Кирьянов И.К., Корниенко С.И. (Пермь)

УКАЗАТЕЛИ К СТЕНОГРАФИЧЕСКИМ ОТЧЕТАМ ГОСУДАРСТВЕННОЙ ДУМЫ КАК ОСНОВА ДЛЯ СОЗДАНИЯ ЭЛЕКТРОННЫХ ИСТОЧНИКОВ ПО ИСТОРИИ ПАРЛАМЕНТАРИЗМА В ДОРЕВОЛЮЦИОННОЙ РОССИИ [1]

Перевод исторических источников в машиночитаемый формат является ключевой проблемой компьютерного источниковедения [2].

Анализ традиционного источника с точки зрения создания его машиночитаемого аналога включает в себя оценку степени его структурированности, уровня формализации и унификации содержащихся в нем данных, характеристику наличествующих семантических связей и возможности их сохранения при создании логической и физической модели машиночитаемого источника.

При реализации проекта по созданию информационной системы "Российские парламентарии в начале XX века" в качестве основы для машиночитаемого источника были использованы указатели к стенографическим отчетам Государственной Думы, содержащие разнообразную информацию о составе и парламентской активности депутатского корпуса, структуре и деятельности Государственной Думы, ее взаимоотношениях с правительственными учреждениями и должностными лицами.

Указатели характеризуются достаточно высоким уровнем структурированности. При этом основные элементы структуры документа в своей основе едины и сохранены применительно к Думе всех четырех созывов. В структуре указателей выделяются два главных раздела: "предметный" и "личный". Предметный содержит перечень всех рассмотренных в Думе вопросов, сгруппированных по статьям. Объемные статьи подразделяются на рубрики. Указатели также включают в себя приложения, представляющие собой достаточно формализованные и унифицированные списки и таблицы: списки членов Государственной Думы по избирательным округам, должностных лиц Государственной Думы, членов Совета Министров; список заявлений о запросах (снабжен именным, предметным и алфавитным указателем); таблица законопроектов, внесенных в Государственную Думу; расписание заседаний Государственной Думы.

Применительно к задаче создания машиночитаемого источника для просопографического исследования особое значение имеет личный указатель. Структурирование, формализация и унификация личных данных депутатов в указателях, положенные в основу машиночитаемого источника, создают благоприятные возможности для решения таких исследовательских задач,

[1] Работа поддержана Российским гуманитарным научным фондом (проект № 03-01-12012в).

[2] См., например: Белова Е.Б., Бородкин Л.И., Гарскова И.М. и др. Историческая информатика. Учебное пособие. М., 1996; Гарскова И.М. Базы и банки данных в исторических исследованиях. М., 1994; Юмашева Ю.Ю. Источниковедческие проблемы создания просопографических баз данных // Информационный бюллетень Ассоциации "История и компьютер". 1992. №7.

как выявление связей между социокультурными и политическими характеристиками как отдельных депутатов, так и парламентских фракций и групп.

В то же время справочно-поисковый характер указателей порождает и определенные трудности при создании на их основе машиночитаемого источника. Они связаны с недостаточной в ряде случаев глубиной и полнотой информации. Путь преодоления этих трудностей заключается в использовании данных из других традиционных источников, прежде всего, стенографических отчетов заседаний Государственной Думы. Для этого указатели создают благоприятные возможности, располагая унифицированным аппаратом ссылок на соответствующую информацию в стенографических отчетах и других документах.

Собственно перевод указателей в машиночитаемый формат осуществлялся на основе сканирования источников на бумажных носителях с помощью планшетных сканеров и последующего распознавания с помощью стандартной программы сканирования и распознавания Fine Reader Professional 6.0.

При осуществлении этой задачи мы, как и другие исследователи, столкнулись с трудностями, связанными с тем, что в источниках используется старорусская орфография и грамматика, а также с особенностями шрифтов, которые использовались типографией Государственной канцелярии. При сканировании создавались достаточно четкие, хорошо читаемые графические образы документов, но при распознавании, даже при условии применения специально создаваемого алфавита и достаточно длительного и трудоемкого процесса распознавания с обучением, которое необходимо было проводить по отношению к каждому новому пакету-источнику, не удалось избежать значительного количества нераспознанных или ошибочно распознанных элементов. В результате пришлось проводить объемную работу по правке распознанного изображения, близкую по затратам к ручному набору имеющегося электронного текста. Исходя из этого, было решено отказаться от сплошного распознавания и работать с графическими образами, прибегая к выборочному распознаванию конкретных, необходимых элементов сканированного источника. Электронные версии документов-источников в качестве архива сохраняются в графическом формате, что при наличии соответствующего программного обеспечения создает возможность беспрепятственного обращения к информации электронной копии источника. Данные из графических файлов источников либо вводились с клавиатуры непосредственно в справочники и основные таблицы баз данных, либо, когда это было необходимо, подвергались фрагментарному распознаванию и после этого копировались в соответствующие справочники и таблицы. На наш взгляд, таким образом удалось оптимизировать затраты на перевод данных источника в машиночитаемый формат.