусматривает семантических правил для идентификатора.
- Система ARK содержит в себе идентификаторы разделов ресурсов и возможные варианты представления различных ресурсов.
- Для активации ARK не требует преобразования к URI, а это значит, что ARK может выполнять роль URI.

Система ARK имеет пять компонентов:

[http://NMAH/]ark:/NAAN/Name[Qualifier]

Такими компонентами являются: необязательный и изменяемый главный порт управления отображением имен NMAH (Name Mapping Authority Hostport); метка «ark:»; номер организации, ответственной за именование NAAN (Name Assigning Authority Number), присвоенное имя (Name) и факультативный (возможно, изменяемый) спецификатор (Qualifier), поддерживаемый ответственной организацией (NMA). Компоненты NAAN и Name совместно образуют постоянный идентификатор объекта. Спецификатор может использоваться для обращения к отдельным частям объекта, к его конкретному представлению или для обеих этих целей. Компонент NMAH указывает службу разрешения ссылок.

Несмотря на то, что резольверы ARK уже введены в действие, в настоящее время общедоступное ПО для системы ARK пока отсутствует. В некоторых отчетах оно описывается как находящееся в стадии разработки.

В.4 Другие системы постоянных идентификаторов

Системы PURL, Handle System и ARK никак не могут считаться образующими полный список инфраструктур постоянных идентификаторов, но они являются хорошими примерами существующих подходов к решению проблемы.

В 2005 году была предложена схема PID-идентификаторов и протокол их разрешения в рамках проекта расширяемого идентификатора ресурсов XRI (eXtensible Pecypc Identifier) [29], разработанного промышленным консорциумом OASIS (Organization for the Advancement of Structured Information Standards). В противоположность системе ARK, идентификатор XRI способствует, например, применению в нем семантических структур, которые конструируются как ряд самоописуемых тегов. В нем также используются специальные символы глобального характера для

индикации неизменяемости и семантического контекста частей идентификатора. Кроме того, в рамках идентификатора XRI допускается слияние идентификаторов из других схем перманентной идентификации.

Сам консорциум W3C не стимулирует применение каких-либо систем разрешения PID-идентификаторов, кроме механизма переадресации с использованием протокола HTTP, как это делается в технологии PURL. Существует проект W3C TAG [25], высвечивающий точку зрения W3C, согласно которой аккуратное применение идентификаторов URI (без использования информации о физическом пути доступа) в конечном итоге должно привести к реальной потребности в использовании PID-идентификаторов во избежание слишком частого возникновения неразрешенных («висячих») ссылок.

Приложение С
(справочное)

Сокращения

Таблица С.1 — Терминологические акронимы

Акроним	Расшифровка сокращения/перевод	Комментарий
APA style	American Psychological Association Американская психологическая ассоциация	Руководство по стилистическому оформлению цитат
ARK	Archival Resource Key Ключ архивного ресурса	Тип инфраструктуры PID-идентификаторов
BICI	Book Item and Component Identifier Идентификатор книги и ее компонентов	Уникальный идентификатор компонентов публикации, в рамках номера ISBN
DARPA	Defense Advanced Research Projects Agency Управление перспективных исследований Министерства обороны	Военная научно-исследовательская организация США
DataCite	Publication and Citation of Scientific Primary Data Публикация и цитирование первичных научных данных	Механизм цитирования первичных научных данных
DCR	Data Category Registry Реестр категорий данных	Формализованный набор лингвистических категорий для ссылок и использования при аннотировании информационных ресурсов
DNS	Domain Name System Доменная система имен	
DOI	Digital Object Identifier Цифровой идентификатор объекта	Постоянный идентификатор оперативно доступного объекта; инфраструктура PID, построенная на основе системы дескрипторов
FTP	File Transfer Protocol Протокол передачи файлов	Технология передачи файлов по протоколу TCP/IP
GATE	General Architecture for Text Engineering Общая архитектура для представления текстов	Технологический инструментарий языка
GOPHER	[не акроним]	Предшественник протокола HTTP
HTTP	Hypertext Transfer Protocol Протокол передачи гипертекстовых данных	Протокол передачи HTML-страниц в сети Интернет
HS	Handle System Система дескрипторов	Тип инфраструктуры PID-идентификаторов
IANA	Internet Assigned Numbers Authority Комитет по цифровым интернет-адресам	Объект, управляющий IP-доменами и зонами DNS
IETF	Internet Engineering Task Force Целевая инженерная группа по развитию Интернета	Группа, разрабатывающая стандарты сети Интернет
IMDI	ISLE Metadata Initiative Программа развития исследований в области метаданных для международной стандартизации разработки языков	Стандарт метаданных на языке XML для формирования языковых ресурсов

