Г. А. Ботвин, **А. Н. Порошин** (Санкт-Петербург, СПбГУ). **Модернизация структуры гипертекста с целью повышения эффективности сетевых информационных ресурсов.**

Развитие инновационной экономики и экономики, основанной на знаниях, предполагает формирование корпоративных информационно-коммуникационных инфраструктур (ИТ-инфраструктур). В широком смысле понятие ИТ-инфраструктуры компании включает корпоративные сети, информационные порталы (сайты), бизнес-приложения, серверы, центры обработки и хранилища данных, а в органах государственной власти и наиболее динамично развивающихся отраслях экономики также и ситуационные центры [1].

Возрастание роли ИТ-инфраструктуры в повседневной хозяйственной деятельности предприятия приводит к необходимости обеспечения стабильного функционирования всего набора ИТ-инструментов. Одним из важнейших из них являются доступные посредством сети Интернет порталы и сайты, представленные в виде электронных гипертекстовых ресурсов (ЭГР). Основой гипертекста является гиперссылка, обеспечивающая его связность, т. е. возможность перехода от заданной якорем-источником точки исходного документа (например, веб-страницы) к целевому документу (или его фрагменту) по задаваемому ссылкой якорю места назначения, Документы обозначаются с помощью специального идентификатора ресурса, на основе которого программа-браузер, связавшись с компьютером-сервером посредством протокола обмена данными HTTP, может определить местонахождение ресурса и обеспечить доступ к нему, т. е. обычно загрузку и просмотр на экране компьютере посетителя сайта.

С развитием информационных технологий ранее применявшийся для определения местоположения данных Единый указатель ресурсов (URL), указывавший конкретное физическое расположение веб-страницы (т. е. путь к ней в файловой системе сервера), был заменен Унифицированным идентификатором ресурса (URI), который теперь определяет уже не конкретный адрес ресурса, а схему его обозначения и зависящее от нее имя. Тем самым URI не всегда указывает на то, как получить ресурс, в отличие от URL, а только идентифицирует его, что позволяет описывать наряду с обычными файлами и абстрактные ресурсы, которые не могут быть непосредственно получены по сети Интернет (например, предприятие, его структурные единицы, виды продукции, товара, услуги и т. п.).

В настоящее время подавляющее большинство мировых информационных ресурсов представлено в документальной текстово-графической форме веб-страниц, а переход по ссылке приводит к необходимости определения их физического размещения в сети, которое очень часто по-прежнему задается URI, представляющим имя файла. Процесс разрешения URI, т. е. его преобразование в адрес файла, не всегда гарантирует получение доступа к документу. Это происходит в связи с устареванием ссылок, содержащих на момент обращения уже неактуальные адреса ресурсов, которые могли быть изменены в связи с переименованием, перемещением или удалением документов.

Такое нарушение ссылочной структуры гипертекста и появление так называемых «битых ссылок» на практике встречается достаточно часто и приводит к невозможности получения требуемой информации, что в условиях ведения современного бизнеса является совершенно недопустимым [2]. Тем самым чисто технические трудности перерастают в серьезную социально-экономическую проблему получения гарантированного доступа к данным.

К настоящему времени существует несколько подходов к решению этой проблемы. Первым следует назвать проект на основе Persistent Uniform Resource Locator (PURL, постоянный единообразный локатор ресурсов). Идея проекта заключается в том, что постоянный идентификатор ресурса PURL указывает не на конкретное место расположения ресурса, а на запись в специальной базе данных PURL. При

обращении к данным производится поиск в этой базе и определение текущего адреса ресурса, после чего выполняется автоматическое перенаправление браузера к целевому файлу (адресу веб-страницы) при помощи стандартных средств HTTP-протокола. В случае изменения адреса ресурса необходимо лишь обновить запись в базе данных, содержащую адрес целевого ресурса, и многочисленные (косвенные) сетевые ссылки на него останутся работоспособными. К достоинствам подхода можно отнести наличие свободно распространяемого программного обеспечения для работы со ссылками на основе PURL, а к недостаткам — его недостаточную популярность, отсутствие единого стандарта и необходимость регулярного обновления базы данных при изменении местоположения целевого ресурса, что требует наличия высокопроизводительных серверов.

