

Классификация документов библиотечных фондов применительно к потребностям процесса оцифровки

Существующие способы классификации документов и печатных изданий, отражающие важные для библиотечной и издательской сфер деятельности особенности, не всегда пригодны для систематизации содержимого библиотечных фондов при планировании работ по оцифровке. В статье делается попытка классифицировать документы по характеру наполнения различными структурными элементами с учетом плотности их размещения.

Ключевые слова: оцифровка, документ, классификация, изображение, внешний вид.

Зачем необходима еще одна классификация?

В литературе по документоведению [2, 3, 4] можно найти разнообразные классификации документов, служащие потребностям специалистов в различных областях библиотечной и издательской деятельности.

Например, фасетная классификация, приводимая в [4], описывает варианты систематизации документов, в основе которых лежат:

- способ записи информации;
- принадлежность знаков записи к определенным знаковым системам;
- форма записи информации и т. д.

В ГОСТ 7.60—2003 [1] перечислены классификации по материальной конструкции, степени периодичности, составу основного текста, целевому назначению. Кроме этого, существуют варианты классификации по характеру оформления, способу полиграфического исполнения и другие.

Однако все перечисленные типологии создавались внутри вполне определенных сфер деятельности — библиотек, архивов, издательств — и отражают рабочую специфику именно этих областей.

В то же время развивающийся уже не первый год процесс оцифровки библиотечных и архивных фондов имеет собственную специфику деятельности — в чем-то совпадающую с общелибротечной, но в чем-то заметно отличную от нее. Поэтому для более эффективного пла-



Михаил Владимирович Дагаев,
главный специалист
отдела поддержки электрон-
ных библиотек
Российской государственной
библиотеки

нирования работ по оцифровке желательно использовать такую классификацию документов, которая в большей степени будет учитывать характерные особенности этого вида деятельности.

Преобразование имеющейся систематики документов к виду, более удобному для использования при классификации продукции оцифровки

Если перечисленные выше классификации свести воедино, то входящие в них основные категории документов можно представить в следующей обобщенной форме:

1. Издания научно-исследовательского и производственно-практического характера.
2. Справочные издания (энциклопедии, справочники, словари).
3. Художественные и научно-художественные (научно-популярные) издания.
4. Издания учебного и учебно-методического характера.
5. Издания для детей и юношества.
6. Нормативно-правовые документы.
7. Издания политического характера.
8. Рекламные издания (проспекты, буклеты, плакаты, пресс-релизы, иллюстрированные каталоги, путеводители).
9. Религиозная литература.
10. Нотные издания.
11. Изоиздания (альбомы, плакаты, репродукции картин, открытки).
12. Картографические издания (карты, планы местности, атласы).
13. Периодические и продолжающиеся издания (газеты, журналы, бюллетени).

Этот список можно преобразовать к виду, более удобному для дальнейшей работы, если перечисленные категории упорядочить с точки зрения рабочих потребностей оцифровки. Для этого документы размещаются в порядке уменьшения своих типичных геометрических размеров (один из значимых параметров при оцифровке). Также вполне разумным действием будет объединение тех категорий, у которых наблюдается заметное сходство во внешнем виде страниц (тоже достаточно значимый параметр оцифровки), даже если с точки зрения общепринятой систематики они относились к различным категориям. Или же наоборот — категории, в которые входят документы, сильно отличающиеся по внешнему виду, разбить на дополнительные подкатегории.

В результате получаем следующий перечень разновидностей документов:

1. Карты, планы местности, атласы.
2. Плакаты.
3. Газеты.
4. Профессиональные журналы.
5. Популярные журналы, рекламные издания.

6. Альбомы, репродукции картин.
7. Нотные издания.
8. Справочники, энциклопедии и словари профессионального характера.
9. Справочники, энциклопедии и словари популярного характера.
10. Издания для детей и юношества.
11. Литература научно-исследовательского и производственно-практического характера.
12. Научно-художественные (научно-популярные) и учебные издания.
13. Издания политического характера, нормативно-правовые документы, художественная литература, религиозная литература, учебно-методические издания, бюллетени.
14. Открытки.