Продолжение таблицы С.1

Акроним	Расшифровка сокращения/перевод	Комментарий
IP	Internet Protocol Интернет-протокол	Протокол сети Интернет, IP-протокол
ISBN	International Standard Book Number Стандартный международный номер книги	Уникальный десятизначный номер, используемый для идентификации книг по ИСО 2108 [31]
ISSN	International Standard Serial Number Международный стандартный номер периодических изданий	Уникальный восьмизначный номер, используемый для идентификации периодических изданий по ИСО 3297 [32]
LEXUS	[не акроним]	Лексический инструментарий, основанный на интернет-технологии
LMF	Lexical Markup Framework Инфраструктура лексической разметки	Стандартизованная инфраструктура электронных словарей
METS	Metadata Encoding and Transmission Standard Стандарт кодирования и передачи метаданных	Стандарт метаданных на языке XML для электронных библиотек
MLA Style Guide	Modern Language Association of America Американская ассоциация современного языка	Руководство по стилю оформления цитируемой информации
MPEG	Moving Picture Experts Group Экспертная группа по движущимся изображениям	Рабочая группа 11 объединенного технического комитета ИСО/МЭК СТК1/ПК29, разрабатывающая стандарты кодирования аудио- и видеосигналов
NAAN	Name Assigning Authority Number Номер, присвоенный организации уполномоченным органом по именованию	Идентификационный номер организации в рамках PID-идентификатора в системе ARK
NCSTRL	Networked Computer Science Technical Reference Library Сетевая библиотека научных публикаций по вопросам вычислительной техники	Распределенная сеть хранения научных публикаций и отчетов факультета вычислительной техники Корнельского университета
NID	Namespace Identifier Идентификатор пространства имен	Уникальный идентификационный код репозитория
NMA	Network Management Application Прикладная система управления сетью	Программное обеспечение, являющееся частью эталонной модели взаимодействия открытых систем
OAIS	Open Archival Information System	Эталонная модель ИСО для системы долговременного хранения данных
OCLC	Online Computer Library Center Оперативно доступный центр компьютеризованных библиотек	Служба электронных библиотек и исследовательская организация, ответственная за систему Дублинского ядра
OWL	Web Ontology Language Язык сетевых онтологий	Семейство языков представления знаний в инфраструктуре описания ресурсов, используемых консорциумом W3C для создания онтологий
PDF	Portable Document Format Формат переносимого документа	Общедоступный стандарт для представления электронных документов
PID	Persistent Identifier Постоянный идентификатор	Постоянная ссылка на оперативно доступный ресурс

Окончание таблицы С.1

Акроним	Расшифровка сокращения/перевод	Комментарий
PII	Publisher Item Identifier Идентификатор продукции издателя	Уникальный идентификатор, используемый некоторыми издательствами научных журналов
PURL	Persistent URL Постоянный URL	Неизменяемая ссылка на URL
RDF	Resource Description Framework Инфраструктура описания ресурсов	Язык описаний для семантической сети
RFC	Request For Comments Запрос комментариев	Процедура IETF, нацеленная на получение комментариев специалистов по предложенному документу
SICI	Serial Item and Contribution Identifier Идентификационный номер периодического издания и его публикаций	Уникальный идентификатор для конкретных томов, статей и других идентифицируемых частей периодического издания
STD-DOI	Publication and Citation of Scientific Primary Data Проект «Публикация и цитирование научных первоисточников»	Стандартный механизм цитирования первичных научных данных
TBX	TermBase eXchange Обмен терминологическими базами данных	Стандарт обмена терминологическими базами систем перевода
TCP	Transmission Control Protocol Протокол управления передачей	
TMF	Terminological Markup Framework Инфраструктура терминологической разметки	См. ИСО 16642:2003 [33]
URI	Uniform Resource Identifier Унифицированный идентификатор ресурса	Компактная строка символов, используемая для идентификации или именования ресурса
URL	Uniform Resource Locator Унифицированный указатель ресурса	URI, который относится к оперативно доступному сетевому ресурсу
URN	Uniform Resource Name Унифицированное имя ресурса	URI, который является именем оперативно доступного сетевого ресурса
XML	Extensible Markup Language Расширяемый язык разметки	Формализованный метаязык, используемый для создания специализированных языков для совместного использования общих структурированных данных
XRI	eXtensible Resource Identifier Расширяемый идентификатор ресурса	Схема и протокол разрешения абстрактных идентификаторов