Вторым подходом является проект Digital Object Identifier (DOI) — идентификатор цифрового объекта. DOI — это стандарт обозначения представленной в сети информации об объекте, в которой содержится указатель его местонахождения (например, URL), его имя (название), прочие идентификаторы объекта (например, ISBN для электронной публикации) и связанный с объектом набор описывающих его метаданных в структурированном и расширяемом виде. DOI отличается от проекта PURL тем, что дополнительно содержит сравнительно большой объем метаданных, необходимых для описания ресурсов стабильных цифровых коллекций типа издательств, библиотек, научных фондов и т. п. Идентификатор DOI является постоянным для документа, тогда как его расположение и другие метаданные могут измениться. Ссылка к онлайн-документу по его DOI обеспечивает более надежный доступ, чем просто обращение по URL, поскольку при изменении местоположения ресурса его администратор должен только обновить метаданные объекта, чтобы возобновить связь с ним по новому адресу. Достоинством проекта DOI можно считать то, что идентификатор DOI идентифицирует как оригинальный единственный объект, а не просто задает его месторасположения. Тем самым реализуется концепция унифицированного идентификатора ресурса, к которой добавляется модель данных и социальная инфраструктура. Имя DOI также отличается от стандартных идентификационных регистров, таких как ISBN, ISRC и т. п.

Одним из серьезных недостатков DOI называют тот, что система не является свободной и доступна лишь на платной основе. Она реализуется через федерацию регистрационных агентств под эгидой Международного фонда DOI, который разработал и управляет этой системой публикации в Интернет с 2000 года. На конец 2009 около 4000 организаций владели приблизительно 43 млн. имен DOI.

Достойным упоминания подходом является поисковая система нарушенных связей на основе программы PageChaser, разработанная японскими специалистами. Идея подхода основана на поиске так называемых авторитетных страниц, ссылки с которых на искомые ресурсы всегда актуальны. При этом используются существующие поисковые машины индексные серверы, результаты которых и предоставляют материал, необходимый для отыскания авторитетных страниц. Неоспоримым достоинством таких систем является независимость от централизованной базы данных — хранилища адресов целевых веб-страниц, а в качестве недостатков следует назвать необходимость периодического запуска системы для поддержания актуальности ссылочной структуры и недостаточно высокую вероятность определения нового местоположения ресурса — на уровне 60–80%.

Авторы предлагают оригинальный подход к решению проблемы нарушенных связей путем модернизации структуры гипертекста, основанный на использовании специализированной БД, управляемой системой управления контентом (CMS) и размещаемой на серверах владельца ресурса. При этом гиперссылка представляется специальным идентификатором внутри операционной среды сайта и обновляется автоматически без участия администратора ресурса. Наличие автоматической динамической многоуровневой обработки таких идентификаторов, эквивалентных косвенной

адресации в языках программирования, а также программной обработки на стороне сервера дескриптивной информации о ресурсе, может обеспечить необходимый уровень дискретизации управления им. Так, появляется возможность свободно и динамически изменять физическое местоположение ресурса, накладывать ограничения на общее и астрономическое время его использования, географическое расположение выдающего запрос пользователя и др.

Предлагаемое решение позволяет обеспечить гарантированный доступ к ресурсам, повышенный уровень их безопасности по сравнению с традиционным подходом, эффективное управление ими на уровне отдельного сервера и сохранение требуемого быстродействия при обращении к данным. Указанный подход может найти широкое применение как при организации доступа к существующим коллекциям цифровых ресурсов, так и при построении распределенных гипертекстовых баз данных в образовании, управлении, бизнесе и во многих других предметных областях.

СПИСОК ЛИТЕРАТУРЫ

1. *QI Xianfeng, LAN Boxiong , GUO Zhenwei* Conceptual Model of IT Infrastructure Capabilityand Its Empirical Justification.
2. *Morishima A.*, *Nakamizo A.*, *Iida T.*, *Sugimoto S.*, *Kitagawa H.* Bringing your dead links back to life: a comprehensive approach and lessons learned. Proceedings of the 20th ACM conference on Hypertext and hypermedia, 2009, Torino, Italy, 21.06–01.07.09, p. 15–24.
3. *Rachman G.* An undeclared war in cyberspace. [Electronic resource] / The Financial Times; 04.10.2010. URL: http://www.ft.com/cms/s/0/539534a0-cfeb-11df-bb9e-00144feab49a. html (date of access 05.10.2010).
4. Информационные системы в экономике : учебный сайт А. Порошина на портале ЭФ СПбГУ. [Электронный ресурс]. URL: http://study.econ.pu.ru/p07 (date of access 05.10.2010).