

СТРУКТУРИРОВАНИЕ ИНФОРМАЦИОННОГО ПРОСТРАНСТВА ТЕХНИЧЕСКОГО УНИВЕРСИТЕТА С ИСПОЛЬЗОВАНИЕМ ПРОЦЕССНОГО ПОДХОДА И СЕМАНТИЧЕСКОЙ ИДЕНТИФИКАЦИИ

УДК 004.65

Геннадий Владимирович Старцев,
к.т.н., доцент каф. Автоматизированных систем управления Уфимского государственного авиационного технического университета (УГАТУ)
Тел.: 8 (347) 273-78-23
Эл. почта: startsev@gmail.com

Мария Анатольевна Шилина,
к.т.н., доцент каф. Автоматизированных систем управления Уфимского государственного авиационного технического университета (УГАТУ)
Тел.: 8 (347) 273-78-23
Эл. почта: maria.shilina@gmail.com

Александр Александрович Бармин,
Аспирант, ассистент каф. Автоматизированных систем управления Уфимского государственного авиационного технического университета (УГАТУ)
Тел.: 8 (347) 273-78-23
Эл. почта: barmin.alexander@gmail.com

Олеся Владимировна Бармина,
Магистрант каф. Автоматизированных систем управления Уфимского государственного авиационного технического университета (УГАТУ)
Тел.: 8 (347) 273-78-23
Эл. почта: obarmina@outlook.com

В статье рассматриваются вопросы необходимости построения единого информационного пространства образовательного учреждения и идентификации слабоструктурированных информационных ресурсов в рамках бизнес-процессов. Предложена многомерная семантическая модель информационного ресурса, комплект математических моделей для описания системных моделей бизнес-процессов и методика идентификации информационных ресурсов. Рассматривается реализация предложенной методики на примере веб-портала выпускающей кафедры технического вуза.

Ключевые слова: CALS-технологии, единое информационное пространство, бизнес-процесс, теория категорий, многомерная модель.

Gennady V. Startsev,
PhD, the Automated Managements Systems Department, Ufa State Aviation Technical University (USATU)
Tel.: 8 (347) 273-78-23
E-mail: startsev@gmail.com

Maria A. Shilina,
PhD, the Automated Managements Systems Department, Ufa State Aviation Technical University (USATU)
Tel.: 8 (347) 273-78-23
E-mail: maria.shiling@gmail.com

Aleksandr A. Barmin,
Post-graduate student, the Automated Managements Systems Department, Ufa State Aviation Technical University (USATU)
Tel.: 8 (347) 273-78-23
E-mail: barmin.alexander@gmail.com

Olesya V. Barmina,
Master student, the Automated Managements Systems Department, Ufa State Aviation Technical University (USATU)
Tel.: 8 (347) 273-78-23
E-mail: obarmina@outlook.com

STRUCTURING OF INFORMATION SPACE OF TECHNICAL UNIVERSITY USING PROCESS APPROACH AND SEMANTIC IDENTIFICATION

Article describes questions of necessity building of common information space for educational institution and identification of weak-structured information resources during business-processes. Proposed multidimensional semantic model of information resource set of mathematical models for description of system models of business processed and method for identification of information resources. Implementation of proposed methods are described based on web-portal of university department.

Keywords: CALS-technologies, common information space, business process, categories theory, multidimensional model.

1. Введение

Опыт российских и зарубежных университетов показывает, что использование современных информационных технологий в деятельности вуза в настоящее время является необходимостью. Требования к автоматизации отражены в стандарте ISO 9001:2011, директивах ENQA, а также в образовательных стандартах третьего поколения и новой редакции Федерального закона «Об образовании». Отмечается, что необходимым условием развития современных университетов является создание и поддержание единого информационного пространства для обеспечения информационной прозрачности и мониторинга, в том числе со стороны государства, повышения качества образования, а также устранения территориальных барьеров (то есть академической мобильности, в том числе виртуальной). В связи с присоединением России к Болонскому соглашению, потребовалось создание новых образовательных стандартов, ориентированных на поддержку концепции «образование в течение всей жизни».

Таким образом, актуально решение проблемы структурирования информационного пространства образовательного учреждения на основе систематизации процессов сбора, обработки и анализа данных об учебном процессе, а также обеспечения их идентифицируемости и прослеживаемости.

