

**Машиночитаемая библиографическая MARC-запись
и математические модели Ципфа и Брэдфорда**

**Machine-Readable Bibliographic MARC Records
and Mathematical Modeling After Bradford and Zipf**

**Машиночитаний бібліографічний marc-запис
і математичні моделі Ципфа і Бредфорда**

М. Л. Халабия

Московский государственный университет культуры и искусств, Москва, Россия

Mariya Khalabiya

Moscow State University of Culture and Arts, Moscow, Russia

М. Л. Халабія

Московський державний університет культури та мистецтв, Москва, Росія

В докладе раскрывается применение математических законов Ципфа и Брэдфорда для анализа описательных полей библиографической записи в MARC-формате.

The paper is devoted to the use of Zipf and Bradford's laws for the analysis of the descriptive fields in a bibliographic MARC record.

У доповіді розкривається вживання математичних законів Ципфа та Бредфорда для аналізу описових полів бібліографічного запису в marc-форматі.

В «Библиотечной энциклопедии», изданной Российской государственной библиотекой в 2007 году, приводится следующее определение понятия «библиографическое описание», которое понимается как «часть библиографической записи, совокупность библиографических сведений о документе, приведенных по определенным правилам, устанавливающим порядок следования областей и элементов и предназначенных для его идентификации и общей характеристики. Оно представляет собой библиографическую модель, оформленную в виде системы взаимосвязанных элементов с предшествующей опознавательной пунктуацией». [1] В свою очередь, библиографическая запись, в состав которой входит библиографическое описание, является элементом библиографической информации, фиксирующей в документальной форме сведения о документе, необходимые для его идентификации и раскрытия содержания в целях библиографического поиска. [2]

MARC-формат, созданный в 1960-х годах, изменил представление о библиографической записи и технологии ее создания. Аббревиатура MARC понимается как машиночитаемая каталогизация и является обозначением коммуникативного формата. Следовательно, под библиографической записью в коммуникативном формате мы можем понимать «совокупность полей, включающую маркер записи, справочник, и поля данных, описывающих одну или несколько библиографических записей как одно целое» [3].

В информатике и математической лингвистике широко известны такие математические модели как законы Ципфа и Брэдфорда, с помощью которых мы хотели бы показать, что при библиографическом описании документов следует, более внимательно, относиться к полям описательных данных MARC-форматов, имеющих огромное значение для пользователя.

Для участия в нашем исследовании были отобраны машиночитаемые библиографические записи, которые создаются в национальных библиотеках Российской Федерации и описывают разные типы и виды документов. Записи, представленные в ЭК РГБ описаны при помощи формата для библиографических данных MARC 21; соответственно в ЭК РНБ они (записи) представлены в российском коммуникативном формате RUSMARC. При изучении библиографических описаний крупнейших библиотек РФ четко прослеживаются две тенденции: с одной стороны, применение при библиографическом описании формата MARC 21, созданного Библиотекой Конгресса США; с другой – использование в практике каталогизации российского коммуникативного формата RUSMARC. Сложившаяся ситуация сказывается на остальных российских библиотеках, которые при внедрении в практику работы новых информационных технологий, ориентируются на РГБ и РНБ, в том числе, и в вопросах каталогизации документных ресурсов. Таким образом, российские

библиотеки используют при создании библиографических записей как формат MARC 21, так и формат RUSMARC. Словом, «битва форматов» представления элементов библиографических данных продолжается. Поэтому нами была сделана выборка из записей, содержащих элементы описательных данных как формата MARC 21, так и RUSMARC.

Под выборкой будем понимать множество случаев (испытуемых, объектов, событий, образцов), которые с помощью определенной процедуры выбраны из генеральной совокупности для участия в исследовании [4].

Было отобраны записи из ЭК РНБ 120 записей в формате RUSMARC, и столько же записей из каталога РГБ в формате MARC 21. В свою очередь, под опытом понимается воспроизведение какого-либо комплекса условий для наблюдения испытуемого явления [5]. В нашем случае под опытом необходимо понимать библиографическую запись, описанную с помощью форматов MARC 21 и RUSMARC. Основным условием для наблюдения выступает частота использования описательных полей библиографической записи при создании библиографического описания. Количество событий соответствует 120 БЗ в RUSMARC и 120 записям – MARC 21. Критерием для отбора событий служил вид описываемого документа.