Выбор критериев для систематизации оцифрованных документов

Однако все сказанное относилось к систематизации печатных оригиналов документов. Если же говорить о них, но уже в оцифрованном виде, то следует учитывать двойную природу любых результатов оцифровки.

С одной стороны, это будут те же самые книги, журналы, карты, альбомы, просто в несколько непривычном виде. С другой стороны, теперь все это суть растровые изображения той или иной разновидности, со всей присущей им спецификой. А значит, их дальнейшее поведение во многом будет определяться закономерностями, которые характерны для работы с изображениями.

Поэтому классификация документов с точки зрения процесса оцифровки должна будет совместить в себе систематику и документов, и растровых изображений.

Растровые изображения можно систематизировать по достаточно большому количеству разнообразных признаков. Для процесса оцифровки наиболее существенное значение имеют следующие из них:

- из какого источника (с помощью какой именно разновидности сканеров или цифровых фотокамер) получено изображение;
- размер изображения (определяется геометрическим размером оригинала и разрешением сканирования);
- глубина цвета (максимально возможное число оттенков, которое можно использовать в данном изображении) — изображения могут быть цветными (Truecolor), в полутонах серого (Grayscale) и черно-белыми (Black&White или Bitonal);
- разрешение устройства, на котором получено изображение;
- формат данных, в котором записано и сохраняется изображение (один из растровых форматов общего назначения — BMP, TIFF, GIF, PNG, JPEG, JPEG 2000, иногда DjVu);
- тип содержимого.

В данном случае имеется в виду не столько то, что составляет информационное *содержание* данного изображения (текст книги, фотографии и рисунки к ним, нотная партитура, карта местности и т. д.), сколько характерные элементы структуры его внешнего вида. Линии рельефа, подписи и цветовая заливка для карты, области плавного перехода оттенков для фотографии, строки литер текста для книги и тому подобное.

Из всех перечисленных признаков только размер получаемого изображения и тип содержимого напрямую связаны с конкретным документом. Выбор источника изображения (оборудования для оцифровки), глубины цвета, разрешения сканирования, формата рабочих файлов обычно задается наиболее типичным наполнением основной массы документов, проходящих оцифровку, т. е. эти параметры характеризуют скорее усредненное состояние обрабатываемых материалов, чем конкретику каждого из них.

Что касается размера получаемого изображения, то он зависит сразу от нескольких факторов: геометрических размеров сканируемого оригинала, разрешения сканирования, глубины цвета. Поскольку последние две величины выбираются усредненно, то брать размер изображения в качестве основы для классификации будет не практично.

Таким образом, для анализа изображений с точки зрения особенностей процесса оцифровки целесообразно построить классификацию на оставшемся свойстве — что из себя представляет содержимое (внешний вид) сканируемого документа.

Построение классификационной схемы

Если проанализировать внешний вид страниц для различных типов документов, то нетрудно заметить, что их содержимое отличается заметным разнообразием верстки, художественного оформления и качества печати.

Характерное содержимое некоторых разновидностей документов

Географические карты и планы местности:

- надписи, выполненные шрифтом мелких кеглей;
- близко расположенные тонкие линии различных цветов;
- большая плотность надписей и линий на единицу площади;
- участки заметного размера, залитые одним и тем же цветом.

Плакаты:

- надписи, выполненные шрифтами больших и очень больших кеглей;
- участки заметного размера, залитые одним и тем же цветом.

Газеты и журналы:

- преобладающее содержимое страниц — текст, набранный небольшими кеглями;
- перепады в несколько десятков кеглей между самыми большими и самыми маленькими литерами;
- полутоновые рисунки (фотографии) с размерами от 1/64 до полной страницы;
- фотографии высокого, профессионального, качества.