Свойства идентификации и прослеживаемости являются необходимыми требованиями для адекватности построения сложных систем. В формализованных (структурированных) областях деятельности (область конструирования, производства и т.п.) эти требования составляют основу соответствующих стандартов (CAD / CAM / CAE).

Проблема прослеживаемости – это, прежде всего, документирование идентифицированных объектов в различной документации предметной области, а, следовательно, и в едином информационном пространстве, являющимся ее отображением.

Как показывает опыт, информационное пространство современного вуза является гетерогенным. В этом случае данные, характеризующие учебный процесс и отдельные его объекты, оказываются распределенными по различным репозиториям и имеют разный формат представления. Количество характеристик информационных объектов, физически представленных в виде записей

в базах данных, превышает пределы обозримости. Следовательно, эффективное удовлетворение информационных потребностей пользователей и оперативный доступ к данным возможны только с использованием развитых информационно-поисковых систем.

2. Обоснование необходимости формирования единого информационного пространства высшего учебного заведения с применением CALS-технологий

Университеты представляют собой консервативные социально-экономические системы. Тем не менее, переход к информационному обществу, а следовательно, изменение способов, сроков обработки информации, привел к значительному увеличению ее объемов, но вместе с тем, к расширению доступа к этой информации, существенному снижению межгосударственных и внутрироссийских барьеров, что требует от университетов непрерывной адаптации к новым условиям. Немаловажным фактором является также постоянное изменение потребностей рынка труда, а также новации в глобальной информационной инфраструктуре [1].

Необходимым условием выполнения требований стандартов серии ГОСТ Р ИСО 9000 и ENQA является высокая степень автоматизации бизнес-процессов и соответствующих процессов управления с целью обеспечения информацией, необходимой для поддержки процессов, жизненного цикла продукции и процессов, составляющих основу создания, поддержания и постоянного улучшения системы качества организации, а также их непрерывный мониторинг.

Доступ к актуальным первичным данным позволяет оперативно получать новую информацию и в условиях изменяющихся бизнес-процессов, новые знания. Однако многократно увеличивающиеся объемы обрабатываемой информации и сложность такой обработки и интерпретации данных требуют использования принципиально новых решений, методов и технологий, использования хранилищ данных и средств много-

мерного анализа, а также развитых поисковых инструментов.

3. Методика идентификации слабоструктурированных информационных объектов на основе семантической аннотации

Информационные ресурсы являются элементами информационного пространства. Информационные ресурсы, представленные в различных хранилищах отличаются по степени структурированности. Так, ресурсы представленные в виде реляционных или нереляционных баз данных являются наиболее структурированными, так как определены все их атрибуты, в том числе и ключевые. Информационные ресурсы, представленные в форме текстовых файлов являются неструктурированными и уникально идентифицируются только именем файла.

Существуют десятки моделей (или языков) представления информационных ресурсов в форме знаний для различных предметных областей. Большинство моделей знаний может быть сведено к следующим классам: продукционные модели; семантические сети; фреймы; формальные логические модели.

В данном случае информационные ресурсы идентифицируются с помощью уникального имени, связи между ресурсами устанавливаются с помощью отношений [2].

Таким образом, в каждом из описанных представлений:

- существует класс объектов Obj_c , обладающих общими свойствами;

- между объектами этого класса существуют отношения – реляционные связи в базе данных, семантические связи в базе знаний, отношения категоризации в файловой системе. Эти отношения имеют различную природу и можно сделать вывод о наличии морфизмов $Hom_c(A, B)$ между объектами $(A, B) \in Obj_c$;

- для каждой пары морфизмов $f \in Hom(A, B)$ и $g \in Hom(B, C)$ определена композиция $f \circ g \in Hom(A, C)$ – объекты информационного пространства связаны и возможен переход от одного объекта к другому, например, с помощью гиперссылок, в этом случае операция композиции ассоциативна;

- для каждого объекта A задан тождественный морфизм $id_A \in Hom(A, A)$ – объекты информационного пространства идентифицируемы и могут быть получены непосредственно по идентификатору [4, 2].

На основе изложенных выше положений можно сделать вывод о существовании категории над множеством объектов и связей информационного пространства (объекты и связи информационного пространства – граф информационного пространства).

