компьютерной графике и машинному зрению Графикон : сб. тр. (Н. Новгород, 10–15 сент. 2001 г.). — Н. Новгород : ННГАСУ, 2001. — С. 227–230.

Кучуганов 1985 — *Кучуганов, В. Н.* Автоматический анализ машиностроительных чертежей / В. Н. Кучуганов. — Иркутск, 1985.

Базы данных и корпусы текстов средневекового французского языка: подходы, проекты, технологии

А. М. Лаврентьев
Институт филологии Сибирского Отделения РАН, Новосибирск, Россия
Ecole Normale Supérieure Lettres et Sciences humaines, Лион, Франция

In this paper we will give a review of the current projects on the Medieval (9^{th} — 15^{th} centuries) French text databases, corpora and electronic editions. We will present: different choices of composition and sources for such databases (full text vs. sample corpora; using critical editions vs. original manuscripts); the principles of markup and the technologies used. Two projects will be considered in more detail: the Base de Français Médiéval (http://bfm.ens-lsh.fr) and the Charrette project (http://lancelot.baylor.edu)

История текстовых баз данных на французском языке восходит к концу 1950-х гг., когда была начата работа над подготовкой словаря «Тезаурус французского языка» (Trésor de la langue française), охватывающего период с XVI по XX век. Текстовая база, получившая название Frantext, активно развивалась в 60-е — 80-е гг., с начала 90-х гг. появилась возможность работы с ней с использованием CD-ROM, а затем и через Интернет. База Frantext приобрела широкую известность и стала активно использоваться лингвистами, литературоведами и историками. В то же время медиевисты оставались лишенными подобного исследовательского инструмента. Между тем тексты являются практически единственным источником данных о языке и литературе средневековья, и использование информационных технологий при их изучении позволило бы получить принципиально новые результаты.

В конце 70-х и 80-х годах возникают первые проекты компьютерного изучения средневековых французских текстов, однако полученные данные долгое время остаются доступными лишь их

создателям. Проекты доступных всему научному сообществу электронных изданий и текстовых баз данных начинают осуществляться в 90-е годы.

С самого начала исследователи столкнулись с двумя принципиальными вопросами, которые необходимо решить перед созданием электронного корпуса средневековых текстов:

1. На какой источник опираться: рукопись или издание?
2. Работать ли с целыми текстами или с фрагментами?

На их решение влияют не только методологические принципы исследователя, но и фактор времени и ресурсов, необходимых для создания корпуса. Следует также учитывать стремительный прогресс компьютерных технологий, позволивший снять многие проблемы, обусловившие в прошлом те или иные решения создателей корпусов.

Среди реализованных в прошлом и продолжающихся проектов текстовых баз на средневековом французском можно встретить все варианты решения данных вопросов.

Кратко охарактеризовав основные проекты корпусов средневековых текстов, разработчики которых объединились в 2004 г. в Консорциум корпусов средневекового французского (CCFM), мы остановимся подробнее на двух из них, представляющих внешне противоположные, но по сути взаимодополняющие подходы.

Первый проект — База средневекового французского (BFM) — был начат в 1989 г. в Высшей нормальной школе гуманитарных наук (ENS LSH) под руководством профессора К. Маркелло-Низья (C. Marchello-Nizia). BFM — полнотекстовая база данных, основанная на современных критических изданиях. Целью проекта было предоставить медиевистам подобный Frantext — мощный инструмент поиска и анализа текстовых данных.

Использование в базе полных текстов произведений объясняется тем, что она изначально была ориентирована на максимально широкий охват данных. Выбор в качестве источника критических изданий обусловлен стремлением в сравнительно короткие сроки сформировать достаточно крупный корпус. При этом для включения в базу отбирались прежде всего издания, подготовленные авторитетными медиевистами, которые строго следовали тексту «базовой» рукописи, исправляя лишь явные фактические ошибки и оговаривая все исправления в сносках.

Пополнение базы осуществлялось благодаря сканированию текстов изданий, а также обменам и безвозмездным вкладам специалистов, подготовивших в ходе своей личной исследовательской работы электронные версии отдельных текстов. Основным периодом, на который первоначально ориентировался проект BFM, был старофранцузский (IX–XIII вв.), однако в базу включались и полученные в дар или в обмен более поздние тексты. На ранних этапах проекта электронные тексты сохранялись в формате «простой текст», в кодировке DOS. С 2001 г. тексты базы постепенно переводятся в формат XML с разметкой, основанной на рекомендациях консорциума TEI.

