УДК 026.06(73)
ББК 78.375.0(7Сое)

Размещение мирового культурного наследия в Интернете методом краудсорсинга*

Статья знакомит с использованием метода краудсорсинга в создании Цифровой коллекции газет Калифорнии и Публичной библиотеки Кембриджа (США). Представлены результаты исследования, проведенного в феврале—марте 2013 г. среди пользователей этих библиотек.

Ключевые слова: *газеты, цифровые коллекции исторических газет, краудсорсинг, коррекция текста, генеалогия, оцифровывание.*

Фредерик Зарндт (F. Zarndt),
председатель Секции газет Международной федерации библиотечных ассоциаций и учреждений, Колорадо, США

Брайян Гейгер (B. Geiger),
директор Цифровой коллекции газет Калифорнии, Университет Калифорнии, Риверсайд, США

Алиса Пэйси (A. Pacy),
архивист Публичной библиотеки Кембриджа, США

Стэфан Боди (S. Boddie),
менеджер-директор Консалтингового центра, Гамильтон, Новая Зеландия

В последние годы слово «краудсорсинг» (англ. crowdsourcing, crowd — «толпа», sourcing — «использование ресурсов») стало часто употребляться. Так, запрос в Интернете от 28 июля 2013 г. выдал 127 ссылок на другие страницы Wikipedia, содержащие это слово, тогда как запрос от 22 января 2010 г. выявил его упоминание только 41 раз [1]. Запрос от 28 июля 2013 г. показал 79 внешних ссылок на страницы Wikipedia, где встречается «краудсорсинг», а запрос от 5 июля 2010 г. — только 10 внешних ссылок.

Самыми большими проектами в мире с использованием метода краудсорсинга являются проекты Wikipedia и Kickstarter. Основанная в 2001 г., Wikipedia содержит более 20 млн статей на 282 языках, написанных и изданных ста тысячами волонтеров — представителями «толпы» по всему миру. Kickstarter (американский проект социальных инвестиций, «народного финансирования»), суть которого заключается в создании программ по сбору денежных средств на развитие различных инновационных проектов, с момен-

* Ф. Зарндт вновь обращается к вопросу применения метода краудсорсинга в процессе освоения мирового культурного пространства. Его предыдущую публикацию см. [1].

Перевод с английского *И.В. Чадновой*, ведущего научного сотрудника НИО библиотековедения Российской государственной библиотеки.

та своего создания в апреле 2009 г. собрал около 350 млн долл. США на развитие 30 тыс. проектов более чем от 2,5 млн членов глобальной «толпы».

На запрос, сделанный 25 января 2010 г., Wikipedia предоставила информацию о 34 краудсорсинговых проектах, в июле 2013 г. этот список вырос до 168 проектов [2], 11 из них относятся к оцифрованным книгам, журналам, рукописям, а также библиотекам.

Большим спросом пользуются оцифрованные коллекции исторических газет.

Сотрудники Цифровой коллекции газет Калифорнии (the California Digital Newspaper Collection, CDNC) и Публичной библиотеки (ПБ) Кембриджа (the Cambridge Public Library, штат Массачусетс) провели исследование среди своих пользователей.

Анкета для опроса пользователей CDNC

1. Вы используете Коллекцию в научных или личных целях (или в обоих случаях)?
2. Считаете ли вы себя специалистом по генеалогии или истории семей?
3. Каковы ваши основные цели?
4. Какой тип информации вы ищете?
5. Участвуете ли вы в работе какого-нибудь онлайнового генеалогического форума?
6. Приблизительно как часто вы посещаете сайт Коллекции?
7. Укажите, пожалуйста, число минут, которые вы обычно затрачиваете на посещение вебсайта Коллекции.
8. Используете ли вы программу коррекции текста в Коллекции?
9. Если вы ничего не знали об этой программе до того, как взяли в руки эту анкету, попытаетесь ли вы исправить текст в будущем?
10. Есть ли у вас свой аккаунт в социальных сетях?
11. Размещали ли вы когда-нибудь статью или информацию из Цифровой коллекции в социальных сетях (Twitter, Facebook)?
12. Мы надеемся улучшить программное обеспечение Veridian, используемое на сервере Коллекции. Какую из опций Вы хотели бы использовать?
13. Пожалуйста, дайте основную демографическую информацию о себе.
14. Каков Ваш возраст?

