К ВОПРОСУ О ФОРМУЛИРОВКЕ ТРЕБОВАНИЙ ДЛЯ ПОСТРОЕНИЯ ИНФОРМАЦИОННЫХ СИСТЕМ НАУЧНО-ОРГАНИЗАЦИОННОЙ НАПРАВЛЕННОСТИ

В. Б. Барахнин, Ю. В. Леонова, А. М. Федотов
Институт вычислительных технологий СО РАН, Новосибирск, Россия
e-mail: bar@ict.nsc.ru, juli@ict.nsc.ru, fedotov@ict.nsc.ru

In this paper, we formulate the fundamental requirements for the construction of informational systems for scientific or managerial purposes, which are based on the author's experience in implementation of the distributed informational system of the SB RAS and the analysis of other scientific informational systems. Problems of the development of a uniform methodology for the formulation of these requirements are considered.

Введение

Доступ к информации — одна из основных проблем, возникающих в деятельности научного исследователя. На первый взгляд может сложиться впечатление, что развитие информационных технологий уже само по себе способно вывести работу с научной информацией на качественно новый уровень, но, к сожалению, это не совсем так. Дело в том, что современные информационные технологии предоставляют исследователю мощный аппарат для работы с *данными*, но не с *информацией* (которая понимается здесь как данные, привязанные к конкретной модели или наделенные семантической структурой [1]). Именно такие данные (особенно структурированные) наиболее интересны, так как, во-первых, данные представляют информационную ценность с точки зрения количества информации лишь тогда, когда они связаны с другими данными [2, 3], а во-вторых, автоматическая переработка информации возможна лишь при наличии ее описания с помощью некоторого алгоритма, т. е. формальной модели данных [4].

Требования научных работников (а также других пользователей ЭВМ), желавших использовать компьютеры для обработки не только данных, но и информации, привели к созданию различных моделей структурирования информации, прежде всего баз данных, и появлению на их основе электронных справочников, каталогов и, наконец, систем автоматического библиотечного обслуживания.

В конце 1980-х годов Международный консультативный комитет по телефонии и телеграфии CCITT (ныне ITU-T) предпринял попытку создания универсальной справочной системы, в основу которой были положены рекомендации (стандарт) X.500 [5]. Однако

эта система реализована далеко не в полном объеме, так как предложенный проект был излишне универсален, что привело к чрезмерной "тяжеловесности" системы.

В середине 1990-х годов возникло осознание необходимости интеграции разнородных научных ресурсов сети Интернет в единые научные информационные системы (НИС), которые позволили бы установить связи между разнородными документами, организовывать единые каталоги документов, а также создавать специализированные системы поиска. К числу наиболее известных систем, функционирующих в настоящее время, относятся CORDIS [6], информационная система "Библиотеки Конгресса США" [7], а также отечественные разработки: eLIBRARY [8], ЕНИП РАН [9], ИРИС СО РАН [10], ИС-Россия [11], Соционет [12], Информика [13] и др. Эти системы в той или иной степени удовлетворяют потребностям исследователей в информации, однако каждая из них страдает определенными недостатками, основным из которых (и общим практически для всех названных НИС) является недостаточность возможностей обеспечения интеграции ресурсов как внутри каждой из систем, так и с внешними системами.

Исходя из накопленного нами опыта создания распределенной информационной системы СО РАН [14] и анализа других НИС (прежде всего перечисленных выше) в настоящей статье мы рассмотрим проблемы выработки единой методологии формулировки требований, предъявляемых к информационным системам научно-организационной направленности.

1. Информационные системы как объект исследований системного анализа

В соответствии с [15] будем рассматривать информационную систему как множество связанных между собой *документов*, т. е. целостных информационных объектов, которые описывают, представляют, отображают или моделируют некоторые сущности реального мира.

Проанализируем информационные системы с использованием методики общей теории систем. Отметим, что классическое определение системы — "множество объектов вместе с отношениями между объектами и между их атрибутами" [16] — основано на тех же понятиях, что и, например, реляционная модель данных [17].

Рассмотрим, как при создании информационных систем реализуются основные системные принципы [18]:

— целостность (зависимость каждого элемента, свойства и отношения от его места и функций внутри целого);

— структурность (возможность описания системы через установление ее структуры, т. е. сетей связей и отношений системы);

— иерархичность (каждый компонент системы, в свою очередь, может рассматриваться как система, а исследуемая система — как компонент более широкой системы);

— множественность описания (посредством использования множества различных моделей);

— взаимозависимость системы и среды (система формирует и проявляет свои свойства в процессе взаимодействия со средой, являясь при этом активным компонентом взаимодействия).

