УДК 681.3.06

СИСТЕМА АВТОМАТИЗИРОВАННОГО ПЕРЕНОСА СОДЕРЖИМОГО ЭЛЕКТРОННЫХ ДОКУМЕНТОВ В БД ИС

А.А. Блажко, С. Ю. Марулин, Ю.А. Дунько

Одесский национальный политехнический университет
65044, Одесса, проспект Шевченко, 1,
тел. 8 (048) 779-75-66, 8 (048) 779 7106
stasfoot@mail.ru

Представлено общую концепцию автоматизированного переноса содержимого электронных документов в БД ИС, разработана технология автоматизированного создания шаблонов документов формата *XLS* и механизм обратного переноса содержательных данных в БД ИС. Проведено апробацию работы технологии на нескольких примерах электронных документов с табличной структурой.

Presented the general conception of the automated carrying over of contents of electronic documents to DB IS, developed the technology of the automated creation templates documents format XLS and the mechanism of return carrying over of the substantial data in DB IS. It is spent approbation of job of technology on several examples of electronic documents with tabular structure.

Введение

Электронные документы (ЭД) – это основные информационные ресурсы предприятия, работа с которыми требует правильной организации. ЭД обеспечивают информационную поддержку принятия управленческих решений на всех уровнях и сопровождают ведение всех бизнес-процессов. Электронный документооборот - это непрерывный процесс движения документов, объективно отражающий деятельность предприятия и позволяющий оперативно управлять им. Множество бумажных копий, длительный поиск нужного документа, потери, дублирующие документы, задержки с отправкой и получением, ошибки персонала составляют не полный перечень проблем, возникающих при плохой организации электронного документооборота. Все это может сильно затормозить, а в исключительных случаях полностью парализовать работу предприятия.

Эффективный электронный документооборот (ЭЭД) является обязательной составляющей эффективного управления предприятием. ЭЭД в многом зависит от хорошо развитой информационной системы (ИС) предприятия, которая в свою очередь не может существовать без единого хранилища данных – базы данных (БД). Многие организации в своем управлении уже используют подобные системы, но их робота оставляет желать лучшего.

Например, в некоторой организации существует множество ЭД, которые создаются и обрабатываются в разных отделах разными лицами, с помощью офисных программ. В процессе работы эти документы с помощью системы ЭД передаются между отделами, где в них вносятся изменения и поправки. Так, к концу определенного периода времени накапливается массив таких документов. Управляющему этим предприятием необходимо провести какую либо статистику накопленного: посмотреть в разрезе те или иные данные, сопоставить разные характеристики, сделать выводы и внести корректировки в процесс. Но сделать перечисленное будет невозможно без единой БД ИС. Таким образом, существует проблема первичного создания БД ИС и проблема поддержки актуальности содержимого БД ИС.

Задача автоматизированного переноса содержимого ЭД в БД

Для уменьшения трудоемкости процесса первичного создания БД ИС и переноса содержимого ЭД в БД ИС предлагается следующая схема (рис. 1).

Как видно из рис. 1 процесс можно разбить на несколько этапов:

1. Параметризация массивов документов/документа.
2. Кластеризация документов.
3. Классификация документа.
4. Создание шаблонов классов документов.
5. Автоматизированный перенос содержательных данных в БД ИС.

На первом этапе необходимо выделить те ключевые характеристики, по которым будем относить тот или иной документ к определенному классу. Типы параметров ЭД:

Числовые: расположение ячейки относительно начала координат (номер строки и номер столбца);

Вещественные: название документа; диапазон ячеек; объединение/разбиение значение ячейки; оформление ячейки (фоновый цвет, метод заливки, выделение границ ячейки линиями, параметры линий: толщина, вид линии, цвет линии; формат текста: жирный, курсив, подчеркнутый, шрифт, размер шрифта, цвет шрифта).

Особый тип: формула.