Таблицы распределенных выборок полей форматов MARC 21 и RUSMARC могут являться вариационными рядами, упорядоченными по убыванию частоты использования полей формата.

Таблица 1

ВАРИАНЦИОННЫЙ РЯД ОПИСАТЕЛЬНЫХ ПОЛЕЙ ФОРМАТА MARC 21

001 – Контрольный номер записи	120
005 – Дата и время последней транзакции	120
008 – Элементы данных фиксированной длины	120
245 – Область заглавий и Сведений об ответственности	120
041 – Код языка	120
040 – Организация-создатель записи	119
260 – Область выходных данных	118
300 – Область количественной характеристики	113
084 – Индексы других классификаций / Индексы ББК	105
017 – Номер регистрации авторского права или обязательного экземпляра	100
852 – Местонахождение единицы хранения	78
650 – Тематическое понятие как добавочный поисковый признак	74
003 – Принадлежность контрольного номера	72
035 – Контрольный номер системы	69
100 – Имя лица как основной поисковый признак	63
020 – ISBN	56
080 – Индекс УДК	35
700 – Имя лица как добавочный поисковый признак	35
500 – Примечание общего характера	31
856 – Местонахождение электронного ресурса и доступ к нему	27
504 – Примечание о наличии библиографических перечней и ссылок	25
490 – Область серии	22
653 – Неконтролируемые КС	19
710 – Наименование коллектива / постоянной организации как добавочный поисковый признак	18

007 – Элементы данных фиксированной длины для физической единицы	18
651 – Географическое название как добавочный поисковый признак	15
773 – Поисковый признак на основную единицу, составной частью которой является описываемый материал	15
044 – Страна публикации/изготовления	14
720 – Неконтролируемое имя / наименование как добавочный предметный поисковый признак	13
246 – Вариант заглавия	12
130 – Унифицированное заглавие как основной поисковый признак	12
034 – Кодированные картографические математические данные	9
534 – Примечания об оригинале	8
546 – Примечание о языке	8
505 – Форматированное примечание о содержании	7
520 – Резюме, аннотация, реферат	7
250 – Область издания	7
255 – Математическая основа картографического произведения	7
541 – Непосредственный источник получения, приобретения	6
538 – Примечания о системных характеристиках и требованиях для ЭР	6
256 – Характеристики электронного ресурса	5
540 – Примечание о правах на использование и воспроизведение	5
506 – Ограничения на доступ к материалу	4
355 – Управление защитой информации от несанкционированного доступа	4
355 – Управление защитой информации от несанкционированного доступа	4
830 – Унифицированное, условное заглавие как добавочный поисковый признак на серию	3
254 – Форма изложения нотного текста	2
555 – Примечания о кумулятивном указателе / вспомогательных указателях	2
022 – ISSN	2
730 – Унифицированное, условное, обобщающее заглавие	1
440 – Область серии / добавочный поисковый признак на заглавие серии	1
247 – Предыдущее / прежнее заглавие продолжающегося ресурса	1
550 – Справка на издающий коллектив	1
774 – Поисковый признак на составную часть	1
776 – Поисковый признак на единицу в другой физической форме	1
048 – Средства исполнения	1
730 – Унифицированное, условное, обобщающее заглавие как добавочный поисковый признак	1
043 – Код географического региона	1

240 – Условное заглавие m	1
340 – Физический носитель	1
110 – Наименование коллектива / постоянной организации как основной поисковый признак	1
306 – Продолжительность проигрывания / время звучания ⁷	1
501 – Примечание о наличии в одной физической единице нескольких библиографических объектов/Владельческий или издательский конволют	1
580 – Справка о связи описываемой единицы с другими материалами	1
242 –Перевод заглавия каталогизирующей организацией	1
610 – Наименование коллектива / постоянной организации как добавочный предметный поисковый признак	1
787 – Поисковый признак на единицу, связанную с описываемой прочими отношениями	1

