

**Межрегиональная аналитическая роспись статей:
методика формирования корпоративной базы данных.
Особенности работы в Научной библиотеке
Челябинского государственного университета**

В своем докладе хотелось бы остановиться на том, как мы начали работу по созданию межрегиональной корпоративной базы данных, какие проблемы решили и какие особенности существуют в нашей библиотеке. Представить немного статистических данных по нынешнему состоянию общей базы, рассказать о трудностях, которые возникли в ходе нашей совместной работы, способах их решения в библиотеке Челябинского государственного университета.

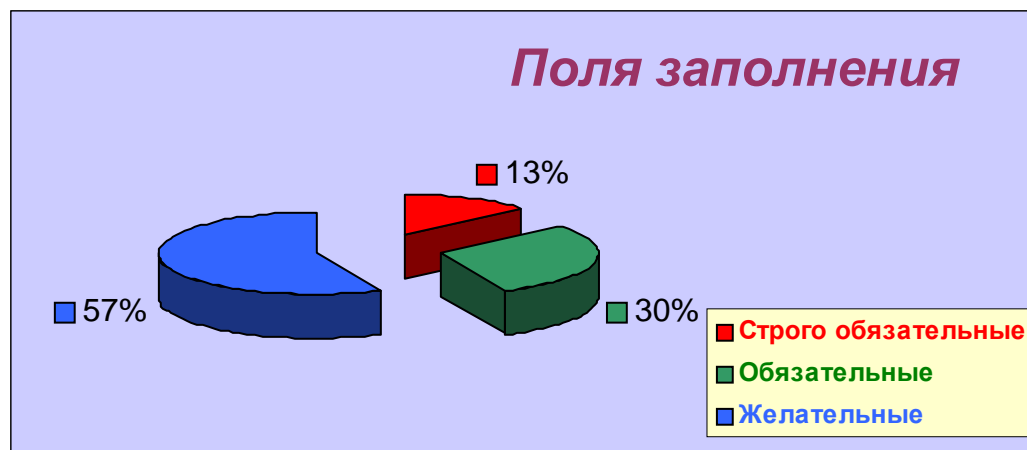
1. Состав записей

В предыдущем выступлении уже говорилось, что программное обеспечение не играет существенной роли при формировании записей корпоративной аналитической базы данных. Важно было договориться о составе полей и о методике заполнения каждого из них.

Во время предварительного этапа работы библиотеки – инициаторы проекта провели сверку форматов записей обоих программных продуктов и анализ перечня полей, необходимых для аналитической росписи периодических изданий. Был составлен список кодов полей формата US MARC для этого перечня. В результате была составлена и представлена всем участникам таблица соответствия заполняемых полей в обеих задачах и выделены три группы полей, используемых для библиографического описания периодического издания.

- **Поля, строго обязательные для заполнения.** Без их заполнения запись о расписываемой статье считается ошибочной.
- **Поля, обязательные для заполнения.** Это поля, которые обязательно надо заполнять, если в расписываемых статьях есть соответствующая информация.
- **Поля, желательные для заполнения.** Это поля, которые заполнять не обязательно, но если информация вносится, то она должна вноситься именно в указанные поля.

Общее количество полей и подполей US MARC, на основе которых создается сводная база данных, - 90. Из них **полей, строго обязательных для заполнения - 12, обязательных для заполнения полей - 27. Полей, желательных для заполнения - 51.**



2. Методика росписи.

Наша библиотека, как и все участники проекта, вместе с перечнем полей получила краткие комментарии по их заполнению. Эти комментарии были составлены на основе методических инструкций и практического опыта работы библиотек-участниц проекта.

Правила заполнения полей

В кратких комментариях к списку заполняемых полей, перечислены правила и особенности заполнения полей, регламентируются общие договоренности между библиотеками-участницами, фиксируются решения, принимаемые на основе общего голосования по каким-либо спорным вопросам. Так, например, было введено дополнительное поле – "Рубрика в журнале" (246g), отдельно выделено поле 600a – "Персоналия". Для совместимости записей в базах данных решено обязательно заполнять поля "Тип фамилии автора" и "Количество незначащих символов" в названии статьи, так как АИБС MARC более строго относится к заполнению контрольных полей в начале группы поля.