Приложение ДА
(справочное)

Сведения о соответствии ссылочных международных стандартов ссылочным национальным стандартам Российской Федерации

Таблица ДА.1

Обозначение ссылочного международного стандарта	Степень соответствия	Обозначение и наименование соответствующего национального стандарта
ИСО 12620:2009	—	*
ИСО 21000-17:2006	—	*

* Соответствующий национальный стандарт отсутствует. До его утверждения рекомендуется использовать перевод на русский язык данного международного стандарта. Перевод данного международного стандарта находится в Федеральном информационном фонде технических регламентов и стандартов.

Библиография

[1] Altman, M. and King, G. A Proposed Standard for the Scholarly Citation of Quantitative Data [в режиме онлайн], *D-Lib Magazine*, **13**(3/4), 2007, ISSN 1082—9873 [просмотр 2010-08-04]. Доступен по адресу: http://www.dlib.org/dlib/march07/altman/03altman.html

[2] Страница Annotea в «Википедии» [в режиме онлайн] [просмотр 2010-08-02]. Доступна по адресу: *http://en.wikipedia.org/wiki/Annotea*

[3] APA. *Style Guide to Electronic References*, June 2007, ISBN: 1-4338-0309-7

[4] Berners-Lee, T., *et al. Uniform Resource Identifier (URI): Generic Syntax*, IETF RFC 3986, январь 2005

[5] Berners-Lee, T., Masinter, L. and McCahill, M. *Uniform Resource Locators*, IETF RFC 1738, декабрь 1994

[6] Brase, J. *Using Digital Library Techniques — Registration of Scientific Primary Data* [в режиме онлайн]. Lecture Notes in Computer Science 3232, pp. 488—494, 2004, ISSN 0302-9743

[7] Dublin Core Metadata Initiative (DCMI), *Terminology* [в режиме онлайн] [просмотр 2010-08-04]. Available from: http://www.ukoln.ac.uk/metadata/dcmi/abstract-model/2004-12-08/#sect-7

[8] Fielding. R., *et al. Hypertext Transfer Protocol — HTTP/1.1*, IETF RFC 2616, июнь 1999

[9] Modern Language Association of America. *MLA Handbook for Writers of Research Papers*. New York: MLA, 7th ed. New York: MLA, 2009

[10] GATE. [просмотр 2010-08-04], Доступен по адресу: http://www.gate.ac.uk

[11] González, R. and Suarez Araújo, C.P., eds. *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Paris: European Language Resource Association. pp. 1321—1326, 2002

[12] Haase, P., Broekstra, J., Eberhart, A. and Volz, R. A comparison of RDF query languages. *Proceedings of the Third International Semantic Web Conference*, Hiroshima, Japan, 2004

[13] IANA. *Approved URI schemes* [в режиме онлайн], [просмотр 2010-08-04]. Доступен по адресу: http://www.iana.org/assignments/uri-schemes.html

[14] International DOI Foundation, *The Digital Object Identifier (DOI) System* [в режиме онлайн], февраль 2001 [viewed 2010-08-04], Available from: http://dx.doi.org/10.1000/203

[15] ИСО 24613:2008 (ISO 24613:2008) Управление лингвистическими ресурсами. Схема лексической разметки
Language resource management — Lexical markup framework (LMF)

[16] ИСО 690:2010 (ISO 690:2010) Информация и документация. руководящие указания по библиографическим ссылкам и цитированию информационных источников
Information and documentation — Guidelines for bibliographic references and citations to information resources

[17] Kemps-Snijders, M., Nederhof, M.-J. and Wittenburg, P. LEXUS, a web based tool for manipulating lexical resources. *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC2006) (pp. 1862—1865) [на компакт-диске]

