

*Зеленина Галина Николаевна,
Научная библиотека
Челябинского государственного университета
Zelenina@csu.ru*

**Особенности методики формирования записей в БД MARC
(о проблемах, связанных с работой в открытых корпоративных проектах).**

Наш проект поддержан на 3 этапе программы «Автоматизация библиотек» Мегапроекта «Пушкинская библиотека». Целью программы является «оказание содействия в создании единого национального библиографического пространства». В качестве одной из основ Программы выбрано следующее положение – «Жесткое соблюдение российских и международных стандартов при описании документа и обмене библиографическими записями. В России приоритетными являются российские стандарты, при взаимодействии с зарубежными организациями – международные».

Слайд с целями

Российским стандартом создания библиографических баз данных является формат RUSMARC. Он же принят стандартом представления и обмена библиографическими записями участников программы «Автоматизация библиотек» Мегапроекта «Пушкинская библиотека».

На сайте RUSLANet представлен документ, который называется «Особенности использования формата RUSMARC в корпоративной среде». Он ориентирован на проекты конкурса «Российские корпоративные библиотечные системы», а также на российские библиотеки, планирующие использовать формат RUSMARC для организации корпоративного взаимодействия.

Слайд с видом структуры формата RUSMARC

Особенности формата RUSMARC.

Структура формата RUSMARC состоит из кодов и правил описания элементов библиографической записи, кодов и правил описания элементов записей авторитетных файлов национальных предметных рубрик, авторов (персональных и коллективных) и так называемой техники связей – специальных правил описания частей записи, связанных с авторитетными файлами. Большое значение в формате отводится индикаторам – специальным указателям механизма техники связей.

База данных MARC, которую мы все вместе создаем, согласно формату RUSMARC должна состоять из записей аналитического описания статей журнала и связанных с ними одной записи описания источника статьи – периодического издания.

Между ними должна быть установлена связь с помощью индикаторов и специальных полей связи.

Особенностью нашей базы данных является то, что создаем мы ее в формате USMARC с помощью различного программного обеспечения. Особенностью большинства программных средств, используемых в нашем проекте, является то, что записи базы данных являются чисто библиографическими. Авторитетные записи источников статей – записей названий журналов - не создаются. Значит, и библиографические записи не имеют индикаторов связи.

Слайд с особенностью нашей БД

Итак, особенности нашей базы данных:

- Библиографические записи в формате USMARC
- Отсутствие авторитетных файлов названий расписываемых периодических изданий
- Наличие собственной методики заполнения конкретных полей библиографического описания.

Слайд с перечнем причин особенностей методик

Анализ списков полей и методик их заполнения на примере с Екатеринбургом

Вопросы создания баз данных, и особенно тех, которые создаются корпоративно силами нескольких библиотек, всегда являются самыми сложными в совместной работе.

Сложности эти обусловлены несколькими причинами:

1. Организационными моментами совместной работы.
2. Составом участников, который зависит от типов библиотек, принимающих участие в совместной работе.
3. Различием в нормативной документации – методике создания и заполнения отдельных полей БД.
4. Различием в лингвистическом обеспечении при создании БД.
5. Разным опытом работы коллективов, принимающих участие в совместной работе.

Работа по анализу состава полей и методик их заполнения в базах данных нашей и екатеринбургской корпораций началась с середины июня этого года, когда А.И. Племнек в Крыму объявил о совместной работе наших двух коллективов.

Методические аспекты совместной работы

В нормативных документах каждого коллектива присутствует документ с перечнем полей и правилами их заполнения.

В коллективе «CONSENSUS OMNIUM» это «Правила заполнения полей в автоматизированной информационно-библиотечной системе (для участников Сводного электронного каталога статей). В коллективе МАРС – это «Список обязательных полей и их кодов для росписи журнальных статей». В обоих документах имеется перечень, коды и названия полей, входящих в запись, статус поля (обязательное или факультативное), приводится краткая инструкция по его заполнению.