Целостность системы проявляется в зависимости каждого объекта, свойства и отношения от его места и функций внутри целого и реализуется посредством использования

единого набора метаданных $M = \bigcup M^i$. Тем самым любой документ d_i системы представляется как $d_i = \langle m_i^{j,k} \rangle$, где $m_i^{j,k}$ — значения элементов метаданных M^j; k — количество значений (с учетом повторений) соответствующего элемента метаданных в описании документа.

Структурность системы обеспечивается оптимальным выбором модели связей между документами, позволяющей адекватно описывать различные аспекты соответствующих межсущностных отношений. Достаточно универсальный характер имеет, например, модель направленных связей [19]. Подробнее, если документ $d_{i'}$ входит в качестве значения элемента M^j метаданных документа d_i, то можно говорить о связи между этими документами вида $M^j\langle d_i, d_{i'}, m_{i,i'}^{l,k} \rangle$, где $m_{i,i'}^{l,k}$ — атрибуты этой связи, являющиеся значениями соответствующих элементов метаданных. Таким образом, выстраиваемые нами отношения фактически переносятся на уровень элементов, определяющих структуру документов.

Иерархичность информационной системы проявляется в том, что она состоит из, вообще говоря, разнородных подсистем, отвечающих тем или иным частным задачам. Документы, описываемые при помощи одних и тех же элементов метаданных, составляющих множество $M_i \subseteq M$, образуют *класс* K_i. Если $M_1 \subset M$, $M_2 \subseteq M$ и $M_1 \subset M_2$, то класс K_2 есть подкласс класса K_1. Документы, входящие в один класс и имеющие одинаковую тематическую направленность, называют *коллекцией*.

Множественность описания системы подразумевает наличие множества различных аспектов построения системы (информационная модель системы, содержательное наполнение, используемые технологии и пр.). Наиболее общий характер имеет описание *модели информационной системы*, которая строится посредством задания классов K_i, определяемых множествами элементов метаданных M_i , и типов возможных связей между классами $M^j\langle K_i, K_{i'} \rangle$ с указанием элементов метаданных $M_{i,i'}^j$, описывающих атрибуты соответствующих связей. Таким образом, для построения модели информационной системы используется комбинация иерархической и реляционной моделей данных.

Взаимозависимость системы и среды интересует нас в плане как разработки модели представления информации (этот вопрос подробно рассмотрен в [20]), так и отражения системой изменений во внешней среде (т. е. актуализации информации), что приводит к необходимости учета мотивации разработчиков системы и возможности использования этой мотивации в течение достаточно длительного времени. Для некоммерческих информационных систем, к которым относятся и НИС, важно иметь механизм, снижающий зависимость системы от изменения мотивации. Таким механизмом может, например, послужить максимальная автоматизация процесса актуализации информации.

2. Содержательное наполнение информационных систем организаций

Сформулируем, какие именно сущности реального мира должны отображать (описывать и т. д.) информационные системы организаций (не обязательно научных).

Любая деятельность человека предполагает определенное противопоставление субъекта и объекта деятельности [18], причем в качестве субъекта деятельности могут выступать как отдельные люди, так и группы (коллективы) людей. В условиях современного общества производственно-технические отношения между людьми возникают, как правило, посредством вхождения этих людей в один коллектив, а характер этих отношений определяется функциями конкретного человека в коллективе. В свою очередь, группы

(коллективы) также могут вступать между собой в те или иные общественные отношения (подчиненности, учредительства и т. п.).

Таким образом, процесс деятельности организации может быть охарактеризован описаниями следующих сущностей:

1) субъекты деятельности — группы, отдельные лица;

2) объекты деятельности — продукты деятельности, акты деятельности.

Между этими сущностями устанавливаются связи:

1) отношения между субъектами и объектами деятельности (группа — объект деятельности, лицо — объект деятельности);

2) отношения между субъектами деятельности — связи типа группа — группа, лицо — группа, лицо — лицо.

Что же касается связей между объектами деятельности, то ввиду их большой специфичности для каждой конкретной сферы деятельности в рамках данной статьи этот вопрос не рассматривается.

Следует отметить, что при создании информационной системы выбор конкретного набора описаний из приведенного выше списка определяется родом деятельности организации. В частности, для научных организаций ввиду персонифицированности труда научных работников требуется отображать подробную информацию о персональном составе (включая возможность вхождения в целый ряд структур, а также отслеживание служебных перемещений) и связи между субъектами и продуктами деятельности. С другой стороны, включение в общую структуру системы информации об актах деятельности (конференциях и проектах), как это предлагается в [21], представляется нам недостаточно целесообразным, так как это излишне "утяжеляет" систему, затрудняя обновление информации. Подобного рода сведения удобнее хранить на новостных лентах (с возможностью доступа и к архивной новостной информации).