При составлении перечня полей с учетом дальнейшего развития проекта были введены поля, которые отсутствовали в стандартных настройках обеих задач. Например, было принято решение о введении в список желательных для заполнения поля 080a – индекс Дьюи и 856u – адрес URL источника.

Одним из основных условий создания корпоративной базы данных является **полная аналитическая роспись** издания.

Полнота расписываемого издания проверяется согласно его содержанию. Объем расписываемого издания регламентируется специально введенным полем 010a. В US MARC - это поле означает контрольный номер записи Библиотеки Конгресса, содержание которого в США является таким же значимым, как ISSN или ISBN.

В нашей корпоративной базе данных это поле характеризует, прежде всего, источник документа: название и номер журнала, порядковый номер записи согласно оглавлению. Все расписываемые издания получают 4-символьный идентификатор издания, который вносится в это поле. Затем идет последняя цифра года, далее проставляется порядковый номер издания в течение года. Последние 3 цифры - порядковый номер статьи. Это поле является ключевым в нашей базе данных. По нему идет сверка полноты поступающих записей от библиотек.

Заполнение поля 520a – «Аннотация» основное условие аналитической росписи в совместной работе. Аннотация призвана дать читателю краткую, точную и полную информацию о содержании статьи.

Поля из групп **обязательных** заполняются согласно нормативным документам, предоставляемых координаторами проекта. Нормативные документы уточняются в начале каждого квартала работы и получают всеми участникам через список рассылки.

К таким документам, как уже говорилось, относятся:

- **список расписываемых журналов** и их 4-символьные обозначения для заполнения поля 010a.
- **Список названий библиотек-участников проекта** и их краткого обозначения для заполнения поля автора записи - 040a.
- **Список рубрик и подрубрик с соответствующими им индексам УДК и ББК.** О списке рубрик и способах его составления уже говорилось в предыдущем докладе. Здесь же хочется отметить, что из опыта информационного обслуживания следует, что чаще всего пользователи ищут необходимую информацию не по классификационным индексам, а по темам, предметно, или по ключевым словам. Именно по этой

причине от участников проекта не требуется строго обязательного заполнения полей всех классификационных индексов.

3. Каков же результат нашей совместной работы.

В предыдущем докладе уже показывалось, как пополнялась корпоративная база данных по месяцам. Состояние базы данных на 11 октября в библиотеке ЧелГУ представлено ниже.

Корпоративная аналитическая база данных (11.10.01, НБ ЧелГУ)			
Общая статистика:	Всего записей в базе	-	16 574
	Просканировано записей	-	16 548
Длина всех записей в US MARC - 14 994 136 байт.			
Величина заполненных данных - 9 909 437 байт (66.62% MARC-записи)			
Средняя длина записи - 906.10 байт.			
Среднее количество полей в записи - 53.94 (из 90).			
Статистика заполнения полей			
	(байт)	Всего полей:	в %
Средняя длина названия издания	- 19.79	16199	98
Средняя длина таблицы для авторов	- 11.75	18787	114
Средняя длина таблицы для заглавия	- 49.46	16549	100
Средняя длина таблицы для кол. автора	- 29.98	2380	14
Средняя длина таблицы для предм. рубрик	- 14.84	42772	258
Средняя длина таблицы для кл. слов	- 15.20	94338	570
Средняя длина таблицы для кода записи	- 9.15	16446	99
Средняя длина таблицы для адреса URL	- 19.17	7667	46
Средняя длина таблицы для УДК, ББК, Дьюи	- 5.30	23705	143
Средняя длина аннотации	- 227.70	14672	89
	Всего:	В среднем в одной записи:	
Кол-во строго обязательных полей	- 187051	0.32%	(11.30 из 12)
Кол-во обязательных полей	- 110345	0.19%	(6.67 из 27)
Кол-во желательных полей	- 87108	0.15%	(5.26 из 51)
Среднее кол-во предметных рубрик в записи - 2.58			
Среднее кол-во ключевых слов в записи - 5.70			
Занесены индексы УДК, ББК, Дьюи в 72.79 % записях			
Занесены адреса URL в 46.33 % записях			

Из отчета полей видно, что не все записи в базе данных удовлетворяют требованиям проекта. Часть записей, например, не имеет названия расписываемого журнала. Отсутствует обязательное поле - код записи.

