Доклад на конференции в Питере

Корпоративная аналитическая база данных статей: методика формирования

В своем докладе мне бы хотелось рассказать о том, как мы начали работу по созданию нашей сводной базы данных, какие проблемы решили. Представить немного статистических данных по нынешнему состоянию базы, рассказать о трудностях, которые возникли в ходе нашей совместной работы и проблемах, которые она выявила.

1. Состав записей.

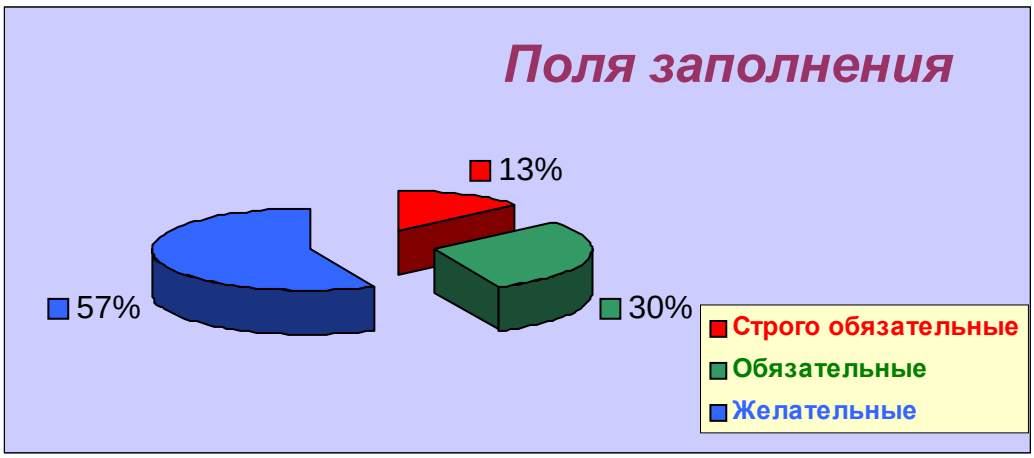
В предыдущем выступлении уже говорилось, что программное обеспечение не играет существенной роли при формировании записей корпоративной аналитической базы данных. Важно было договориться о составе полей и о методике заполнения каждого из них.

Во время предварительного этапа работы библиотеки - инициаторы проекта провели сверку форматов записей обоих программных продуктов.

Был проведен анализ перечня полей, необходимых для аналитической росписи периодических изданий, составлен список кодов полей формата US MARC для этого перечня. Записи, формируемые обоими программными пакетами, были переданы для загрузки друг другу для тестирования работы программ, формирования выходных форм. В результате была составлена и представлена всем участникам проекта таблица соответствия заполняемых полей в обеих задачах (представить таблицу). На основе такого анализа были выделены

- **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 (диаграмма 2).**



2. Методика росписи.

Все участники проекта вместе с перечнем полей получили краткие комментарии по их заполнению (показать инструкцию). Эти комментарии были составлены на основе методических инструкций и практического опыта работ библиотек - участницах проекта.

2.1. Правила заполнения - в краткой инструкции.

В кратких комментариях перечислены правила, особенности заполнения обязательных полей, регламентируются общие договоренности между библиотеками - участницами, фиксируются решения, принимаемые на основе общего голосования по каким-либо спорным вопросам. Так, например, было принято решение о заполнении названия журнала в поле с кодом 773t, 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 полей обеих задач. Например, было принято решение в введении в список желательных для заполнения поля 080a - индекс Дьюи и 856u - адрес URL источника. Некоторые библиотеки их аккуратно заполняют.

2.2. Полнота росписи.

Одним из основных условий участия в корпоративном проекте является **полная аналитическая роспись** издания.

Полнота расписываемого издания проверяется согласно его содержанию. Объем расписывамого 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 - название и номер журнала, порядковый номер записи согласно оглавлению. Все расписываемые издания получают 4-символьный идентификатор издания, который вносится в это поле (показать отрывок из инструкции по этому полю). Затем идут 2 последние цифры года, затем проставляется порядковый номер издания в течение года. Последние 3 цифры - порядковый номер статьи. Это поле является контрольным в нашей базе данных. По нему идет сверка полноты поступающих записей от библиотек.

2.3. Заполнение обязательных полей - нормативное.

Поля из групп **обязательно заполняемых полей** заполняются согласно нормативным документам, предоставленных координаторами проекта. Нормативные документы уточняются у нас в начале каждого квартала работы и предоставляются всем участникам через список проекта.