В настоящее время проведена работа по сверке этих документов. Результаты сравнения приведены в таблицах

Показать таблицу сравнения полей. (fields.doc)

Слайд с итоговой таблицей сравнения

Таблица 1. Сравнение состава полей

N п/п	Характеристика	Consensus Omnium	МАРС
1.	Общее возможное количество полей записи БД	82	87
2.	Количество строго обязательных полей	13	14
3.	Количество обязательных полей	26	26
4.	Количество желательных (факультативных) полей	43	47
5.	Полностью совпадают коды полей	В 77 полях	
6.	Полностью совпадают типы полей	В 68 полях	
7.	Полностью совпадают методики заполнения полей	В 59 полях	
8.	Добавить коды	4 полей: 010a - ?????? 080a, 082, 246g	5 полей: 008, 246l,505a, 8561,8562
9.	Изменить код у	2 полей 998a->600a, 245b->520a	-
10.	Изменить статус поля у количества полей: Из них строго обязательных Обязательных	8	1
		4 010a - ????? 080a, 520a, 650x	1 040b
		4	0
11.	Требуется согласование методик для заполнения	В 10 полях: -> Е-бург – 040a, 856u -> МАРС - 080a, 091a, 650a, 650x, 773g, 773t Совместно - 653a, 520a	

Как видно, большинство полей в обеих базах данных совпадают. Для совместимости баз данных от нашей корпорации в перечень полей необходимо ввести следующие поля: 008 – кодированная информация, 2461 – Другая форма заглавия, где скопированы сведения из поля «Параллельное заглавие», 505a – форматированное содержание, куда копируются сведения о содержании из поля «Примечание» в CONSENSUS OMNIUM, а также индикаторы для поля «Адрес в интернет» – 856u.

По результатам голосования последние 4 поля фактически утверждены и вносятся в список полей для заполнения.

Поле 008 практически не требуется внутри-используемых нами АИБС, но очень важно для обмена данными. По нему также происходит контроль заполнения данных, в частности, когда и кем создана запись, которая выводится в статистических отчетах анализа соответствия записей форматам RUSMARC и Z39.50. Поэтому заполнять это поле строго обязательно, так как это требование ко всем записям баз данных участников программы «Автоматизация библиотек».

Для его заполнения как в АИБС «Библиотека» так и в АИБС МАРК используются шаблоны для заполнения элементов этого поля. Некоторые элементы можно заполнять автоматически. Это можно поручить конвертору при выгрузке записей или роботу.

Для 9 полей требуется согласование методик заполнения.

Например, 040a – код библиотеки-создателя записи. У нас – это кодированное поле, в Екатеринбурге – полное название библиотеки, которое заполняется групповой корректировкой через внешний словарь. Предлагается и нам использовать эту же методику, так как при выводе БД в интернет кодированное название библиотеки будет не понято пользователями.

В поле 856u – адрес в интернет по правилам необходимо заносить полностью тип протокола передачи, например, <http://www.lib.csu.ru>. Предлагается поручить контроль и дописывание типа протокола до адреса ресурса, если он отсутствует, также роботу. Основное отличие в заполнении этого адреса у нас и Екатеринбурга то, что мы выводим адрес журнала в интернет, а Екатеринбург – полный адрес ресурса, выводящий на его полный текст.

При заполнении 5 полей предлагается Екатеринбургу использовать нашу методику. Ввести строго обязательное поле 080a – индекс УДК, 520a – аннотацию, 650x – предметную подрубрику. В группе 773 в название источника мы вносим полное название журнала из внешнего словаря, Екатеринбург допускает сокращения. Нам кажется, имеет смысл при составлении паспортов изданий брать название журнала с его обложки. При составлении описания выходных данных (номер журнала, том, страница), Екатеринбург допускает пробелы в поле. Это также требует согласования.

При выработке методики заполнения полей «Ключевые слова» и «Аннотация» требуется совместная работа.

При предварительной беседе накануне нашего круглого стола выяснилось, что практически все поля для совместной работы согласованы.

Проблемы создания и использования нормализованной лексики.