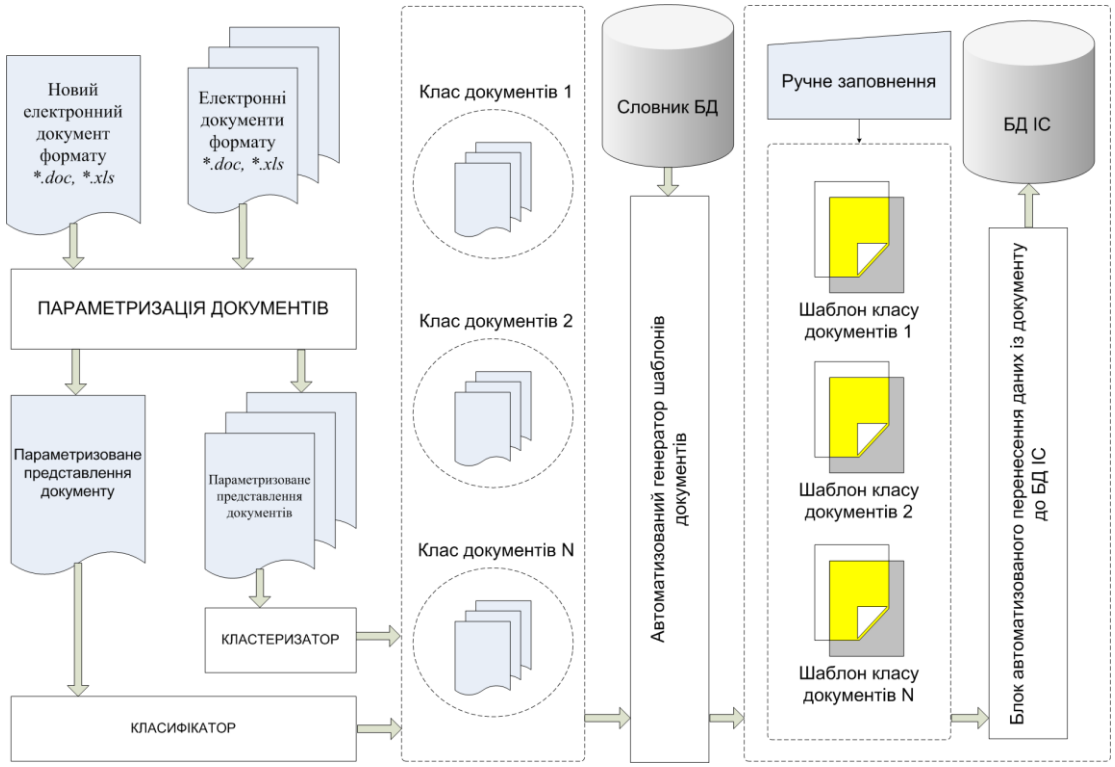

Рис. 1. Общая схема технологии

Наряду с числовыми и вещественными параметрами можно выделить метаданные которые описывают документ “снаружи”. Метаданные – это информация о данных [1]. Этот термин в широком смысле слова используется для любой информации о данных: именах таблиц, колонок в таблице в реляционных базах данных, номер версии в файле программы. Метаданные – это структурированные данные, представляющие собой характеристики описываемых сущностей для целей их идентификации, поиска, оценки, управления ими [2]. Иными словами – набор допустимых структурированных описаний, которые доступны в явном виде, предназначение которых может помочь найти объект [3]. Этап параметризации в системе автоматизированного переноса содержимого ЭД является ключевым, так как хорошо выбранные критерии или параметры дают нам возможность правильно отнести ЭД к нужному кластеру.

В зависимости от рода параметров выбирается алгоритм кластеризации.

Кластеризация документов. Обзор методов

Кластеризация документов — одна из задач информационного поиска. Цель кластеризации документов – автоматическое выявление групп семантически похожих документов среди заданного фиксированного множества документов. Следует отметить, что группы формируются только на основе попарной схожести описаний документов, и никакие характеристики этих групп не задаются заранее, в отличие от классификации документов, где категории задаются заранее.

Существует различные методы кластеризации:

1. **K-средние (K-means).** Наиболее популярный метод кластеризации. Алгоритм представляет собой модификацию EM-алгоритма для разделения смеси гауссиан. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k. Действие алгоритма таково, что он стремится минимизировать среднеквадратичное отклонение на точках каждого кластера:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

где k – число кластеров, Si – полученные кластеры, $i = 1,2,...,k$ и μi – центры масс векторов $x_j \in S_i$.

2. **Графовые алгоритмы кластеризации.** Обширный класс алгоритмов кластеризации основан на представлении выборки в виде графа. Вершинам графа соответствуют объекты выборки, а рёбрам попарные расстояния между объектами $p_{ij} = \rho(x_i, x_j)$. Достоинством графовых алгоритмов кластеризации является наглядность, относительная простота реализации, возможность вносить различные усовершенствования, опираясь на простые геометрические соображения.