К таким документам относятся:

- **список расписываемых журналов** и их 4-символьные обозначения для заполнения поля 010a (показать список).
- **Список названий библиотек-участников проекта** и их краткого обозначения для заполнения поля автора записи - 040a (показать список библиотек)
- **Список рубрик и подрубрик с соответствующими им индексам УДК и ББК** (показать список рубрик).

2.4. Лингвистическое обеспечение росписей.

Двойная систематизация в нашем проекте не случайна. Мы специально не стали требовать от библиотек - участников проекта строго придерживаться какой-то одной системы классификации, так как в нашем проекте участвуют библиотеки разных ведомств, с различной спецификацией фондов, каждая со своим опытом работы. Мы также не требуем, чтобы участники проекта, готовя записи, обязательно делали систематизацию по всем предложенным системам.

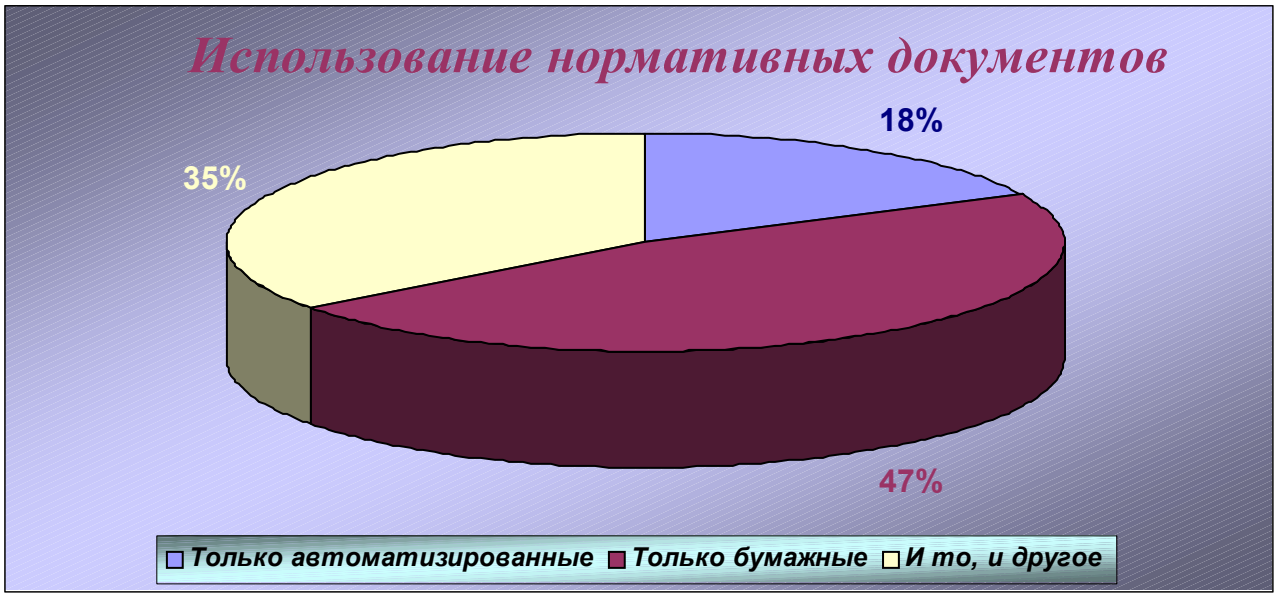
В начале организационных работ было предложено в качестве альтернативы взять классификацию Дьюи. Но, поскольку, библиотеки совершенно не имели опыта работы с ней, то следующим предложенным вариантом стал Авторитетный файл предметных рубрик Российской Национальной библиотеки. Он построен на основе новой редакции ББК. Но и эта системы не была принята в качестве лингвистической основы нашей корпоративной базы данных. Причиной этому, как уже отмечалось, являются традиции работ в конкретных библиотеках, использование при систематизации документа не только ББК, но и УДК, и ГРНТИ. И второй является устаревшая техника, на которой работают некоторые библиотеки участницы. Базу РНБ невозможно прочитать на dos-машинах.

Наша команда библиотек-участниц была бы очень признательна коллегам из Российской Национальной библиотеки, если бы они оказали нам помощь, и помогли перевести свой файл авторитетных рубрик в одноуровневый машиночитаемый список, чтобы мы могли использовать его в нашей работе.