Вопросы лингвистического обеспечения совместно создаваемых баз данных являются самыми сложными. Ведь основная задача обоих наших проектов – создание качественного информационного продукта, который бы удовлетворял запросы пользователей каждой библиотеки-участницы.

На первый взгляд кажется, что достаточно соблюдать общепринятые форматы, стандарты и все будет в порядке. Но как показывает опыт, в совместной работе даже внутри одной корпорации, возникают существенные разночтения и толкования одних и тех же положений, инструкций, методик.

Одной из причин таких толкований - различный контингент обслуживаемых в библиотеках читателей, которые приходят с разными уровнями запросов и требуют разного уровня их выполнения. Сказывается также личный опыт библиотекарей, занимающихся информационным обслуживанием. Немаловажную роль играет также объем и качество информационных материалов, которые находятся в распоряжении библиотекарей для выполнения запросов пользователей.

Коллектив «CONSENSUS Omnium» весь информационный материал расписываемых изданий предметизирует рубриками ББК, количество которых невелико и определяется основными разделами таблиц для областных библиотек. Количество рубрик – 48. Затем содержание уточняется набором ключевых слов, для которых разработаны «Краткие правила заполнения». В них перечислено, что является ключевыми словами, на основе чего и как они формулируются, в каком падеже и числе они должны вводиться. Приводятся указания для различного заполнения ключевых слов (собственных имен, исчисляемых и неисчисляемых существительных, аббревиатур, и т.д.). Количество обязательных ключевых слов в записи должно быть не менее 5, достаточных для описания содержания статьи.

В нашем проекте пока должно быть использовано не менее 3 ключевых слов. Для его заполнения также используются краткие правила, где перечислены аналогичные пункты. Тем не менее, расхождения в заполнении этого поля существуют. Наверное, согласование методики заполнения этого поля, а также аннотации, будут самыми сложными.

Хорошо известно, что при создании баз данных лингвистическое обеспечение является основой эффективного поиска в них. В обоих коллективах, как в «CONSENSUS Omnium», так и в МАРС эти вопросы выделены в нормативных документах. Нормативное заполнение полей, которые составляют основу поискового аппарата в обеих базах данных, закреплено документально и является основой для проверки качества росписей, поступающих от библиотек.

В корпорации «CONSENSUS Omnium» нормативно заполняемыми являются поля 040a – создатель записи, код языка каталогизации и язык текста (040b, 041a, 041h), индексы ББК (091a), 650a основная рубрика, 700e – роль лиц, 7004 – код авторского отношения, 773t – название источника. Информация для заполнения этих полей берется из, так называемых, внешних словарей, которые готовятся корпоративно и предоставляются всем участникам.

В МАРСе количество нормативно заполняемых полей практически столько же. Дополнительными к перечисленным являются индексы УДК, Дьюи (080a, 082a) и названия подрубрик (650x).

Слайд со списком словарей и откуда они готовятся

Для заполнения нормативных полей целесообразно использовать так называемые внешние словари. Сейчас методический совет предлагает следующие нормативно используемые словари:

- Словарь кодов журналов
- Словарь названий журналов
- Словарь кодов библиотек с их названиями
- Словарь названий рубрик
- Словарь названий подрубрик
- Словарь индексов УДК
- Словарь индексов ББК
- Словарь индексов Дьюи

Они готовятся методическим советом из 2 нормативных документов: файла - списка журналов и закрепленных за ними библиотек, и файла – рубрикатора. Сейчас имеется инструкция для подготовки необходимых файлов словарей для всех типов программного обеспечения, а также их подключения к системам. Эта инструкция вместе со словарями высылается в список рассылки после утверждения нормативных документов.

Имеет смысл провести исследование о целесообразности создания и использования нормализованных словарей ключевых слов, гео и хроно - рубрик. Предлагается контроль нормативно заполняемых полей поручить роботу. В таблице сравнения полей указаны поля, для которых следует проводить робот-контроль... (fields.doc)

Двусторонний анализ БД

Кроме анализа нормативных документов проведена двусторонняя работа по оценке качества заполнения баз данных.