В CDNC с февраля по март 2013 г. было собрано 555 ответов на анкету, в то время как ПБ Кембриджа получила только 30 ответов. Основными пользователями коллекций в обеих библиотеках являются специалисты по генеалогии и истории семей: 82% для ПБ Кембриджа и 66% — для CDNC.

Анализ результатов опроса показал, что респондентами в ПБ Кембриджа являлись люди не моложе 30 лет, а в CDNC только менее 5% респондентов имели возраст свыше 30 лет.

На вопрос «Считаете ли Вы себя специалистом по генеалогии или истории семей?» большинство респондентов ответили утвердительно (табл. 1).

Таблица 1

Ответы респондентов на вопрос «Считаете ли Вы себя специалистом по генеалогии или истории семей?»

	CDNC	ПБ Кембриджа
Да	66,31% (368)	82,14% (23)
Нет	33,69% (187)	17,86% (5)
Всего	555	28

Распределение пользователей обеих коллекций по возрасту представлено на рис. 1.

В обоих исследованиях выяснялось, какие материалы представляют для пользователей наибольший интерес. Люди, занимающиеся генеалогией, ищут некрологи, семейные объявления (о рождении детей, свадьбах), биографическую информацию (табл. 2).

Таблица 2

Ответы респондентов на вопрос «Какой тип информации Вы ищете?»

	CDNC	ПБ Кембриджа
Биографические сведения	410	20
История сообщества	375	19
Некрологи	366	24
Объявления о свадьбе	307	20
Объявления о рождении	306	19
Судебные уведомления	276	15
Реклама	207	12

Сотрудники CDNC интересовались, пользуются ли респонденты генеалогическими службами (например, FamilySearch, Ancestry.com, RootsWeb). Результаты показали, что около 60% участников опроса пользуются либо одной, либо несколькими из них.

CDNC (http://cdnc.ucr.edu) — это самый большой, находящийся в открытом доступе архив оцифрованных газет Калифорнии. Коллекция содержит около 60 тыс. выпусков газет (с 1846 г. и по сей день) и 550 тыс. страниц текста. Проект управляется и поддерживается Центром библиографических исследований Университета Калифорнии (Риверсайд). Партнеры проекта — Национальная программа оцифровывания газет, Национальный фонд поддержки гуманитарных наук, Библиотека Конгресса США, Институт музейных и библиотечных услуг. Партнерами CDNC являются также учреждения Калифорнии, помогающие в оцифровке и добавлении контента в архив.