3. **Алгоритмы семейства FOREL.** Алгоритмы семейства FOREL использует критерий F, основанный на гипотезе компактности: в один таксон должны собираться объекты, "похожие" по своим свойствам на некоторый «центральный» объект. В результате получаются таксоны сферической формы.

4. **Иерархическая кластеризация или таксономия.** Метод кластерного анализа – разбиения заданной выборки объектов (ситуаций) на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

5. **Нейронная сеть Кохонена.** Класс нейронных сетей, основным элементом которых является слой Кохонена. Слой Кохонена состоит из адаптивных линейных сумматоров («линейных формальных нейронов»). Как правило, выходные сигналы слоя Кохонена обрабатываются по правилу «победитель забирает всё»: наибольший сигнал превращается в единичный, остальные обращаются в ноль. Слой Кохонена состоит из некоторого количества n параллельно действующих линейных элементов. Все они имеют одинаковое число входов m и получают на свои входы один и тот же вектор входных сигналов $x = (x_1,...x_m)$. На выходе j-го линейного элемента получаем сигнал

$$y_j = w_{j0} + \sum_{i=1}^{m} w_{ji} x_i ,$$

где w_{ji} – весовой коэффициент i-го входа j-го нейрона, w_{j0} – пороговой коэффициент.

После прохождения слоя линейных элементов сигналы посылаются на обработку по правилу «победитель забирает всё»: среди выходных сигналов y_j ищется максимальный; его номер $j_{max} = \text{argmax}_j\{y_j\}$. Окончательно, на выходе сигнал с номером j_{max} равен единице, остальные – нулю. Если максимум одновременно достигается для нескольких j_{max}, то либо принимают все соответствующие сигналы равными единице, либо только первый в списке (по соглашению). «Нейроны Кохонена можно воспринимать как набор электрических лампочек, так что для любого входного вектора загорается одна из них» [5].

6. **EM-алгоритм.** Алгоритм, используемый в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей, в случае, когда модель зависит от некоторых скрытых переменных. Каждая итерация алгоритма состоит из двух шагов. На E-шаге (expectation) вычисляется ожидаемое значение функции правдоподобия, при этом скрытые переменные рассматриваются как наблюдаемые. На M-шаге (maximization) вычисляется оценка максимального правдоподобия, таким образом увеличивается ожидаемое правдоподобие, вычисляемое на E-шаге. Затем это значение используется для E-шага на следующей итерации. Алгоритм выполняется до сходимости. Часто EM-алгоритм используют для разделения смеси гауссиан.

Пусть X – некоторые из значений наблюдаемых переменных, а T – скрытые переменные. Вместе X и T образуют полный набор данных. Вообще, T может быть некоторой подсказкой, которая облегчает решение проблемы в случае, если она известна. Например, если имеется смесь распределений, функция правдоподобия легко выражается через параметры отдельных распределений смеси. Положим p – плотность вероятности (в непрерывном случае) или функция вероятности (в дискретном случае) полного набора данных с параметрами Θ: $p(X,T|\Theta)$. Эту функцию можно понимать как правдоподобие всей модели, если рассматривать её как функцию параметров Θ. Заметим, что условное распределение скрытой компоненты при некотором наблюдении и фиксированном наборе параметров может быть выражено так:

$$p(T|X,\Theta) = \frac{p(X,T|\Theta)}{p(X|\Theta)} = \frac{p(X|T,\Theta)\,p(T|\Theta)}{\int p(X|\hat{T},\Theta)\,p(\hat{T}|\Theta)\,d\hat{T}},$$

используя расширенную формулу Байеса и формулу полной вероятности. Таким образом, необходимо знать только распределение наблюдаемой компоненты при фиксированной скрытой $p(X|T,\Theta)$ и вероятности скрытых данных $p(T|\Theta)$.

7. **Метод опорных векторов (SVM – support vector machines).** Основная идея метода опорных векторов — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей наши классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

Критерии адекватности для методов кластеризации.

1. Обработка вновь поступающих документов не должна существенным образом изменять результат кластеризации.

2. Устойчивость: незначительные ошибки в описании объектов могут вызывать также незначительные изменения в результатах кластеризации.

3. Независимость результата кластеризации от исходного порядка на множестве объектов.