В настоящее время основной корпус текстов BFM насчитывает 74 текста общим объемом около 3 000 000 словоупотреблений. Эксплуатация базы осуществляется через Интернет, с помощью системы Weblex, разработанной С. Эйденом (S. Heiden). Доступ к базе предоставляется бесплатно всем ученым, преподавателям и студентам, принимающим условия ее использования.

Второй проект — комплексное мультимедийное издание рукописной традиции романа Кретьена де Труа «Рыцарь телеги» (“Chevalier de la charrette”), получившее название «Проект Charrette», осуществленное в 1990–2003 гг. в Принстонском университете под руководством профессора К. Ютти (K. Uitti).

Его целью было создать принципиально новое издание средневекового текста, совмещающее в себе критически выверенный текст с фотографиями и детальными транскрипциями всех сохранившихся рукописей произведения, обогащенное морфологической и филологической разметкой и снабженное мощной поисковой машиной.

Первоначально транскрипции рукописей в проекте Charrette осуществлялись в формате SGML в соответствии с рекомендациями TEI. В 2001 г. было принято решение перевести их в формат XML с дополнительной сверкой с оригиналами и внесением дополнительной информации. Параллельно на основе критического издания была создана база данных поэтических фигур.

База данных по поэтическим фигурам, дипломатические транскрипции и фотографии рукописей доступны на сайте проекта.

В рамках Консорциума CCFM в настоящее время ведется выработка единых норм кодирования и описания средневековых

французских текстов. Одной из целей консорциума является создание единого портала, позволяющего осуществлять поиск во всех базах данных его участников. Совместными усилиями решаются проблемы морфологической разметки и лемматизации текстов, ведется исследование типологии текстов, призванное обеспечить более адекватную оценку репрезентативности существующих корпусов и наметить приоритетные направления их расширения.

Список литературы

Heiden S., Guillot C. Capitalisation des savoirs par le web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de Français Médiéval // Ancien et moyen français sur le Web, enjeux méthodologiques et analyse du discours / Kunstmann, P. (ред.). — Ottawa: Les éditions David, 2002. — С. 77–92.

Heiden, S., Lavrentiev, A. Ressources électroniques pour l'étude des textes médiévaux : approches et outils // Revue française de linguistique appliquée. — 2004. — IX(1). — С. 99–118.

Kunstmann, P. Ancien et moyen français sur le Web: textes et bases de données // Revue de linguistique romane. — 2000. — 253–254. — С. 17–42.

Проект INCA: цели, проблемы, тенденции и ИТ-решения по капитализации знаний

С. Г. Маслов
Ижевский государственный технический университет, Россия

The project on the knowledge capitalization system in the form of constructive electronic publication fund for technical as well as natural-scientific educational and research activities is considered. The project objectives and tasks are denoted. A realization approach based on the original component basis and synthesis of the system, cybernetic and synergetic approaches is suggested

Бурное развитие человечества происходило в периоды появления новых форм представления и методов переработки информации и знаний, а также синтеза существующих технологий. На современном этапе такой формой становятся электронные издания, накапливаемые в электронных библиотеках, на сайтах, на порталах и в электронных каталогах. Однако сложившегося разнообразия явно недостаточно, потому что оно концентрирует исследователя на уже известных ему формах, но существующих на других носителях. Это чаще всего приводит к избытку некачественной информации или дезинформации, к ситуации, когда излишек информации вреден точно так же, как и ее недостаток.

В работе рассматриваются современные проблемы, тенденции и ИТ-решения, которые инициировали *проект INCA* (фонд конструктивных электронных изданий в области инженерно-технического образования и научно-исследовательской деятельности) и составили основу его концептуальных архитектурных решений.

Фонд конструктивных электронных изданий — это стратегическое направление поддержки общества, основанного на знаниях, на базе непрерывности потока информации и знаний в цепи: *исследование и творчество* (анализ, генерация и синтез знаний); *образование* (распространение информации и знаний, формирование