

УДК 061.1:81(477)(082)

Дорожинська А.В.

Український мовно-інформаційний фонд НАНУ

ОЦИФРУВАННЯ ТЕРМІНОЛОГІЧНОГО СЛОВНИКА

Одним із завдань сучасної комп'ютерної лексикографії є створення цифрових словників, серед яких не останнє місце займають багатомовні термінологічні словники. Для більшості таких паперових праць в українській термінографії не існує цифрових аналогів, саме тому актуальне завдання їх оцифрування.

Сьогодні існує широкий набір інструментальних засобів для відображення у цифру словників різних типів, але не випрацювано технології для універсального вирішення цього завдання.

Серед усього словникового розмаїття був обраний для оцифрування «Словник української біологічної термінології» [1] (за словами авторів, цей словник є першою в україністиці лексикографічною працею нового покоління, що охоплює найуживанішу біологічну термінологію українською, російською та англійською мовами та містить тлумачення термінів).

Наш підхід полягає у поетапному перетворенні тексту словника у веб-сайт. Базові технологічні етапи: текст у PDF-форматі, HTML-файл, побудова концептуальної моделі лексикографічної системи Словника, побудова XML-файла, база даних, сайт.

Першим кроком було перетворення PDF-файлу у простий HTML-файл, що містить лише візуальну розмітку. Нам пощастило отримати PDF опублікованої книги (процес був би набагато складнішим, якби ми мали тільки друкований текст; у цьому випадку багато ресурсів треба було витратити на розпізнавання та коректуру сканованого тексту).

Наступним, але основним етапом, було побудова концептуальної моделі лексикографічної системи словника (за теорією лексикографічних систем В. Широкова [2]), яка була покладені в основу XML-структури словникової статті. Вся подальша робота будується на XML-файлі. Маркування тексту словника тегами XML було виконано програмними засобами. Наступний крок — формування бази даних та створення сайту.

Використані джерела

1. Словник української біологічної термінології. – К.: КММ, 2012. – 746 с.
2. *Широков В.А.* Комп'ютерна лексикографія : Монографія ; УМІФ – К. : Наук. думка, 2011. – 351 с.