Энциклопедии, справочники, словари:

- преобладающее содержимое страниц — текст, набранный небольшими кеглями;
- примечания, набранные предельно малыми кеглями, вплоть до 6 пунктов;
- преимущественная ориентация на черно-белые иллюстрации, с небольшим объемом цветных.

Издания научно-исследовательского и производственно-практического характера, научно-художественные (научно-популярные), художественные, учебные:

- основная масса текста набрана одним и тем же шрифтом и кеглем;
- если содержание документа носит не технический характер, могут практически отсутствовать иллюстрации.

Перспективы, буклеты:

- основная масса текста набрана небольшими кеглями;

- полутоновые рисунки (фото) с размерами вплоть до полной страницы.

Открытки:

- полутоновые рисунки и фото предельно миниатюризированного размера, в которых при этом сохраняются мелкие детали изображаемого.

Многоплановый характер содержимого может затруднить процесс классификации оцифрованных документов, поэтому, прежде чем приступить к нему, желательно как-то унифицировать разнообразие компонент внешнего вида страниц.

С точки зрения своего содержимого изображения, получаемые при оцифровке, можно условно разделить на две большие категории. К первой относятся изображения или участки изображений, где цвета идут с плавными переходами; а ко второй — изображения, где между цветами наблюдаются резкие переходы (образуются явно прорисованные границы).

Первая категория обычно обозначается как полутоновая графика. Если абстрагироваться от ее содержания и обращать внимание только на структуру, то она имеет однотипный вид — мягкие пастельные переходы между оттенками. Разница лишь в сравнительных размерах участков, заполненных тем или иным цветом, и в степени размытости/сжатости переходов «цвет—цвет».

Вторая категория не столь однородна. Ее можно разделить на четыре разновидности:

- текст — строки символов однотипного вида, примерно одинакового (в пределах одной строки) размера;

- штриховая графика — рисунки, состоящие из множества близко проведенных линий (примером может служить классическая гравюра или карандашные рисунки, выполненные в том же стиле);

- чертежи, графики, диаграммы, блок-схемы и т. п. (общепринятого названия не имеет, часто обозначается как «чертеж»);

- плакатная графика, наиболее характерная деталь — большие участки изображения, залитые одним и тем же цветом, с резкими переходами цветов между соседними участками.

Кроме перечисленного, в качестве дополнительных классификационных элементов следует учитывать плотность размещения структурных элементов в анализируемом изображении, а также необходимость разделения изображения на основную (содержательную) часть и фон.

Понятие «плотность размещения» введено для того, чтобы иметь возможность оценивать степень детализации типичных страниц тех или иных разновидностей документов — насколько они насыщены мелкими деталями, близко проведенными линиями и т. д. По степени детализации определяется разрешение, с которым необходимо производить

сканирование, и оборудование для его выполнения. Выражение «плотность размещения» обозначает:

- для текста — плотность литер в строке и строк на странице; в более общем виде определяется величиной кегля основного объема текста;

- для штриховой графики — плотность линий в рисунке;

- для «чертежа» — плотность линий на графике, диаграмме, чертеже и т. п., количество элементов «чертеж» на единицу площади изображения;

- для плакатной графики — количество областей с резкими переходами «цвет—цвет» на единицу площади изображения;

- для полутоновой графики — размер отдельных рисунков/фотографий, размещенных на странице, относительно с размером страницы.

Взяв за основу приведенную выше обобщенную классификацию документов, можно связать каждую из ее категорий с теми разновидностями структурных элементов, которые будут для нее характерны.

Полученные результаты представлены в виде таблицы.

Чем может быть полезна приведенная классификация? Как отмечалось, при планировании работ по оцифровке библиотечных фондов необходимо определиться с тем, на каком оборудовании и с какими рабочими параметрами (разрешение, глубина цвета) желательно сканировать отобранные материалы. Традиционную сложность для высококачественного сканирования представляют объекты, содержащие большое количество мелких, плотно расположенных деталей, близко проведенных линий и т. п. Характерный пример — географические карты, планы местности, сложнодетализованные гравюрные рисунки прошлых веков, особенно отпечатанные в виде книжных миниатюр.

