У статті розглядається інформаційна технологія автоматизації створення бібліографічних описів

Ключові слова: бібліографічний опис, природномовні тексти, мова спеціалізованої розмітки XML, процесор обробки текстів

В статье рассматривается информационная технология автоматизации создания библиографических описаний

Ключевые слова: библиографическое описание, естественно-языковые тексты, язык специализированной разметки XML, процессор обработки текстов

The article describes the information technology automation creation of bibliographic descriptions

Keywords: bibliographic description of natural language texts, specialized markup language XML, text processor

УДК 681.3.01

РОЗРОБКА ЗАСОБІВ АВТОМАТИЗАЦІЇ СТВОРЕННЯ БІБЛІОГРАФІЧНИХ ОПИСІВ

О.В. Касілов
Кандидат технічних наук, доцент
Національний технічний університет «Харківський політехнічний інститут»
вул. Фрунзе, 21, м. Харків, Україна, 61002
Контактні тел.: (057) 707-66-84
E-mail: o.kasilov @ hotmail.com

Вступ

Згідно з Указом Держкомітету України з питань технічного регулювання і споживчої політики в Україні діє міждержавний стандарт ДСТУ 7.1:2006 «Система стандартів з інформації, бібліотечної та видавничої справи.

Бібліографічний запис, бібліографічний опис. Загальні вимоги та правила складання». Цей стандарт є базовим для складання бібліографічного опису всіх видів документів.

Постановка задачі

Огляд існуючих в бібліотеках України програмних засобів обліку бібліотечного фонду показав необхідність розробки засобів автоматизації створення бібліографічних записів та описів. Національна наукова бібліотека ім. В. І. Вернадського має каталоги поточних надходжень і каталоги спеціалізованих фондів в електронному вигляді, починаючи з 1994 р, сканований генеральний алфавітний каталог налічує 5 млн зображень карток.

З метою підвищення ефективності роботи автоматизованих бібліотечних систем необхідно створити інформаційну технологію спеціалізованої обробки текстів природної мови шляхом удосконалення процесу перетворення бібліографічних записів та описів (БЗО) згідно нових стандартів (ДСТУ 7.1:2006) [1] на базі модифікованої мови розмітки XML (eXtensible Markup Language) [2].

Основна частина

Бібліографічний опис містить бібліографічні відомості про документ, наведені за певними правилами, які встановлюють наповнення та порядок проходження областей та елементів, і призначені для ідентифікації і загальної характеристики документа. Бібліографічний опис є основною частиною бібліографічного запису.

Бібліографічний запис може включати також заголовок, терміни індексування (класифікаційні індекси і предметні рубрики), анотацію (реферат), шифри зберігання документа, додаткові бібліографічні довідки.

Розглянемо бібліографічний опис як елемент лексикографічних систем.

Застосуємо інформаційне моделювання лексикографічних ефектів. Під цим моделюванням розуміється наступне: інформаційна модель, що відповідає певній системі S довільної природи, має відображати лексикографічні аспекти системи S. В результаті такого моделювання формується лексикографічна система LS, що відображає елементи системи S, при цьому отримують структуровану лексикографічну систему, відповідну деякій початковій системі [3].

БЗО, як і словник, слід розглядати як інформаційну систему, в якій за допомогою поліграфічного виконання в паперовій копії або за допомогою елементів розмітки в електронному форматі позначені певні лінгвістичні ефекти. Ними можуть бути: шрифтові виділення, позиційні розміщення, спеціальні позначення та ін., які відіграють роль ідентифікаторів відповідних інформаційних змін.

Розробка програмних комплексів, орієнтованих на вирішення завдань обробки специфічних текстів, вимагає набору адекватних формальних інформаційних моделей систем, які могли б відіграти роль початкової концептуальної бази для програмування лексикографічного процесора або його складових частин. Інформаційні системи спеціального типу вимагають дотримання єдиної методології проектування. В першу чергу визначається загальна архітектура системи, розробляються складові частини цієї архітектури, зв'язки і відображення між ними. Відповідно до ANSI/X3/SPARK, в архітектурі інформаційної системи виділяються три рівні опису даних: концептуальний, внутрішній і зовнішній [3].

Бібліографічний опис згідно стандарту [1] має наступні області: 1. область заголовка і відомостей про відповідальність; 2. область видання; 3. область специфічних відомостей; 4. область вихідних даних; 5. область фізичної характеристики; 6. область серії; 7. область примітки; 8. область стандартного номера (чи його альтернативи) та умов доступності.

Області опису складаються з елементів, які поділяються на обов'язкові та факультативні. В описі можуть бути або тільки обов'язкові елементи, або обов'язкові і факультативні.