Для сравнения были взяты росписи 65 номеров 12 наименований научных журналов за 2001 год. При анализе баз данных учитывались полнота росписи издания по сравнению с его оглавлением, а также средние данные по записям: общая длина, количество полей, средние данные по основным поисковым элементам. Отдельно анализировались данные по поисковым возможностям записей - предметным рубрикам, каталогизационным индексам и ключевым словам. Результаты сравнения представлены в таблицах, приведенных ниже.

Слайд с таблицей сравнения БД

Таблица 2. Сравнение полноты баз данных

N П/п	Сравниваемые издания	Количество статей	
		БД MAPC	БД Consensus Omnium
1.	Банковские технологии.-2001.-NN1-11.	278	145
2.	Библиотекведение.-2001.-NN4-6.	85	43
3.	Вестник Московского университета. Серия 6. Экономика.-2001.-NN1-6.	50	46
4.	Вестник Московского университета. Серия 7. Философия.-2001.-NN1-6.	50	59
5.	Вестник Московского университета. Серия 8. История.-2001.-NN1-6.	47	46
6.	Вестник Московского университета. Серия 9. Филология.-2001.-NN1-6.	132	103
7.	Вестник Московского университета. Серия 10. Журналистика.-2001.-NN1-5.	60	56
8.	Вестник Московского университета. Серия 11. Право.-2001.-NN1-6.	54	52
9.	Вестник Московского университета. Серия 12. Политические науки.-2001.-NN1-6.	51	48
10.	Вестник Московского университета. Серия 13. Востоковедение.-2001.-NN1,3-4.	26	21
11.	Вестник Московского университета. Серия 14. Психология.-2001.-NN1-4.	41	38
12.	Вестник Московского университета. Серия 18. Социология и политология.-2001.-NN1-4.	56	46
	Итого статей:	930	703

Слайд с таблицей сравнения полноты росписей

Таблица 3. Сравнение полноты росписей изданий

N п/п	Сравниваемые издания	Количество статей		
		По оглавлению журнала	БД МАРС	БД Consensus Omnium
1.	Банковские технологии.-2001.-NN 1-11.	288	278	145
2.	Библиотечное ведение.-2001.-N 5.	32	30	19
3.	Вестник Московского университета. Серия 6. Экономика.-2001.-N 2.	9	9	9
4.	Вестник Московского университета. Серия 7. Философия.-2001.-N 2.	9	9	10
5.	Вестник Московского университета. Серия 8. История.-2001.-N 1.	11	11	9
6.	Вестник Московского университета. Серия 9. Филология.-2001.-N 2.	25	24	15
7.	Вестник Московского университета. Серия 10. Журналистика.-2001.-N 5.	7	8	8
8.	Вестник Московского университета. Серия 11. Право.-2001.-N 2.	11	10	11
9.	Вестник Московского университета. Серия 12. Политические науки.-2001.-N 6.	10	10	9
10.	Вестник Московского университета. Серия 13. Востоковедение.-2001.-N 3.	6	6	6
11.	Вестник Московского университета. Серия 14. Психология.-2001.-N 4.	9	8	8
12.	Вестник Московского университета. Серия 18. Социология и политология.-2001.-N 3.	10	10	10
	Итого статей :	139	135	114

Значительное отличие в количестве записей по сравнению с оглавлением журнала объясняется различием в нормативных документах корпораций. В CONSENSUS ONMIUN есть журналы, которые расписываются полностью, а есть и такие, статьи для росписи которых выбираются библиографами. К сожалению, признак выборки статей – неизвестен.

Были проанализированы массивы (росписи) обеих баз на перечень основных полей в записях. Результаты приведены в таблице 3.

Слайд с таблицей сравнения состава записей

Таблица 4. Сравнение состава записей в записях

Сравниваемый параметр	База данных МАРС	База данных Consensus Om- nium
Среднее кол-во полей в записи	53.9	60.8
Кол-во предметных рубрик в записи	2.8	2.7
Кол-во ключевых слов в записи	5.6	7.3
Индексов в записи	1.1	1
Процент проиндексированных записей	72 %	100 %
Длина предметных рубрик (байт)	14.9	18.2
Длина ключевых слов (байт)	16.3	19.6
Длина каталогизационных индексов (байт)	5.1	3.1

Основным результатом этого сравнения является то, что базы обоих проектов практически одинаковы. Основное отличие состоит только в том, что база данных МАРС является аннотированной.