Множество информационных ресурсов организации представим в виде подмножеств структурированных, слабоструктурированных и неструктурированных информационных ресурсов:

$$D = (D_S \cup D_H \cup D_U), \quad (1)$$

где D – множество информационных ресурсов информационного пространства,

D_S – подмножество структурированных информационных ресурсов,

D_H – подмножество слабоструктурированных информационных ресурсов,

D_U – подмножество неструктурированных информационных ресурсов.

Подмножество структурированных информационных ресурсов можно представить в виде модели в нотации IDEF1X. В этом случае информационные ресурсы являются сущностями, а отношения между ними – связями. Сущности имеют уникальное имя и содержат ключевые и неключевые атрибуты. Ключевые атрибуты однозначно идентифицируют каждый экземпляр сущности в рамках информационного пространства, что позволяет отследить его положение в бизнес-процессе.

Модель структурированного информационного пространства можно представить в виде ориентированного графа, где узлами выступают сущности, а дугами – отношения между ними:

$$D_S = (E_d, R_d), \quad (2)$$

где E_d – множество сущностей предметной области,

R_d – множество отношений между сущностями.

Каждая сущность в структурированном информационном пространстве представляется в виде уникального имени и совокупности атрибутов:

$$E_d^i = (n, A_K, A_d), \quad (3)$$

$$A = A_d \cup A_k,$$

где n – наименование сущности,
 A – атрибуты сущности: A_d – неключевые атрибуты сущности,
 A_k – ключевые атрибуты сущности.

Слабоструктурированные информационные ресурсы содержат в себе структурированную и неструктурированную части. Структурированная часть информационного ресурса может быть представлена в виде информационной модели, неструктурированная часть – в форме онтологии:

$$D_H = (D_S \cup O), \quad (4)$$

где D_H – слабоструктурированный информационный ресурс,
 D_S – структурированная часть информационного ресурса,
 O – представление неструктурированной части в форме онтологии.

Модель онтологии имеет вид:

$$O = (T, R, F), \quad (5)$$

где O – онтология,
 T – термины предметной области, которые определяют онтологию O ,
 R – конечное множество отношений между терминами предметной области,
 F – конечное множество функций интерпретации заданных на терминах или отношениях онтологии O .

Неструктурированные информационные ресурсы целиком представляются в виде онтологии.

Для представления неструктурированных информационных ресурсов с помощью онтологии необходимо определить термины предметной области. Термины можно выделить путем индексирования всех информационных ресурсов или из тезауруса моделей бизнес процессов, выполняющихся в информационном пространстве.

Тезаурус представляет собой множество упорядоченных пар тер-

минов предметной области $t_i \in T$ и их описаний $d_i \in Desc, i = 1..N$:

$$Thes = \{(t_i, d_i)\}, \quad (6)$$

где T – множество терминов предметной области;
 $Desc$ – множество описаний терминов предметной области.

В тезаурус модели бизнес-процесса входят термины всех входных и выходных данных, нормативных документов и механизмов, участвующих в бизнес-процессе. На основе терминов предметной области формируется базис информационного пространства, в рамках которого могут быть идентифицированы неструктурированные ресурсы.

Классификацию неструктурированных информационных ресурсов можно провести на основе терминов, присутствующих в них. Стоит отметить, что классификация вариативна в зависимости от выбранных признаков классификации, но может быть сужена за счет использования моделей бизнес-процессов в качестве основания для классификации. В этом случае информационный ресурс может быть представлен в виде вектора в пространстве базиса предметной области:

$$R = \{t_1, t_2, \dots, t_n\}. \quad (7)$$

Представим модель бизнес-процесса также в виде подмножеств терминов онтологии предметной области:

$$BP = (I, C, O, M) = (T_i, T_c, T_o, T_m), \quad (8)$$

где I – входные данные,
 T_i – термины, связанные со входными данными,
 C – нормативные документы,
 T_c – термины, связанные с нормативными документами,
 O – выходные данные,
 T_o – термины, связанные с выходными данными,
 M – механизмы и исполнители бизнес-процесса,
 T_m – термины, связанные с механизмами и исполнителями.