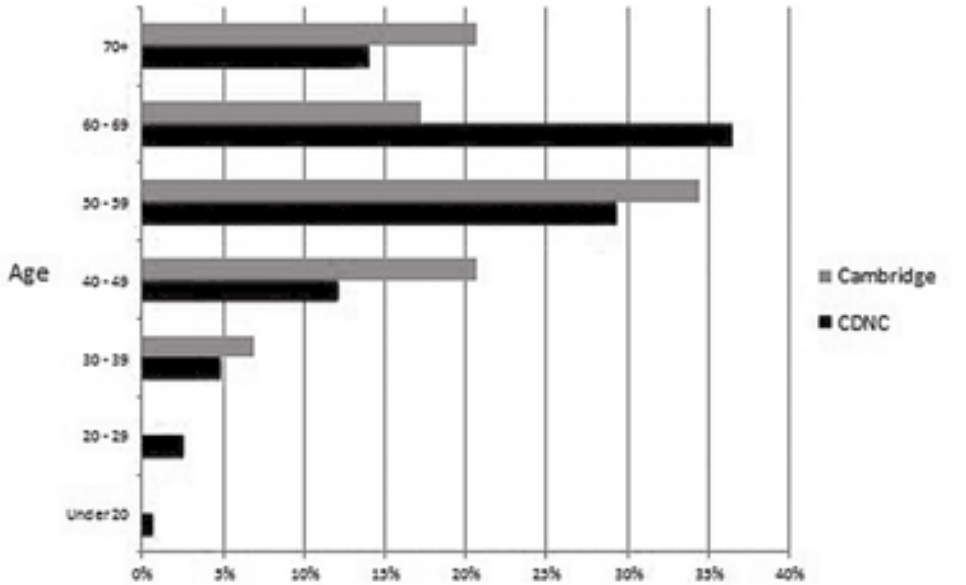

Рис. 1. Распределение (в %) респондентов по возрасту

Работы по оцифровке газет Калифорнии начались в 2005 г., когда Центр библиографических исследований стал первым партнером из участников Национальной программы оцифровывания газет. В октябре 2007 г. начал свою работу официальный интернет-сайт CDNC, с 2009 г. использующий программное обеспечение (ПО) Veridian. В августе 2011 г. сотрудники CDNC с участием разработчиков ПО Veridian обеспечили пользователям процесс коррекции текста внутри самого архива (user text correction, UTC). Это позволило им регистрироваться и затем создавать компьютерный текст. Спустя месяцы было зарегистрировано более 1300 человек, около 600 из которых исправили более одного миллиона строк текста. На сайте с самого начала поддерживается опция Google Analytics, тем самым анализируется средняя продолжительность посещений.

Понятие «гражданское архивирование» (citizen archivy) является новым для библиотек и архивов, но оно быстро стало трендом во всем мире. Гражданское архивирование предполагает вовлечение общества в создание архивных коллекций посредством Интернета, предусматривается улучшение или добавление материала к уже существующим онлайновым историческим коллекциям. Первой начала использовать этот процесс Национальная библиотека Австралии. В 2009 г. она создала свою цифровую коллекцию Trove — интерактивный портал национальных исторических документов, который совместно с Администрацией национальных архивов и записей составил единую систему поиска. Любой человек, занимающийся гражданским архивированием, используя специальную панель, может написать текст, разместить тэги и фотографии, загрузить различные картинки. ПБ Кембридж также начала развивать свои цифровые коллекции с помощью гражданского архивирования. В марте 2013 г. библиотека отмечала первую годовщину со дня основания Коллекции исторических газет Кембриджа (the Historic Cambridge Newspaper Collection, http://cambridge.dlconsulting.com) — своего единственного цифрового интерактивного проекта. Сотрудничая с организацией DL Consulting, командой инженеров-программистов и системных администраторов из Новой Зеландии, библиотека оцифровала и разместила в открытом доступе все газеты Кембриджа, свободные от авторского права, включая такую газету, как Cambridge Chronicle — самую старую еженедельную газету в США. ПО Veridian позволяет пользователям корректировать оцифрованный