Слайд с замечаниями к записям МАРС

Основные замечания коллег из Екатеринбурга к записям корпорации МАРС.

1. Очень много орфографических ошибок, опечаток.
2. Отсутствуют единые решения по заполнению полей «Географическая рубрика», «Хронорубрика».
3. Для заполнения ключевых слов не используются словари нормализованной лексики, методика составления ключевых слов не соответствует принятой к корпорации «CONSENSUS Omnium».
4. Замечания по несоответствию ГОСТу при заполнении полей в отдельных записях (например, в некоторых записях инициалы стоят перед фамилией авторов).
5. В отдельных записях отсутствуют предметные рубрики и ключевые слова (В МАРСе ключевые слова могут не составляться на нормативные документы).

... «К несомненным достоинствам этой БД нужно отнести действительно очень полную роспись статей каждого номера, а также заполнение поля «Аннотация»...» - это цитата из отзыва библиографов Свердловской областной библиотеки.

Слайд с замечаниями к записям Е-бурга

Основные замечания коллектива МАРС к записям корпорации «CONSENSUS Omnium».

1. Отсутствие аннотаций
2. Очень крупные рубрики.
3. Отсутствие подрубрик
4. Очень короткие индексы ББК и УДК
5. Неполная роспись издания – есть пропуски в описании статей журналов, согласно оглавлению.
6. Номер и название части в некоторых записях пишутся в поле «Продолжение заглавия», а не в специальные поля группы 773.
7. Записи формируются не в строгом соответствии формата USMARC. Например, в поле ББК или в ключевых словах в данных присутствует знак - разделитель повторяющихся полей в АИБС МАРК. Это делает невозможным полноценный поиск по этим полям.

8. Поле георубрика выделена в отдельную группу, что увеличивает длину записи в целом.

Все библиотеки, ответившие на просьбу проанализировать записи, отметили следующие достоинства записей корпорации «CONSENSUS Omnium»:

- Большое количество ключевых слов
- Практически нет орфографических ошибок и опечаток.

Главными в вопросах методики заполнения полей являются, как это ни странно, не требования соблюдения стандартов и форматов, а психологические моменты. Очень важно научиться слушать, слышать и прислушиваться к мнениям друг друга, тактично соблюдать условия ведения переговоров, не считать работу своего коллектива главной только потому, что другие делают ее несколько иначе. Ведь оба наших коллектива уже доказали способность работать. Нам всем есть чему учиться друг у друга, просто надо выработать критерии независимой оценки качества работы.

Слайд с основными выводами

Итак, основные выводы.

1. При подготовке записей корпоративной базы данных каждому участнику проекта МАРС следует строго соблюдать общепринятые нормы заполнения полей.
2. При предоставлении записей в список рассылки каждой библиотеке обязательно проводить проверку записей сотрудником, который не готовил записи, на предмет выявления орфографических ошибок, а также на соответствие методики заполнения каждого поля.
3. При подготовке записей использовать машиночитаемые словари для заполнения нормативных полей.
4. Для выполнения условий к участникам 3 этапа программы «Автоматизация библиотек» и вывода сводной базы данных в интернет следует дополнить список полей для заполнения базы данных необходимыми полями. Методическому совету подготовить необходимые изменения в нормативные документы к 1 декабря.
5. Для улучшения качества базы данных попросить технический совет подготовить программное обеспечение для автоматического контроля полноты и качества составления записей при поступлении записей в базу данных робота. Техническое задание подготовить к 1 декабря и представить в список рассылки.

6. Обсудить и утвердить через список рассылки методику заполнения полей «Ключевые слова» и «Аннотация».
7. Обратиться в научно-технический совет АРБИКОНа с просьбой анализа качества базы данных МАРС и рубрикатора и высказывания замечаний специалистами ведущих библиотек России.