Также, каждую функцию модели бизнес-процесса можно описать в виде множества терминов:

$$F_j = (T_i^j, T_c^j, T_o^j, T_m^j), (T_i^j, T_c^j, T_o^j, T_m^j) \in T^j. \quad (9)$$

Используя (9) и (7) можно идентифицировать документы, относя-

щиеся к каждой конкретной функции бизнес-процесса, то есть обеспечить их прослеживаемость в рамках бизнес-процесса.

Использование семантической идентификации не позволяет однозначно идентифицировать информационный ресурс, соответствующий запросу, но позволяет выделить группу ресурсов, удовлетворяющих запросу в определенной степени. Степень соответствия между информационным ресурсом и запросом можно определить на основе семантического расстояния между ними.

Запрос к информационному пространству представляется в виде совокупности множеств терминов T_q и булевых операций между ними R_q и может быть представлен в дизъюнктивной нормальной форме:

$$q = (T_q, R_q) = q_{dnf} = \bigvee_{i=1..N} q_i \quad (10)$$

где q_i – i -я конъюнктивная компонента запроса q .

Пусть i – индекс термина t_i из таксономии O ($i = 1, \dots, M$), r^j – информационный ресурс, принадлежащий информационному пространству D , а $w_i^j \geq 0$ – вес, ассоциированный с парой (t_i, r^j) .

Для каждого термина t_i , который не входит во множество терминов, связанных с информационным ресурсом r^j , его вес равен нулю: $w_i^j = 0$. Информационный ресурс будем рассматривать как вектор $r^j = (w_1^j, w_2^j, \dots, w_M^j)$.

Введем также инверсную функцию g , соответствующую индексу термина t_i , которая определяется следующим образом: $g_i(r^j) = w_i^j$.

Мера близости информационного ресурса r^j и запроса q – $\text{sim}(r^j, q)$ определяется выражением:

$$\text{sim}(r^j, q) = \begin{cases} 1, & \text{если } \exists q_i: (q_i \in \\ & \in q_{dnf}) \wedge (\forall k: g_k(q_i) = \\ & = g_k(r^j)) \\ 0 & \text{иначе} \end{cases} \quad (11)$$

То есть $\text{sim}(r^j, q)$ принимает значение 1, если существует такая конъюнктивная компонента q_i , входящая в дизъюнктивную нормальную форму q_{dnf} , что инверсная каждого термина k данной конъюнктивной компоненты совпадает с этой же инверсной функцией для информационного ресурса r^j . В противном

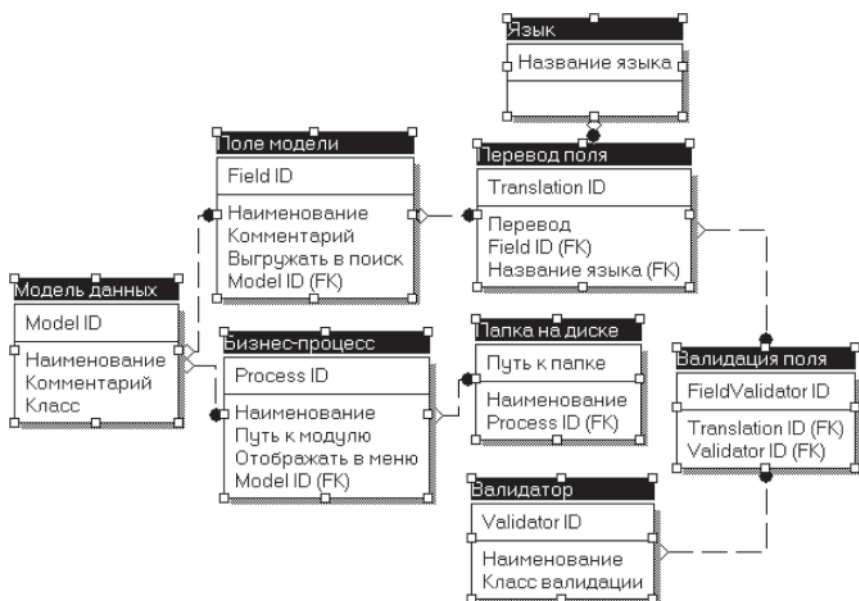


Рис. 1. Фрагмент информационной модели метаданных об объектах предметной области

случае $\text{sim}(r^j, q)$ оказывается равной 0. Таким образом, если $\text{sim}(r^j, q) = 1$, то информационный ресурс r^j является релевантным запросу q [3].