Интересны цифры по статистике так называемых информационно-содержательных полей. В каждой записи в среднем по 2.5 рубрик. Обычно бывают заполнены обязательно поля "рубрика", "под-рубрика" и одна из их разновидностей – «гео-» или «хроно-» подрубрики. Каждая запись в среднем сопровождается более, чем 5 ключевыми словами. Почти 73 % базы данных – систематизировано с помощью классификационных систем, т.е. в записях имеются индексы ББК, УДК, Дьюи.

Интересны показатели в последней строчке. Но это – тема особого выступления. Пока можно лишь сказать, что 46% наших росписей предварительно готовы для создания системы электронной доставки документа.

Мы попытались проанализировать тенденцию нашей совместной работы. Следующая таблица показывает качественное изменение записей корпоративной базы данных на протяжении трех кварталов.

<i>Изменение качества записей в БД</i>			
<i>Сравниваемые параметры (в среднем в записи)</i>	<i>I кв.</i>	<i>II кв.</i>	<i>III кв.</i>
Количество полей	48,02	51,14	51,73
Длина записи (байт)	817,55	868,06	903,03
Длина аннотации (байт)	212,26	224,33	244,76
Количество предметных рубрик	2,27	2,54	2,72
Количество ключевых слов	5,11	5,51	6,03

Как видно, с приобретением опыта работы увеличивается количество полей и объем записи, ее поисковые возможности – растет количество предметных рубрик и ключевых слов. Увеличивается объем аннотации – наиболее полно раскрывается содержание документа. Все это повышает информационную ценность росписей.

Приведенная ниже гистограмма показывает динамику соответствия групп полей в записях с течением времени. Участники проекта внимательнее стали относиться к правилам заполнения полей, от квартала к кварталу увеличивается количество строго заполняемых и обязательных полей.

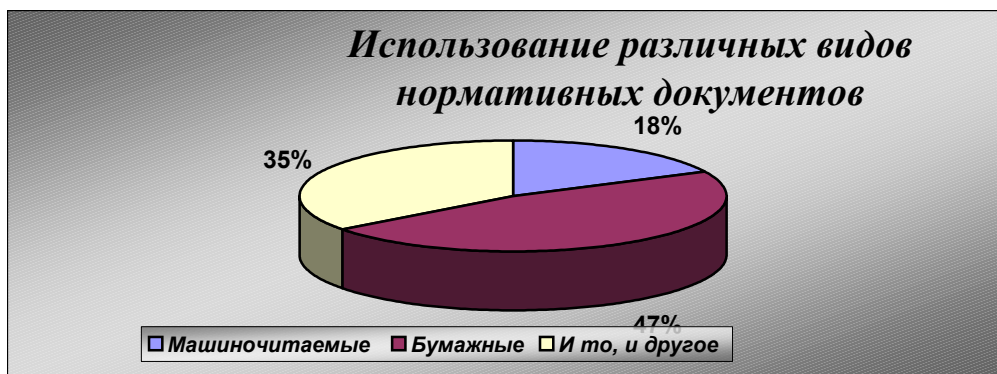
Динамика соответствия групп полей в записях



Надо сказать, что в таблице и гистограмме представлены так называемые исходные записи.

Как же происходит формирование базы данных? В разных библиотеках по-разному подходят к ее созданию, но при этом обязательно используются инструкции и нормативные документы.

После подведения итогов работы в III квартале из 17 библиотек, ответивших на вопрос о способе использования нормативных документов, только 3 из них (18%) используют машиночитаемые нормативные файлы. Это, в основном, библиотеки, в штате которых имеются программисты. Около половины (8 из 17 – 47%) заносит информацию только с бумажных носителей. 35% ответивших (6 библиотек) – используют как машиночитаемые, так и бумажные формы.