Поэтому выбор метода кластеризации зависит от конкретно поставленной задачи.

Третий этап технологии подразумевает создание единого шаблона для каждого выделенного класса. Созданные на предыдущем этапе классы ЭД содержат множество однотипных документов, но их структура и оформление порой очень разнятся. Для того, что выделить общие структуры документа и описать данные, характерные только этому классу, и тем самым уйти от неоднозначного описания ЭД разными структурными подразделениями организации используется блок автоматизированного создания шаблонов, к которому подключается словарь БД. Такой подход позволит выработать единые правила оформления документа и предоставит пользователям системы заполнять готовые шаблоны, а не придумывать их самостоятельно. Следует отметить тот факт, что обработка созданных и заполненных таким образом шаблонов значительно упростит задачу переноса данных из ЭД в БД ИС.

Четвертым этапом работы данной технологии является непосредственный перенос данных ЭД в соответствующие таблицы БД ИС.

Представление шаблонной модели документа

Для решения задачи автоматизированного переноса содержимого ЭД в БД ИС предложено два метода.

Первый метод основан на реляционном представлении структуры документа. Для этого ЭД описан в виде множества

$$\langle D, SD, SQI \rangle$$

Множество *D* (*dialog*) можно представить как четверку

$$\langle \ln, v, qr, id\_doc \rangle$$

где *ln* – название элемента *Label* окна диалога; *v* – порядковый номер элемента *Label* окна диалога; *qr* – запрос, на выбор данных для элемента *ComboBox* окна диалога; *id_doc* – уникальный идентификатор документа.

Это множество используется для создания окна диалога с пользователем, который задает критерии создания шаблона документа.

Множество данных *SD* представим как семёрку вида

$$\langle st, sl, d, sh, tp, vl, id\_doc \rangle$$

где *st, sl* – координаты начального место расположения блоков данных в документе; *d* – направление распространения блоков данных в документе; *sh* – шаблон разбивки блоков данных документа; *tp* – тип значения поля *vl*; *vl* – запрос или конкретная фраза; *id_doc* – уникальный идентификатор документа.

Значение параметра *d* есть {*C,D,U,R,L*}, определяют направление распространения блоков данных в документе. “*C*” – центрируемое, “*D*” – вниз, “*U*” – вверх, “*R*” – вправо, “*L*” – влево.

Значение параметра *sh* может быть любым символов, который будет разделять фразы на лексемы. Например если значение sh=” ”(пропуск), то фраза “*Ivanov Ivan Ivanovich*”, разделится на лексемы *Ivanov, Ivan, Ivanovich.*

Значение параметра *tp* есть {*C, S*}, определяют является ли значения поля *vl* константой (*C*) или запросом (*S*).

Значение параметра *vl* может быть запросом на выборку или словом.

Множество SQI представим как тройку вида

$$\langle id, q, id\_doc \rangle$$

где *id (identification)* – идентификатор запроса *q*; *q (query)* – запрос на внесение данных в таблице БД ИС; *id_doc* – уникальный идентификатор документу.

В табл. 1 приведен пример описания ЭД с табличной структурой.

Таблица 1. Пример описания ЭД

№	Stroka	Stolbec	Napravlenie	Shablon	Type	Value
1	1	1	C	S	C	FIO
2	1	2	C	S	C	SEM
3	2	1	D		S	Select “фамилия”||’’||“имя”||’’||“отчество” from students where “группа”=$1
4	1	3	R	-	S	Select “семестр”||’-’||“предмет”||’-’||“видконтроля” from predmet where “группа”=$1 and “семестр”=$2 and “видконтроля”=$3

Такое описание документа позволяет создавать шаблон документа по определенным критериям, задаваемыми пользователем, а также переносить данные из этого документа в БД ИС. На рис. 2 показан шаблон документа “Учебная ведомость”, описанный в табл. 1.

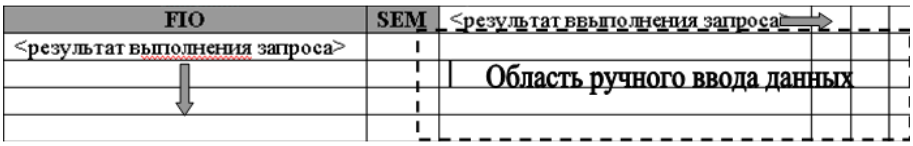

Рис. 2. Структура ЭД с табличной структурой

Алгоритм создания шаблона ЭД. Использует структуру данных, приведенную в табл. 1.