4. Реализация информационного поиска в контенте корпоративного веб-портала выпускающей кафедры технического вуза

Автоматизированная информационная система образовательного учреждения в процессе своей де-

ятельности накапливает существенный объем данных. Эти данные находятся в разнородных хранилищах и представлены в разнообразных форматах. Использование порталных технологий на основе сервис-ориентированной архитектуры позволяет обеспечить централизованный доступ к информационному пространству с использованием тонкого клиента – веб-браузера. Использование информационно-поисковых систем в качестве сервиса

информационного пространства обеспечивает оперативный поиск и доступ к информации.

Обмен данными между сервисами корпоративного веб-портала выполняется с использованием XML-веб-сервисов. Информационно-поисковая система является сервисом, выполняющим две функции:

1. Создание и обслуживание поискового индекса объектов информационного пространства.
2. Выдача ссылок на объекты информационного пространства, удовлетворяющие поисковому запросу.

Каждый бизнес-процесс выпускающей кафедры представляет собой отдельную задачу веб-портала, доступ к которой регламентируется с помощью системы прав доступа. Создание индекса объектов информационного пространства требует доступа к задачам веб-портала с правами супер-пользователя – все объекты информационного пространства должны попасть в поисковый индекс.

Фрагмент информационной модели, описывающий метаданные о сущностях предметной области приведен на рисунке 1.

На основе метаданных экземпляры объектов предметной области сохраняются в XML-документы, которые затем отправляются в поисковый сервис.

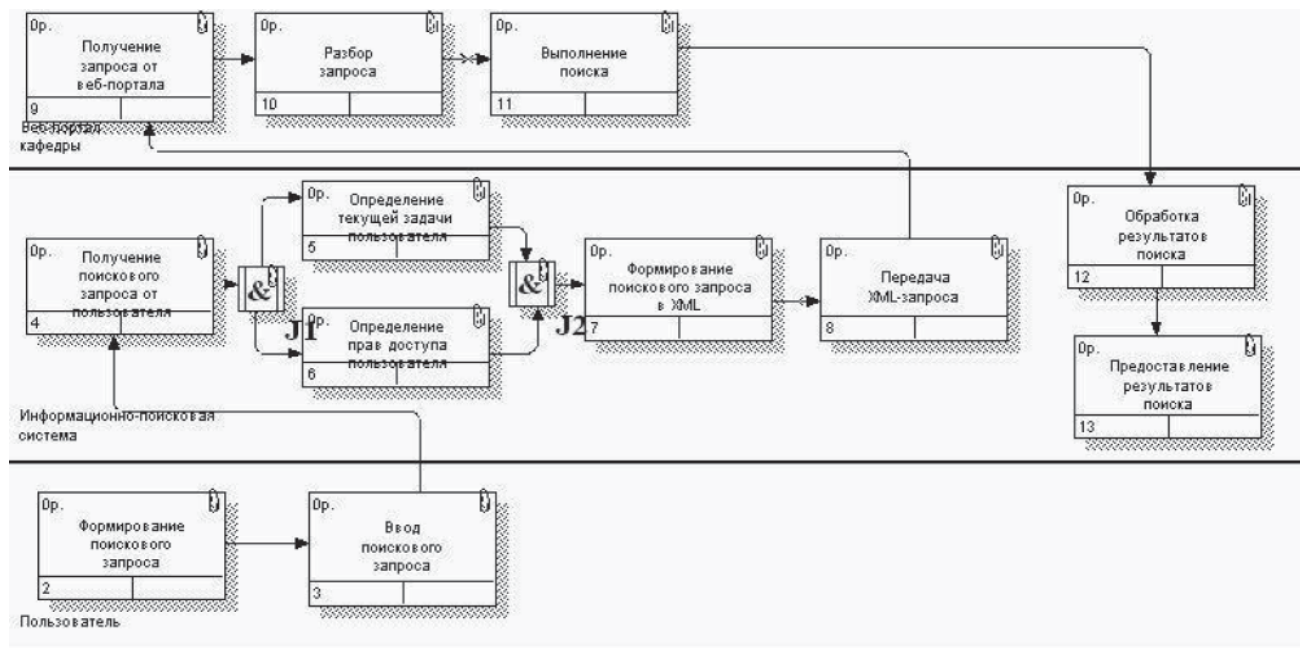


Рис. 2. Динамическая модель взаимодействия информационно-поисковой системы и сервисов веб-портала

Получение результатов поиска также происходит путем обмена XML-сообщениями между веб-порталом и сервисом информационного поиска. XML-сообщения со ссылками на объекты предметной области в соответствии с метаданными, хранящимися в базе данных веб-портала преобразуются в доступный для пользователя вид.