Пока же выходом из ситуации стал обобщенный рубрикатор, построенный на основе УДК и ББК ***(показать рубрикатор)***. В ходе совместной работы ведется его постоянная доработка. Мы выносим на обсуждение всех участников предложения по дополнению и изменению, а также расширению различных уровней индексов рубрикатора. После принятия решения все участники получают его последнюю версию, чтобы использовать в работе. Кроме этого, опыт обслуживания пользователей показывает, что чаще всего читатели ищут необходимую информацию не по библиотечным индексам, а по темам, предметно, или просто по ключевым словам. Именно по этой же причине мы не требуем от участников проекта обязательного заполнения полей всех классификационных индексов.

2.5. Использование нормативных документов.

При формировании записи библиотеки-участницы используют различные способы работы с нормативными документами (показать диаграмму)



При подведении итогов работы в 3 квартале из 17 библиотек, ответивших на вопрос о способе использования нормативных документов только 3 (18%) используют автоматизированные нормативные файлы. Это, в основном, библиотеки, в штате которых имеются программисты. Около половины (8 из 17, 47%) заносят информацию только по бумажным нормативным документам. 35% ответивших на вопрос (6 библиотек) используют возможности своих АИБС по автоматическому заполнению полей, наряду с бумажными нормативными документами.

Например, в нашей библиотеке из нормативных документов формируются машиночитаемые словари, которые используются при формировании записи. Они подключены к полям : коды библиотек-участниц, коды и названия журналов, индексов УДК, ББК, Дьюи, перечней рубрик и подрубрик. Эти словари значительно ускоряют ввод данных, позволяют избежать ошибок ручного ввода, являются нормативными при проверке получаемых записей от библиотек-участниц. При изменении нормативных документов - словари актуализируются.

Заполнение остальных полей в основном регламентируется краткой инструкцией, предоставленной координаторами всем участникам проекта.

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**
Просканированно 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 - 53.97

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 этого отчета видно, что не все записи удовлетворяют требованиям проекта.

Некоторые, например, не имеют название источника, т.е. названия расписываемого журнала. Некоторые - без обязательного поля - кода записи. Следует искать и анализировать, почему некоторые записи без основного требования проекта - аннотации. Этот вопрос очень долго обсуждался на предварительном этапе работы.

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 ББУ, 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 (будут вопросы - показать наш каталог…).

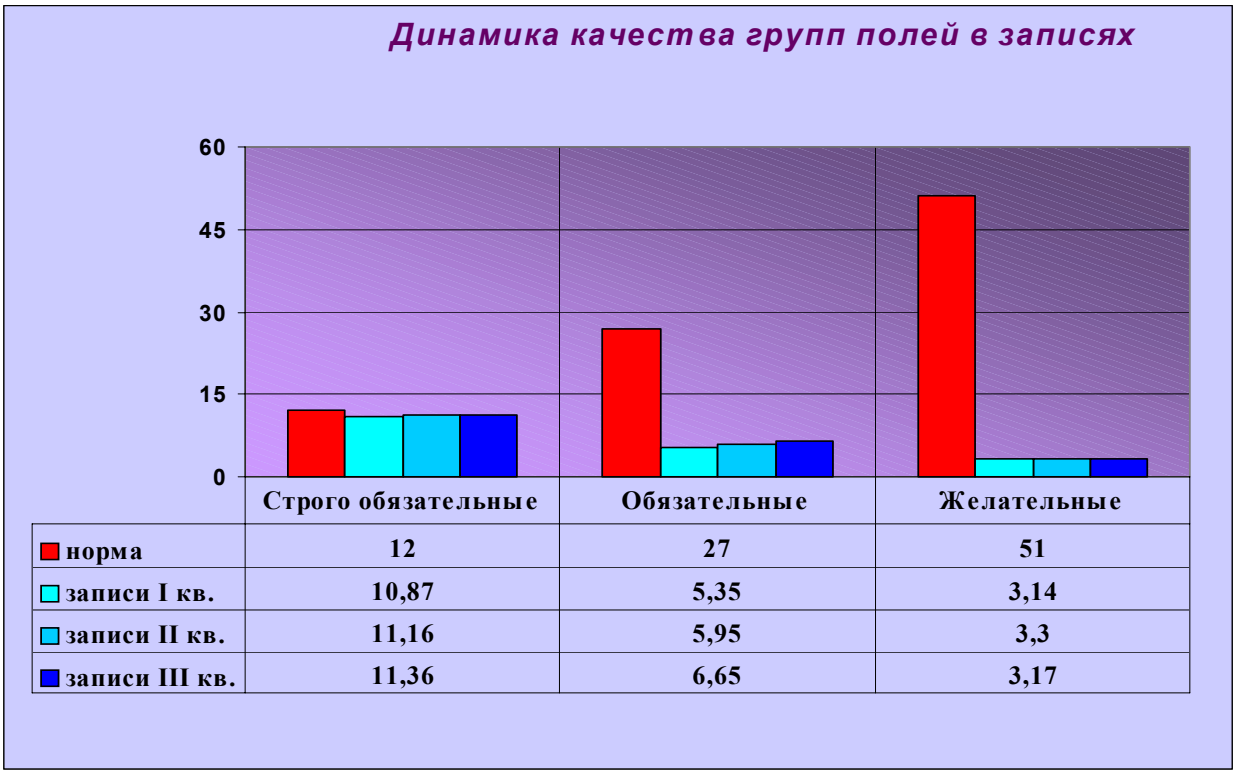
Мы попытались проанализировать тенденцию нашей совместной работы. Следующая таблица показывает качественное изменение записей корпоративной базы данных по кварталам.