Основным результатом деятельности научных организаций являются новые научные знания. Как правило, эти знания становятся достоянием общества путем создания информационного продукта — публикаций в научных изданиях (печатных или электронных). Особенность такого продукта заключается в том, что его создателю, заинтересованному в максимально широком распространении продукта, обычно неважно, какую форму носит это распространение — коммерческую или бесплатную, в отличие, например, от информационных продуктов художественного характера (книг, музыкальных произведений, фильмов и т. п.), создатели которых нацелены на их коммерческое распространение. Поэтому оптимальная форма представления публикации в информационной системе научного сообщества — размещение документа-описания публикации с указанием url-адреса документа-представления (т. е. полной электронной версии данной публикации) или наиболее подробного варианта документа-описания.

Таким образом, содержательное наполнение НИС, адекватно удовлетворяющее информационные потребности пользователей, должно отвечать следующим требованиям:

— наличие подробной информации о структуре, включая группы, не входящие в основную административную структуру;

— наличие подробной информации о персонах, причем сведения об отношении персоны к структуре не утрачивают актуальность даже после прекращения данного отношения;

— максимально подробное представление информации о предмете деятельности, причем соответствующие документы не утрачивают актуальность с течением времени;

— наличие сохраняющих актуальность связей между персонами и объектом деятельности.

3. Основные подсистемы НИС и подходы к их реализации

В настоящее время в нашей стране не существует информационных систем, отражающих в достаточно полном объеме деятельность какой-либо научной корпорации. Действующие информационные системы, как правило, имеют целью подробно описать один из трех названных выше типов сущностей (группы, персоны, объекты деятельности), характеризующих деятельность научного сообщества. Например, сайты "Члены Российской академии наук" [22] или "Научные сотрудники — математики СО РАН" [23] содержат персональную информацию. В базе данных "Организации СО РАН" [24] представлена организационная структура Отделения, а в различных библиотечных системах или в базах данных инновационных разработок [25] описываются объекты деятельности.

При построении информационной модели, лежащей в основе систем узкой тематики, сущности соответствующего типа становятся независимыми, а сущности других типов — зависимыми. Это обстоятельство способно оказать значительное влияние на выбор конкретной модели данных. Так, для разработки информационных систем, отражающих различные аспекты деятельности персон, целесообразно использовать подход, развитый в сетевых операционных системах и приложениях, которые используют справочники для хранения информации о пользователях. Независимыми сущностями таких справочников являются персоны, объединяемые в группы. Поскольку между персонами нет прямых отношений подчиненности, справочники имеют плоскую структуру. Примером приложений являются почтовые клиенты MS Outlook Express и Netscape Communicator, ориентированные на схему данных X.500 [5].

В подходе, реализуемом в справочных системах организаций, а также системах, подобных "желтым страницам", независимая сущность — это организация. Система упорядочения организаций связана с административным и территориальным делением, что подразумевает жесткую иерархическую структуру справочника. В общей схеме служб справочника для организаций играет большую роль протокол LDAP [26]. Поддержка стандарта LIPS (Lightweight Internet Person Schema) [27] обеспечивает относительную стандартизацию общей схемы представления персональных данных, таких как имя, организация, сертификат и контактная информация.

Наконец, хранение больших объемов информации об объектах научной деятельности организуется, как правило, с использованием протокола Z39.50 [28]. Независимая сущность таких систем — это объект деятельности. Персона (или организация) приобретает уровень словаря, помогающего идентифицировать персону как субъект данной деятельности, а не как отдельную сущность. В качестве стандарта описания данных используется схема GILS [29] — общие описания информационных ресурсов. Важное достоинство протокола Z39.50 — это возможность организации атрибутивного поиска, что позволяет, в частности, искать документы из разных коллекций, имеющих один или несколько общих атрибутов.

Таким образом, можно выделить два основных подхода к организации информации в НИС: *иерархический* (характерный для протокола X.500 и его "облегченной" версии LDAP) и *горизонтальный* (характерный для протокола Z39.50). К сожалению, каждый из этих подходов страдает определенными недостатками. Отсутствие в Z39.50 возможности построения иерархической структуры приводит к дублированию информации, относящейся к объектам с общими свойствами. С другой стороны, отсутствие в X.500 горизонтальных связей влечет необходимость повторения записей, описывающих объекты, связанные с тем или иным объектом.

Возникает проблема установления связей (перечисленных в п. 1) между документами, относящимися к разным составным частям системы, а также в отдельных случаях и между документами, относящимися к одной и той же составной части системы (например, между документами, описывающими организацию и ее неструктурные подразделения). Тем самым становится актуальной разработка технологии идентификации, спецификации и визуализации горизонтальных отношений между документами сущностями. С этой целью нами предложена модель направленных связей [19], в которой выстраиваемые отношения фактически переносятся на уровень элементов, определяющих структуру документов.