В Научной библиотеке Челябинского государственного университета, например, в процессе формирования записи используются машиночитаемые словари, подготовленные из нормативных документов. Эти словари подключены к следующим полям: коды библиотек-участниц, коды, названия журналов, индексов УДК, ББК, Дьюи, перечней рубрик и подрубрик. Их применение значительно ускоряют ввод данных, позволяют избежать ошибок ручного ввода. Чаще всего ими являются:

1. Орфографические ошибки, опечатки.
2. Различные варианты заполнения одного и того же поля.

Подробнее остановимся на втором. На примере журнала «Вестник Московского университета. Сер.13. Востоковедение». При заполнении появляются такие варианты:

Вестник МГУ. Востоковедение.

Вестник Московского университета. Сер. Востоковедения.

Следует отметить, что такие названия источника в записи поступают от одной и той же библиотеки, а бывали случаи разночтений в названии журнала в росписи одного номера. И эта ситуация характерна и для других полей.

4. Редактирование корпоративной базы данных

После полной росписи очередного номера журнала, записи выгружаются и высылаются, как уже отмечалось, всем участникам корпоративной работы.

После получения записей они распаковываются и загружаются в рабочую базу данных. Затем записи обычно просматриваются, и сотрудники библиотеки принимают решение о дальнейшем их предоставлении читателям.

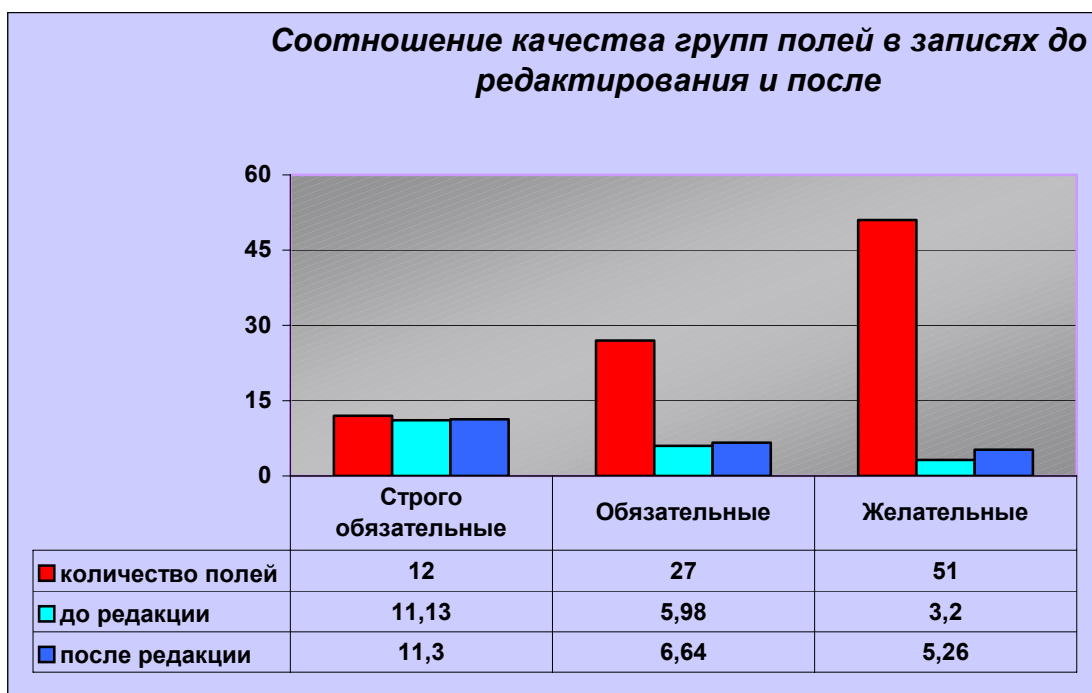
Но перед этим они редактируются. Это отметили 15 из 17 ответивших библиотек-участниц. В двух библиотеках на эту работу просто физически не хватает времени.