Классификация документов по характеру наполнения различными структурными элементами (с учетом плотности их размещения) позволит заранее выделять те разновидности, которым надо уделить повышенное внимание и предусмотреть для них иные условия сканирования, чем для основного потока.

После того как выполнено сканирование, начинается следующий этап оцифровки — размещение документа на сайте. Для этого изображения преобразуются в один из форматов со сжатием (JPEG, JPEG 2000, TIFF G4, DjVu и т. д.). Однако разные виды структурных элементов могут по-разному реагировать на сжатие тем или иным форматом. Решение, вполне допустимое для книги с текстом и формулами, может оказаться не слишком удачным для технического справочника с графиками, чертежами и диаграммами. А то, что будет хорошо для современного художественного альбома, может совсем не подойти для альбома такого же типа, но выпущенного сто лет назад.

По этим причинам при работах по оцифровке имеет смысл подходить к документам с несколько

Классификация документов по входящим в них основным структурным элементам

Разновидности документов	Структурные элементы/плотность их размещения					Фон
	Графика			«Чертеж»	Текст	
	Полутоновая	Штриховая	Плакатная			
Карты, планы местности, атласы			+ / низкая	+ / высокая	+ / высокая и очень высокая	цветной
Плакаты			+ / низкая		+ / низкая	цветной
Газеты	+ / от высокой до средней		+ / от высокой до средней	+ / средняя	+ / высокая и очень высокая	белый
Профессиональные журналы	+ / средняя		+ / от высокой до средней	+ / от высокой до средней	+ / высокая	белый
Популярные журналы, рекламные издания	+ / средняя	+ / высокая	+ / средняя	+ / средняя	+ / высокая	белый, цветной
Альбомы, репродукции картин	+ / средняя	+ / высокая			+ / высокая	белый
Нотные издания				+ / средняя	+ / средняя	белый
Справочники, энциклопедии и словари профессионального характера	+ / от высокой до средней			+ / от высокой до средней	+ / высокая	белый
Справочники, энциклопедии и словари популярного характера	+ / от высокой до средней	+ / высокая		+ / средняя	+ / высокая	белый
Литература научно-исследовательского и производственно-практического характера				+ / средняя	+ / высокая	белый
Научно-художественные (научно-популярные) и учебные издания	+ / средняя	+ / высокая		+ / средняя	+ / высокая	белый
Издания политического характера, нормативно-правовые документы, художественная литература, религиозная литература, учебно-методические издания, бюллетени					+ / высокая	белый
Издания для детей и юношества	+ / средняя	+ / средняя	+ / низкая		+ / средняя	белый
Открытки	+ / от высокой до средней					отсутствует

иной меркой, чем в обычной библиотечной жизни, и основное внимание обращать не столько на то, что из себя представляет *содержание* данного издания, сколько на то, как выглядит *содержимое* страниц этого издания, из каких структурных элементов оно складывается.

Список источников

1. Издания. Основные виды. Термины и определения: ГОСТ 7.60—2003. — Введ. 2004—07—01. — М.: Изд-во стандартов, 2004. — 42 с. — (Система стандартов по информации, библиотечному и издательскому делу).
2. Столяров Ю.Н. Библиотекведение, библиографоведение, и книговедение, как единая научная специальность / Ю.Н. Столяров. — Орел: Орлов. гос. ин-т искусств и культуры, 2007. — 266 с.
3. Типология изданий. — М.: Кн. палата, 1990. — 231 с.
4. Швецова-Водка Г.Н. Общая теория документа и книги: учеб. пособие / Г.Н. Швецова-Водка. — М.: Рыбари; Киев: Знання, 2009. — 487 с.