Заключение

Резюмируя вышесказанное, можно сформулировать следующие основные требования к НИС.

1. Научная информационная система должна содержать информацию об организации и ее структуре, о персонах и о предмете деятельности организации, т. е. содержать в качестве основных подсистем биографический словарь, подробную адресно-телефонную книгу и библиографический каталог, а также вспомогательные подсистемы (новостную ленту, архив событий и т. п.).

2. Каждая из этих подсистем создается с использованием моделей данных и технологий, наиболее предназначенных для хранения информации данного типа.

3. Между документами, хранящимися в разных подсистемах, устанавливаются связи, имеющие ряд атрибутов. Для описания этих связей используется модель информационной системы, включающая описание классов, определяемых соответствующими множествами элементов метаданных, и типов возможных связей между классами с указанием элементов метаданных, описывающих атрибуты соответствующих связей.

Список литературы

[1] Математический энциклопедический словарь. М.: Сов. энциклопедия, 1988. 847 с.

[2] Колмогоров А.Н. Три подхода к определению понятия “количество информации” // Проблемы передачи информации. 1965. Т. I, вып. 1. С. 3–11.

[3] Колмогоров А.Н. Теория информации и теория алгоритмов. М.: Наука, 1987. 303 с.

[4] Ляпунов А.А. О некоторых общих вопросах кибернетики // Проблемы кибернетики. 1958. Вып. 1. С. 5–22.

[5] RFC 1274 — The COSINE and Internet X.500 Schema // http://www.networksorcery.com/enp/rfc/rfc1274.txt

[6] The CERIF (Common European Research Information Format) Standard // http://www.eurocris.org/en/taskgroups/cerif/new_6/new_0/ /C%3A%5CDocuments+and+Settings%5Ceg53%5CDesktop%5CCERIF_2000_part2.pdf

[7] Библиотека Конгресса США. http://www.loc.gov/

[8] Научная электронная библиотека eLIBRARY.RU http://elibrary.ru/defaultx.asp

[9] ЕНИП РАН. http://www.ras.ru/

[10] База данных "Организации и сотрудники СО РАН". http://www.sbras.ru/sbras/db/

[11] Университетская информационная система РОССИЯ. http://www.cir.ru/index.jsp

[12] Соционет. http://socionet.ru/

[13] Государственный НИИ информационных технологий и телекоммуникаций "Информика". http://www.informika.ru

[14] Информационная система СО РАН. http://www.sbras.ru

[15] Шокин Ю.И., Федотов А.М., Леонова Ю.В. Принцип динамического формирования документов в информационных системах на примере интегрированной распределенной информационной системы (ИРИС) СО РАН // Тр. Четвертой Всерос. науч. конф. "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", Дубна, 15–17 октября 2002 г. Т. 2. С. 159–169.

[16] Холл А.Д., Фейджин Р.Е. Определение понятия системы // Исследования по общей теории систем. М.: Прогресс, 1969. С. 252–282.

[17] Codd E.F. A relational model of data for large shared data banks. Comm. ACM 13,6. (June 1970). P. 377–387.

[18] Философский энциклопедический словарь. М.: Сов. энциклопедия, 1983. 840 с.

[19] Барахнин В.Б., Леонова Ю.В. Информационная модель отношений между документами в информационной системе // Вычисл. технологии. 2005. Т. 10. Спецвыпуск. С. 129–137.

[20] Гуськов А.Е. О модели информационных цифровых систем // Там же. С. 58–70.

[21] Бездушный А.Н., Кулагин М.В., Серебряков А.А. и др. Предложения по наборам метаданных для научных информационных ресурсов // Там же. С. 29–48.

[22] Сайт "Члены Российской академии наук". http://www.ras.ru/members.aspx

[23] Сайт "Научные сотрудники — математики СО РАН". http://www.sbras.ru/sbras/math_soran/

[24] Сайт "Организации СО РАН". http://www.sbras.ru/sbras/db/dep.phtml?3++rus

[25] Сайт "Перечень важнейших разработок СО РАН, предлагаемых для широкого использования". http://www.sbras.ru/win/sbras/main-work.html

[26] RFC 2251 — Lightweight Directory Access Protocol (v3). http://www.faqs.org/rfcs/rfc2251.html

[27] RFC 2798 — Definition of the inetOrgPerson LDAP Object Class. http://www.faqs.org/rfcs/rfc2798.html

[28] Жижимов О.Л., Мазов Н.А. Принципы построения распределенных информационных систем на основе протокола Z39.50. Новосибирск: ИВТ СО РАН, 2004. 361 с.

[29] Global Information Locator Service (GILS). http://www.gils.net/

Поступила в редакцию 26 сентября 2006 г.