Динамическая модель взаимодействия пользователя, веб-портала и сервиса информационного поиска представлена на рисунке 2.

Описанное выше решение реализовано с использованием объектно-ориентированного языка программирования PHP 5, системы управления базами данных MySQL, сервера приложений Apache. В качестве сервиса информационного поиска выступает типовое решение Apache Solr.

Заключение

Свойства идентификации и отслеживаемости являются необходимыми требованиями для адекватности построения сложных систем. В формализованных (структурированных) областях деятельности (область конструирования, производства и т.п.) эти требования составляют основу соответствующих стандартов (CAD / CAM / CAE). Единое информационное пространство является ключевым компонентом концепции CALS. Данные, содержащиеся в едином информационном пространстве, относятся, в том числе, и к системе менеджмента качества. Таким образом, можно сделать вывод, что структурирование единого информационного пространства, систематизация сбора, обработки данных ведут к повышению качества и оператив-

ности информационной поддержки управления, а, следовательно, повышению эффективности управляемых процессов.

Семантическое аннотирование слабоструктурированных объектов информационного пространства позволяет идентифицировать их в рамках бизнес-процессов образовательного учреждения, организовать информационную поддержку жизненного цикла образовательного процесса на всех его этапах жизненного цикла. Таким образом, устанавливается соответствие между моделями бизнес-процессов и их экземплярами в информационном пространстве вуза.

Использование информационно-поисковых систем позволяет в автоматизированном режиме производить идентификацию объектов информационного пространства, выполнять их поиск и навигацию.

Применение информационно-поисковой системы в качестве сервиса корпоративного портала позволяет выполнять оперативный поиск информации в различных источниках. Доступ к корпоративному portalу осуществляется с помощью тонкого клиента – веб-браузера, доступного практически на любых устройствах. Таким образом, можно получать оперативную информацию об объектах информационного пространства.

Литература

1. Тихомирова Н.В., Минашкин В.Г., Дубейковская Л.Г. Образовательный процесс в электронном университете: условия и направления трансформации // Высшее образование в России: науч.-пед. журнал. – 2011. – №2. – С. 3-11

2. Кук Д., Бейз Г. Компьютерная математика: Пер. с англ. – М.: Наука, гл. ред. физ.-мат. лит., 1990. – 384 с.

3. Куликов Г.Г., Старцев Г.В., Бармин А.А. Подход к построению информационно-поисковых систем для систем электронного документооборота // Актуальные проблемы в науке и технике. Т. 1. Информационные и инфокоммуникационные технологии: сб. науч. тр. 8-й Всерос. зимн. shk.-sem. аспирантов и молодых ученых (Уфа, 14–16 февр. 2013). Уфа: УГАТУ, 2013. С. 405.

4. M.Barr, C.Wells. Category Theory for Computing Science: Prentice Hall; 1 edition (July 6, 1990), 350 p.

References

1. Tihomirova N.V., Minashkin V.G., Dubeikovskaya L.G. Educational process in digital university: conditions and directions of transformation // Vysshee obrazovanie v Rossii: nauch.-ped. zhurnal. – 2011. – №2. – S. 3-11

2. Kuk D, Beiz G. Computer mathematics: translation from english. – M.: Nauka, gl. red. fiz.-mat. lit., 1990. – 384 s.

3. Kylikov G.G., Startsev G.V., Barmin A.A., Approach in building information retrieval systems for content-management systems// Aktualnye problemy v nauke i tehnike. T. 1. Informacionnye i infokommunikacionnye tehnologii: sb. nauch. tr. 8-j Vseros. zimn. shk.-sem. aspirantov i molodyh uchenyh (Ufa, 14–16 fevr. 2013). Ufa: UGATU, 2013. S. 405.M.

4. M.Barr, C. Wells. Category Theory for Computing Science: Prentice Hall; 1 edition (July 6, 1990), 350 p.