В нашей библиотеке редактирование записей осуществляется в три этапа:

1. При загрузке файлов в промежуточную базу данных с использованием конверторов. Основное назначение это конвертора – автоматическое исправление постоянно повторяющихся ошибок, связанных в основном с особенностями программного обеспечения системы МАРК или «хроническими» ошибками некоторых участников.
2. На втором этапе записи редактируются так же с помощью конверторов, если они содержат какие-либо особые ошибки. Например, неверно введен или совсем не заполнен код записи, название журнала, поле 040a – код библиотеки-автора записи и т.п. Возможности программного обеспечения «Библиотеки» 4, на котором работает библиотека Челябинского государственного университета, позволяют делать контекстную замену данных в определенных полях, перекодировать коды полей US MARC, добавлять данные в поля записи, если их совсем нет, удалять поля и т.д. Конверторы могут быть использованы как ко всей базе данных, так и к ее части.
3. На третьем этапе происходит визуальный просмотр полученных записей и ручное их исправление, если после двойного конвертирования записи продолжают содержать ошибки, например, опечатки.

В выявлении ошибок помогают программы, которые разработали сотрудники нашего отдела автоматизации. Эти программы значительно облегчают и ускоряют процесс редактирования базы данных, позволяют наглядно просмотреть оглавление журнала и таким образом сверить его с оригиналом. В конце каждого месяца выдается сводная ведомость полученных периодических изданий, которая формируется по строго обязательным полям: коду журнала, коду библиотеки, названию и номеру журнала.

Наглядно результаты редактирования при помощи конверторов и вручную представлены на следующей гистограмме.



Здесь видно как увеличилось число всех видов полей после редактирования.

Работа по формированию и редактированию базы данных отнимает много времени, но все это необходимо для одной цели – наиболее полного и качественного обслуживания пользователей. Как это делается в библиотеке Саратовского государственного технического университета, будет рассказано в следующем докладе.

Все же в конце выступления хочется сделать предварительные выводы по методике формирования корпоративной аналитической базы данных:

- ✓ При формировании корпоративной базы данных в каждой библиотеке используется своя традиционная методика работы. Но есть и общие правила работы, которых все участники стараются придерживаться.
- ✓ Основное внимание в корпоративной базе данных уделяется полям, раскрывающим содержание расписываемого документа. За основу систематизации базы данных взят сводный рубрикатор УДК-ББК с расшифровкой рубрик и подрубрик.
- ✓ Создаваемая корпоративная аналитическая база данных активно используется всеми участниками проекта – все библиотеки отметили важность совместной работы, ее исключительную значимость в информационном обслуживании своих пользователей.

В процессе работы, конечно же, будут возникать новые проблемы – их не бывает только у тех, кто ничего не делает. В той части, которая касается методики заполнения базы данных, они очевидны.

- Это, прежде всего, качественное лингвистическое обеспечение базы данных. Нам нужны развернутые рубрикаторы предметных рубрик и подрубрик. Это может быть принятый в качестве российского стандарта авторитетный файл предметных рубрик РНБ, либо рубрикатор, разработанный совместными усилиями. Нашей коллективной команде необходимо приобретать машиночитаемые таблицы ББК, УДК, Дьюи, адаптировать их под наши пакеты, чтобы применять в работе.
- Необходим единый словарь ключевых слов, для подготовки записей. Проблема методики его создания пока открыта.
- Очень остро встает проблема формирования библиографических записей в нашей корпоративной базе данных в связи с принятием ГОСТа 7.80-2000, касающегося заголовка библиографического описания. Этот ГОСТ принят в связи с машиночитаемой обработкой документов и ориентирован на российский стандарт MARC – на RUSMARC.
- Встанет в скором будущем проблема сотрудничества с другими корпоративными проектами, которая касается в основном, переконвертирования нашей базы данных в формат – RUSMARC.

Мы помним об этих проблемах. Но сидеть и ждать, пока их кто-то за нас решит, нет возможности - к нам обращаются читатели, им нужна оперативная информация.