[18] Kunze, J. Towards Electronic Persistence Using ARK Identifiers. *Proceedings of the 3rd ECDL Workshop on Web Archives* [в режиме онлайн], August 2003 (PDF) [просмотр 2010-08-04]. Доступен по адресу: *http://bibnum.bnf.fr/ecdl/2003/proceedings.php?f=kunze*

[19] Kunze, J. and Rodgers, R.P.C. *ARK Persistent Identifier Scheme* (интернет-проект [в режиме онлайн], обновлен: май 2008) [просмотр 2010-08-04]. Доступен по адресу: http://tools.ietf.org/id/draft-kunze-ark-15.txt

[20] METS. *Metadata Encoding and Transmission Standard* [в режиме онлайн] [просмотр 2010-08-04]. Доступен по адресу: http://www.loc.gov/standards/mets/

[21] Moats, R. *URN Syntax*, IETF RFC 2141, май 1997

[22] Pfeiffer, S., *et al. Specifying time intervals in URI queries and fragments of time-based Web resources* [в режиме онлайн], 2005 [просмотр 2010-08-04]. Доступен по адресу: http://www.annodex.net/TR/draft-pfeiffer-temporal-fragments-03.html

[23] Shafer, K., *et al. Introduction to Persistent Uniform Resource Locators (PURL)* [в режиме онлайн], 1996, [просмотр 2010-08-04]. Доступен по адресу: http://www.isoc.org/inet96/proceedings/a4/a4_1.htm

[24] Sun, S., Lannom, L. and Boesch, B. *Handle System Overview*. IETF RFC 3650, ноябрь 2003

[25] Thompson, H. and Orchard, D. *URNs, Namespaces and Registries*, проект под редакцией W3C TAG [в режиме онлайн], 2006-08-17 [просмотр 2010-08-04]. Доступен по адресу: http://www.w3.org/2001/tag/doc/URNsAndRegistries-50

[26] Tipster Text Program [в режиме онлайн] [просмотр 2010-08-04]. Доступен по адресу: http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/

[27] Troncy, R., *et al. Use cases and requirements for Media Fragments* [в режиме онлайн], *W3C Working Draft*, 2009-04-30 [просмотр 2010-08-04]. Доступен по адресу: http://www.w3.org/TR/2009/WD-media-frags-reqs-20090430/

[28] Wittenburg, P., Peters, W. and Broeder, D. Metadata Proposals for Corpora and Lexica. In: R. GONZELEZ and C.P. SUAREZ ARAUJO, eds. *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Paris: European Language Resource Association, 2002, pp. 1321-1326

[29] XRI. *Extensible Resource Identifier* [в режиме онлайн] [просмотр 2010-08-04]. Доступен по адресу: http://www.oasis-open.org/committees/xri

[30] W3C. *Cool URIs for the Semantic Web*: W3C Interest Group Note 31 [online] March 2008. SAUERMANN, L. and CYGANIAK, R., eds., 2008 [просмотр 2010-08-04]. Доступен по адресу: http://www.w3.org/TR/cooluris/

[31] ИСО 2108:2005 Информация и документация. Международный стандартный книжный номер (ISBN)
(ISO 2108:2005) Information and documentation — International standard book number (ISBN)

[32] ИСО 3297:2007 Документация. Международный стандартный номер серийного издания (ISSN)
(ISO 3297:2007) Information and documentation — International standard serial number (ISSN)

[33] ИСО 16642:2003 Применение компьютера в терминологических целях. Структура терминологической разметки
(ISO 16642:2003) Computer applications in terminology — Terminological markup framework

УДК 001.4:006.354 ОКС 01.140.20 Т58

Ключевые слова: менеджмент языковых ресурсов, постоянная идентификация, устойчивый доступ, постоянный идентификатор PID, терминологический ресурс, цитирование и указание ссылок на языковые ресурсы

Редактор *Т.С. Никифорова*
Технический редактор *В.Н. Прусакова*
Корректор *В.Е. Нестерова*
Компьютерная верстка *В.И. Грищенко*

Сдано в набор 02.03.2015. Подписано в печать 23.03.2015. Формат 60x84 1/8. Гарнитура Ариал. Усл. печ. л. 4,18.
Уч.-изд. л. 3,35. Тираж 31 экз. Зак. 1329.

Издано и отпечатано во ФГУП «СТАНДАРТИНФОРМ», 123995 Москва, Гранатный пер., 4.
www.gostinfo.ru info@gostinfo.ru