Затем были применены такие математические модели как законы Ципфа и Брэдфорда. Под законом Ципфа понимается эмпирическая закономерность распределения частоты слов естественного языка, а именно если все слова языка (длинного текста) упорядочить по убыванию частоты их использования, то частота n -го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n (так называемому рангу этого слова). Например, второе по использованности слово встречается примерно в два (2) раза реже чем первое, третье – в 3 раза ниже чем первое [6]. Исходя из его формулировки, мы можем представить описательное поле библиографической записи как отдельную лексическую единицу, созданную при помощи библиографического языка. Так, если обратить внимание на выборку записей, то можно заметить, что существует ряд полей, которые имеют достаточно высокую частоту использования в БЗ. Следует отметить, что упорядоченность задается ранжированием (порядком размещения) наименований элементов по частоте их появления в порядке ее убывания. Такая упорядоченная совокупность наименований полей формата MARC, которая используются в записи элементов, являются ранговым распределением. Распределения, которые изучал Ципф – это типичные примеры ранговых распределений. Оказалось, что вид рангового распределения, его строение характеризуют ту совокупность документов, к которой относится данное ранговое распределение. Вместе с тем, закон Брэдфорда является специфическим случаем распределения Ципфа для периодических изданий по науке и технике. Он сформулирован следующим образом: «эмпирическая закономерность распределения публикаций по изданиям, согласно которым в списке научных журналов, расположенных в порядке убывания числа статей по заданному вопросу, можно выделить три зоны, содержащие равное число статей по заданному вопросу. При этом первая зона – профильные журналы, непосредственно посвященные заданному вопросу; Вторая зона – журналы, частично посвященные заданному вопросу; Третья зона – журналы, тематика которых далека от заданного вопроса [7].

Несмотря на внесенные поправки, модель Брэдфорда не отражает разнообразие реальных распределений. Это несоответствие объясняется тем, что Брэдфорд сделал свои выводы, основываясь на выборе массивов, относящихся только к узким тематическим областям.

Однако указанная математическая модель наряду с моделью Ципфа является одним из важных элементов нашего исследования. Он нам необходим для того, что выявить зоны рассеивания полей библиографических записей и предположить, что существуют группы рассеяния описательный полей библиографической записи по определенным зонам.

Данные статистические ранговые распределения можно представить в виде обычной гистограммы, которую можно аппроксимировать непрерывной убывающей кривой распределения. Для его большей наглядности построим график зависимости $\ln p_r = f(\ln r)$, где p_r – относительная частота поля с рангом r или доля полей библиографических записей с рангом r .