Шаг 1. Анализ поля *value* структуры *SD*, где *type* = '*S*'. Поле *value* содержит запрос, из которого выделяются лексемы, необходимые для создания структуры *D*, т.е. для создания окна диалога с пользователем, где он задает необходимые критерии создаваемого шаблона.

Шаг 2. Анализ поля *Stroka*, *Stolbec* и *value* структуры *SD*, где *type* = '*S*' и где *Napravlenie* = '*C*'. Поле *value* содержит константы в виде слов, которые будут появляться в документе как статические элементы.

Шаг 3. Пользователь задает критерии создания шаблона в диалоговом окне (рис. 3).

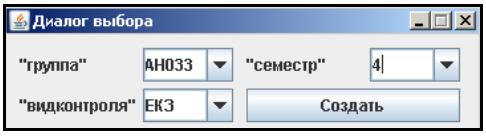

Рис. 3. Диалоговое окно выбора критериев

Шаг 4. Анализ поля *Stroka*, *Stolbec*, *Napravlenie* и *value* структуры *SD*, где *type* = '*S*' и где *Napravlenie* ≠ ≠ '*C*'. Ответы на запросы, находящиеся в поле *value*, располагаются в ЭД согласно значениям поля *Stroka*, *Stolbec* и в направлении, указанном в параметре *Napravlenie*. В результате генерируется шаблон документа, готовый для дальнейшего заполнения (рис. 4).

	A	B	C	D
1	FIO	SEM	4-ПРОФЕСІЙНА ІНОЗЕМНА МОВА-ЕКЗ	4-ФІЗИКА-ЕКЗ
2	ЄСЄВ МИХАЙЛО СЕРГІЙОВИЧ		95	83
3	БОНДАР АРТЕМ ОЛЕКСАНДРОВИЧ		77	66
4	ВАРУХА ОЛЕГ ГЕННАДІЙОВИЧ		92	71
5	ГУЩІН МИКОЛА МИКОЛАЙОВИЧ		69	83
6	ДАВИДОВ МИХАЙЛО ОЛЕКСАНДРОВИЧ		71	98
7	ДРОБОТ АЛЬОНА СЕРГІЇВНА		83	66
8	ЄРМОЛАЄВ ЮРІЙ МИКОЛАЙОВИЧ		93	80
9	ІВАНОВ ВІТАЛІЙ ВАЛЕРІЙОВИЧ		72	65
10	КОСТИЛЄВ ВАДИМ ІГОРОВИЧ		61	92

Рис. 4. Фрагмент шаблона ЭД

Шаг 5. После внесения пользовательских данных содержимое ЭД переносится в БД ИС. В результате чего появляется соответствующая таблица-рис. 5.

	fio character varying	predmet character varying	estimate character
1	ЄСЄВ МИХАЙЛО СЕРГІЙОВИЧ	4-ПРОФЕСІЙНА ІНОЗЕМНА МОВА-ЕКЗ	95
2	ЄСЄВ МИХАЙЛО СЕРГІЙОВИЧ	4-ФІЗИКА-ЕКЗ	83
3	ЄСЄВ МИХАЙЛО СЕРГІЙОВИЧ	4-ОБ`ЄКТНО-ОРІЄНТОВАНЕ ПРОГРАМУВАННЯ-ЕКЗ	62
4	БОНДАР АРТЕМ ОЛЕКСАНДРОВИЧ	4-ПРОФЕСІЙНА ІНОЗЕМНА МОВА-ЕКЗ	77
5	БОНДАР АРТЕМ ОЛЕКСАНДРОВИЧ	4-ФІЗИКА-ЕКЗ	66
6	БОНДАР АРТЕМ ОЛЕКСАНДРОВИЧ	4-ОБ`ЄКТНО-ОРІЄНТОВАНЕ ПРОГРАМУВАННЯ-ЕКЗ	64

Рис. 5. Фрагмент таблицы БД ИС

Второй метод основан на описании структуры документа с использованием технологии *XML*. *XML* - текстовый формат, предназначенный для хранения структурированных данных, обмена информацией между программами. Цель создания *XML* – обеспечение совместимости при передаче структурированных данных между разными системами обработки информации, особенно при передаче таких данных через Интернет [4].