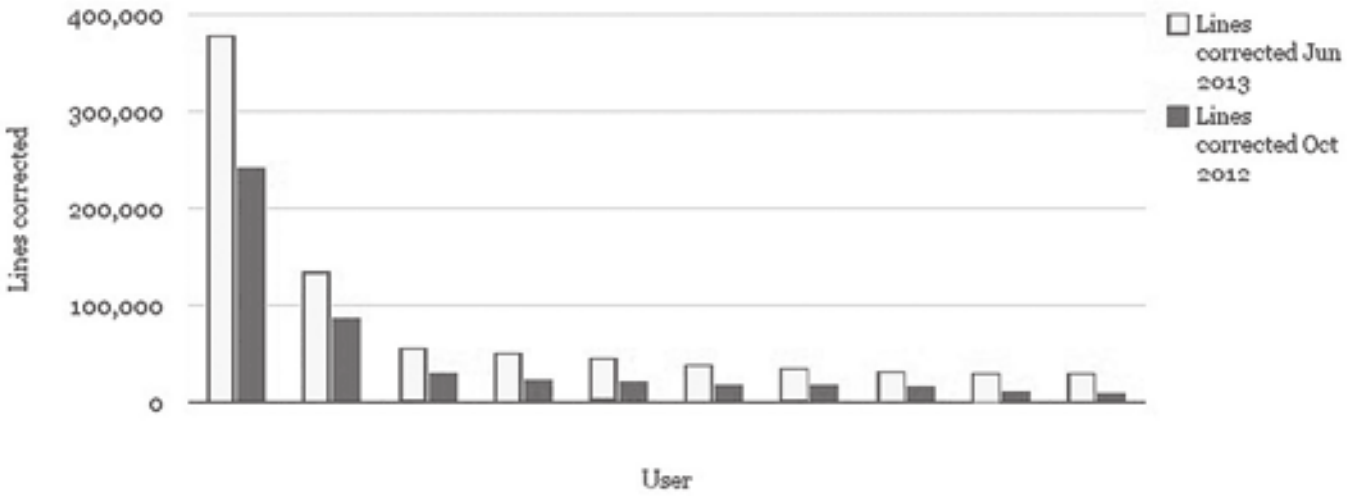

Рис. 2. Показатели продуктивности работы корректоров CDNC в октябре 2012 г. и июне 2013 г.

текст. Число строк, исправленных каждым пользователем, подсчитывается, а в специальной рубрике «Text Corrector Hall of Fame» на домашней странице указываются имена тех, кто исправил наибольшее число строк.

За год пользователи исправили 40 тыс. строк нечитаемого текста газет, их число возросло до 1 тыс. строк в неделю. Успех этого проекта показывает, что общество имеет реальное желание быть вовлеченным на местном уровне в процесс сохранения исторических ресурсов и предоставления их в пользование. Библиотека учитывает это в своей деятельности.

Комментарии в блогах волонтеров — корректоров текста в оцифрованных коллекциях газет — свидетельствуют о том, что они посвящают большую часть своего свободного времени созидательной деятельности, от которой получают пользу другие люди (в частности, в виде более точного текста статей исторических газет).

Насколько мотивированы корректоры текста? На рис. 2 показано число строк исправленного текста, отредактированного десятью разными корректорами CDNC (показатели июня 2013 г.). Некоторые корректоры чрезвычайно продуктивны, причем их продуктивность со временем не уменьшается.

Процесс коррекции OCR текста можно доверить и аутсорсинговым агентствам. Как это сделали в Австралии, Новой Зеландии, Сингапуре и Калифорнии. Но так как сам процесс является затратным, то все корректирование было ограничено газетными заголовками, или же, как например, в проекте Trove, к ним были добавлены первые четыре строчки некоторых статей. Стоимость труда волонтеров, занятых исправлением текста оцифрованных газет, представлена в табл. 3.

Таблица 3

Стоимость труда волонтеров

	Число исправленных строк	Стоимость труда волонтеров (долл. США)*
ПБ Кембриджа	43 671	873
CDNC	1 273 000	25 460
Trove	101 766 326	2 035 326

* Из расчета 0,5 долл. США за 1 тыс. знаков.

Каждый волонтер цифровых коллекций CDNC, ПБ Кембриджа, Trove нашел что-то ценное для себя в процессе коррекции текста. Вопреки трудностям измерения результатов работы, краудсорсинг в эпоху Интернета является очень простым способом взаимодействия библиотек с сообществом, которое они обслуживают.

Список источников

1. *Зарндт Ф.* Размещение в Интернете мирового культурного наследия методом краудсорсинга // Библиотековедение. — 2013. — № 1. — С. 76—84.
2. Category: Crowdsourcing [Electronic resource]. — Mode of access: http://en.wikipedia.org/wiki/Category:Crowdsourcing
3. Crowdsourcing the world's cultural heritage: Part II [Electronic resource]. — Mode of access: http://www.ifla.org/files/assets/newspapers/Singapore_2013_papers/day_2_06_2013_ifla_satellite_zarndt_et_al_crowdsourcing_the_worlds_cultural_heritage-_part_ii.pdf
4. List of crowdsourcing projects [Electronic resource]. — Mode of access: http://en.wikipedia.org/wiki/List_of_crowdsourcing_projects

Контактные данные:
e-mail: frederick@frederickzarndt.com, bgeiger@ucr.edu, apacy@cambridgema.gov, stefan@dlconsulting.com