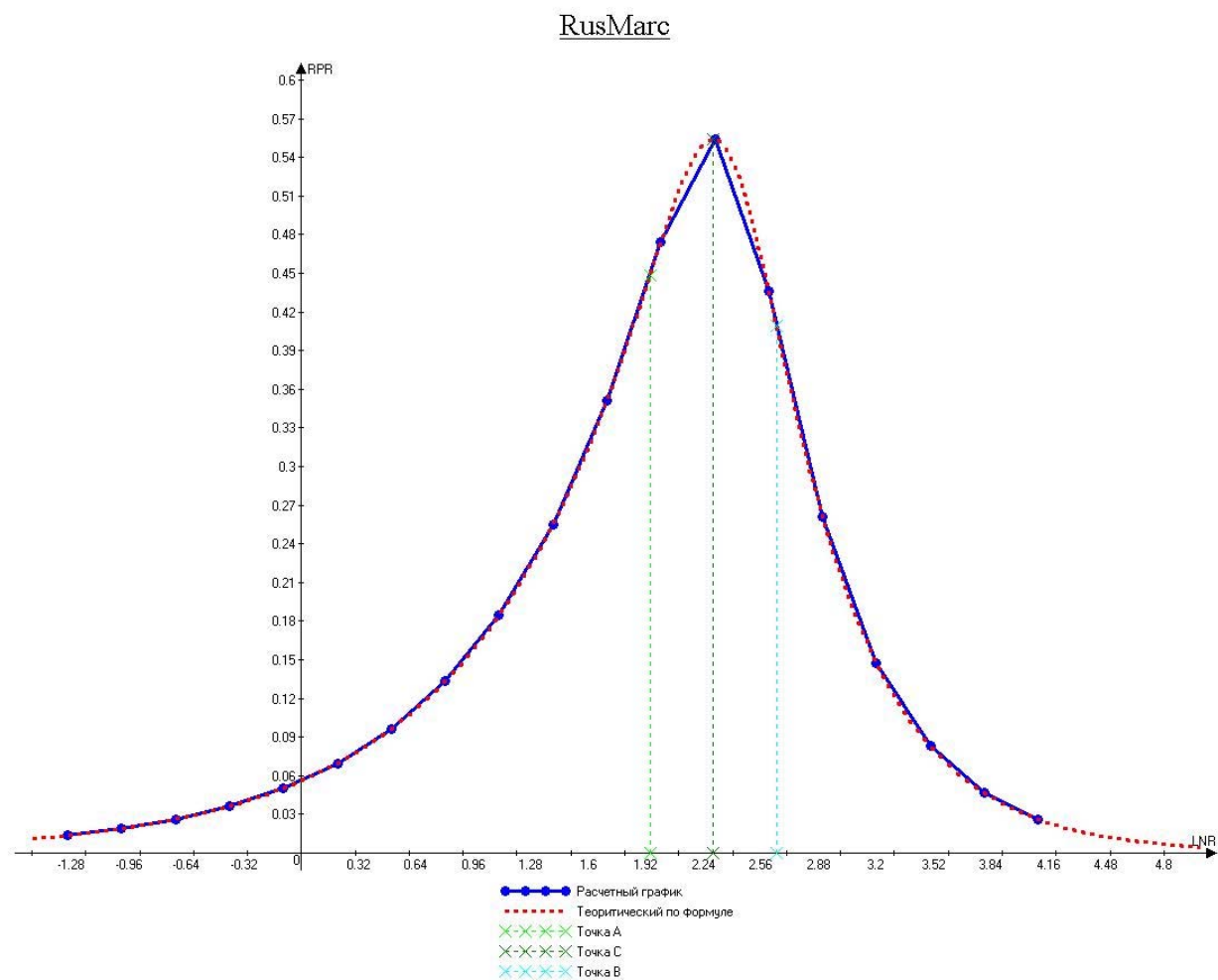


Рис. 2. График распределения полей формата RUSMARC

На представленных графиках видно, что условно можно выделить 4 зоны рассеивания описательных полей форматов MARC 21 и RUSMARC. При этом в первую группу в обоих форматах примерно около 10 полей, во вторую и третью – 10, а в четвертую – около 20 полей.

Из проведенного исследования видно, что на ядро библиографической записи приходится около 10 описательных полей. Как показывает практика каталогизации документов большая часть данных в определенных полях является очень важной для пользователя. Отношение каталогизатора к ядру библиографической записи, исходя из вышесказанного, не должно быть поверхностным. Необходимо избегать ошибок в этих описательных полях.

Литература

1. Библиотечная энциклопедия / Рос. гос. б-ка; [редкол.: гл. ред. Ю. А. Гриханов, науч. сост. е. И. Ратникова, Л. Н. Уланова и др.]. – М. осква: Пашков дом, 2007. – 1299 с.: портр., цв. ил. [1]
2. Фокеев В. А. Библиографическая наука и практика: терминологический словарь / В. А. Фокеев. – Санкт-Петербург: Профессия, 2008. – 269, [1], с. [2,3]
3. Нешиной В. В. Элементы теории обобщенных распределений: [монография] / В. В. Нешиной; Учреждение образования «Белорус. гос. ун-т культуры и искусств. – Минск: Респ. ин-т высш. шк., 2009. – 202, [1] с. [4,6,7]
4. Кибзун А. И., Горяинова Е. Р., Наумов А. И. и др. Теория вероятностей и математическая статистика: базовый курс с примерами и задачами / А. И. Кибзун, Е. Р. Горяинова, А. И. Наумов, А. Н. Сироткин; под ред. А. И. Кибзуна. – Москва: Физматлит, 2002. – 224 с. [5].