Основой похода служит язык разметки *XML*, тегами которого описывается структура документа с табличной структурой. На рис. 6 показан абстрактный пример описания ЭД с табличной структурой.

```
<doc>
        "название"
        <head0>
                "константная часть заголовка $переменная$"
        </head0>
        <info>
                <head1>
                        "константная часть заголовка $переменная$"
                </head1>
                <tab_info>
                        "название таблицы $переменная$"
                </tab_info>
        </info>
        <conclusion>
                "вывод"
        </conclusion>
</doc>
```

Рис. 6. Абстрактное описание ЭД языком *XML*

Для описания более сложных документов может понадобиться следующий набор тегов, значения которых выведено в табл. 2.

Таблица 2. Значение тегов описания ЭД

Тег	Значение	Описание
uniteTable	no/yes	Объединять или не объединять таблицы в документе
changeHeadName	старое значение:=новое значение/*	Изменяет название заголовков таблицы
myHeadName		Определяет новые заголовки таблицы по порядку
treeTable	no/yes	При "no" ячейки таблицы будут восприниматься так как они есть, при "yes" ячейки таблицы будут анализироваться и обьеденённые ячейки разбиваться на несколько ячеек
headType	headName:=varchar(50)/*/?	Устанавливает тип заголовков таблицы
Info tree_type	true/false	Определяет порядок раз бора документа (если true, то древовидный с поддержкой вложенности тегов)

Если необходимо вставить символ конца строки в описание тега, то нужно вставить константу - "LINE_END" (без кавычек). Если необходимо, чтобы значение всего параграфа использовалось как значение, то нужно использовать ключевое слово ALL_INFO следующим образом: ALL_INFO$class#1$ALL_INFO", где "class" это название колонки в БД, а значение после символа '#', обозначает, с какого параграфа.

Общая структура документа, описанного с помощью структуры данных *XML*, показана на рис. 7.

Заголовок уровня 0			
Заголовок уровня 1			
Заголовок таблицы			
Поле 1	Поле 2	…	Поле N
Значение 11	Значение 12	…	Значение 1N
…	…	…	…
Заголовок таблицы			
Поле 1	Поле 2	…	Поле N
Значение 11	Значение 12	…	Значение 1N
…	…	…	…
Заголовок таблицы			
Поле 1	Поле 2	…	Поле N
…	…	…	…
Заголовок уровня 1			
Заголовок таблицы			
Поле 1	Поле 2	…	Поле N
Значение 11	Значение 12	…	Значение 1N
…	…	…	…

Рис. 7. Структура ЭД с *XML* описанием

Выводы

Построена общая технология первичного заполнения БД ИС и поддержки актуальности на этапе ее эксплуатации. Описаны методы кластеризации документов. Также была реализована задача создания шаблонной модели ЭД с табличной структурой, что значительно упрощает механизмы переноса содержимого ЭД в БД ИС. В дальнейшем планируется модифицировать представленную модель ЭД. Учитывая особенности каждого метода кластеризации, предполагается синтезировать свой собственный метод, который позволит качественно выделять в классы ЭД определенной спецификации. В симбиозе с кластеризацией необходимо будет определиться с методами классификации, которые при появлении нового ЭД позволят относить его к тому или иному кластеру. Но наибольшее внимание будет уделено механизмам согласования структур данных ЭД и структур таблиц БД ИС с целью корректного переноса информативных данных из ЭД в БД ИС. Таким образом, будет достигнута конечная цель – уменьшения затрат времени на первичное заполнение и поддержку актуальности БД ИС.

1. *Воройский Ф.С.* Информатика. Новый систематизированный словарь-справочник (Вводный курс по информатике и вычислительной технике в терминах). — 2-е изд., перераб. и доп. – М.: Изд-во Либерия, 2001. – С. 536.
2. *Task* Force on Metadata. Summary Report. // American Library Association. – 1999. – T. June.
3. *D. C. A. Bultermann.* Is It Time For a Moratorium on Metadata? // IEEE MultiMedia. – 2004. – T. Oct-Dec.
4. *Дэвид Хантер, Джефф Рафтер и др.* XML. Базовый курс = Beginning XML. – М.: Вильямс, 2009. – 1344 с.
5. *Уоссермен, Ф.* Нейрокомпьютерная техника: Теория и практика = Neural Computing. Theory and Practice. – М.: Мир, 1992. – 240 с.