Изменение качества записей в БД

Сравниваемые параметры (в среднем в записи)	***I кв.***	***II кв.***	***III кв.***
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**
Процент систематизированных записей	**65,73 %**	**69,69 %**	**76,57 %**

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

На следующей схеме показано, как изменялись группы основных полей записей.



	Строго обязательные	Обязательные	Желательные
норма	12	27	51
записи I кв.	10,87	5,35	3,14
записи II кв.	11,16	5,95	3,3
записи III кв.	11,36	6,65	3,17

Приведенная гистограмма показывает, что участники проекта внимательнее стали относиться к правилам заполнения полей, от квартала к кварталу увеличивается количество строго заполняемых и обязательных полей.

Объяснения этому - очевидны.

Во-первых, сказывается опыт работы. Библиографы постепенно освоили инструкцию по заполнению полей, научились пользоваться нормативными документами. Некоторым в ходе работы были указаны типичные ошибки при заполнении. Особенно это помогает "новичкам" в период их вступления в совместную работу. В подписанном всеми нами договоре не зря прописана процедура принятия в члены нашего корпоративного проекта. Все "новички" проходят у нас испытательный тест. Перед тем, как они получат всю базу за год, каждый из "кандидатов" должен предоставить на общий суд свою работу - подготовить росписи закрепленных изданий по нашей методике, выслать их в коллективный список. Все участники проекта просматривают их, некоторые высказывают свои замечания по качеству записей авторам росписи.

Общение в нашем списке рассылки - не формально. Это внимательное отношение к проблемам друг друга, тактичное указание на ошибки, оперативная помощь в конкретном деле. Мы, практически, не знаем друг друга лично, у каждого из нас сформировался собственный стиль работы, своеобразный имидж. Иногда, по обратному адресу электронной почты получаемых записей решается, что от этой библиотеки записи можно не проверять, а к другой "посылке " следует отнестись немного внимательнее.

Во-вторых, за время коллективной работы не только библиографы накопили опыт работы. Программисты библиотек - участниц подготовили из текстовых файлов нормативных документов машиночитаемые словари, встроили и адаптировали автоматизированные средства нормативной лексики в программные пакеты. Наиболее "автоматизированно -

продвинутые" библиотеки для редактирования получаемой базы используют конверторы. Все это безвозмездно передается всем нуждающимся.

По итогам работы за квартал многие из нас высказываются о том, кто как работает в своей библиотеке. Кто-то просто читает эти отчеты и принимает информацию к сведению, а кто-то просит совета и обязательно его получает. Мы - очень разные, но, если работаем вместе, то стараемся помогать друг другу.

В предыдущем докладе уже говорилось о том, как мы думаем работать дальше. Хочется повторить, что все участники проекта 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, как уже упоминалось, принятый в качестве российского стандарта авторитетный файл предметных рубрик РНБ, либо рубрикатор, разработанный совместными усилиями. Необходимо приобретать машиночитаемые таблицы ББК, УДК, Дьюи, адаптировать их под наши пакеты, чтобы применять в работе.

Необходим единый словарь ключевых слов, из которого бы выбирались слова при подготовке записей. Проблема методики его подготовки пока открыта.

Очень остро встает проблема формирования библиографических записей в нашей корпоративной базе данных 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 RUS MARC.

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.

Таким образом, по результатам совместной работы можно сделать следующие выводы:

- **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.**

- **Общая работа в проекте поднимает профессиональный уровень всех ее участников, библиографов, программистов, пользователей.**

- **Совместная работа порождает новые проблемы, которые мы все сообща может разрешить.**