Для формування бібліографічного опису використовується запропонована пунктуація, що виділяє складові елементи і області або укладає їх. Її вживання не пов'язане з нормами мови.

В якості запропонованої пунктуації виступають знаки пунктуації та математичні знаки: крапка і тире, крапка, кома, двокрапка, крапка з комою, три крапки, коса риса, дві косі риси, круглі дужки, квадратні дужки, знак плюс, знак рівності. В кінці бібліографічного опису ставиться крапка.

Позначимо структурні елементи бібліографічного опису наступним чином:

$$d_i < M_{i1}, M_{i2}, M_{i3}, M_{i4}, M_{i5}, M_{i6}, M_{i7}, M_{i8} > ,$$

де d_i – біографічний опис; M_{i1} – область заголовка і відомостей про відповідальність; M_{i2} – область видання; M_{i3} – область специфічних відомостей; M_{i4} – область вихідних даних; M_{i5} – область фізичної характеристики; M_{i6} – область серії; M_{i7} – область примітки; M_{i8} – область стандартного номера (чи його альтернативи) та умов доступності.

Деякі з перерахованих елементів можуть бути одноелементними і навіть порожніми, тобто можуть бути відсутніми в біографічному опису.

Розглянувши структуру бібліографічного опису, сформулюємо набір правил для перетворення вхідних даних (бібліографічний опис) у вихідні дані (бібліографічний опис ДСТУ 7.1:2006 в XML запису).

$$ПР_n\left(T_n^j\right) = R_n,\ \ j = 1, M\ ,$$

де $ПР_n$ – програма, що здійснює перетворення; T_n^j – складова частина бібліографічного опису; R_n – результат перетворення; j – номер складової частина бібліографічного опису.

Досліджено існуючі варіанти розмітки документів, що відповідають попереднім редакціям ГОСТу. Розмітка документа (бібліографічного опису) має на меті: виділення смислових частин (логічних елементів) документа і зв'язків між ними (структурна розмітка); визначення дій, які мають бути здійснені з цими елементами.

Мова розмітки повинна визначати ряд спеціальних інструкцій, правил і угод для опису структури елементів документа і відносин між елементами цієї структури. Спеціальні інструкції, їх ще називають маркерами або тегами, в структурованих документах повинні певним чином кодуватися, тобто виділятися серед основного тексту. Їх головне призначення – служити інструкціями управління для програмних засобів обробки структурованих текстів.

Враховуючи функціональні можливості сучасних бібліотечних систем, такі, як експорт даних з різноманітних форматів, зокрема XML, вирішено формувати вихідні дані в модифікованому вигляді в XML форматі. При цьому XML накладає деякі обов'язкові вимоги до розмітки даних елементів бібліографічного опису.

На рівні метамови визначаються внутрішні атрибути розмітки, базові набори символів, правила при-

власнення імен, зарезервовані слова, допустимі відхилення від стандарту (відсутність кінцевого тегу) відповідно до XML-стандарту [2]. Оптимальним є дотримання вимог стандарту XML. Кодування даних задається за бажанням користувача.

Доцільно використовувати Unicode/ISO 10646 для запису даних.

На синтаксичному рівні визначаємо точні назви тегів і синтаксичні правила їх подання. На цьому рівні відповідність синтаксичному стандарту може бути перевірено за допомогою аналізу тексту, тобто формальним способом. На семантичному рівні необхідно гарантувати однозначність розмітки та її інтерпретаційної частини.

Для зняття неоднозначностей при перенесенні текстової інформації між різними системами необхідно вказати для користувача належність використовуваних тегів частинам документа, в нашому випадку – структурним елементам бібліографічного опису.

Такі правила в основному визначаються в супроводжуючих довідниках користувача.

Висновки

Вирішення задач розробки засобів автоматизації створення бібліографічних описів дозволяє максимально ефективно інтегрувати систему в існуючі бібліотечні системи та спеціалізовані Web ресурси.

Досліджено структуру даних бібліографічних описів в форматах попередніх ГОСТів і розроблено засоби перетворення бібліографічних описів згідно ГОСТу ДСТУ 7.1:2006 у спеціальну електронну форму. Побудовано концептуальну математичну модель бібліографічного опису.

Розроблено вимоги до мови розмітки бібліографічного опису, розвинуто можливості мови XML з метою застосування її для розмітки структурованих текстів у вигляді бібліографічного опису.

Література

1. ДСТУ ГОСТ 7.1:2006. Бібліографічний запис, бібліографічний опис. Загальні вимоги та правила складання : метод. рекомендації з впровадження [Текст] / уклали: Галевич О. К., Штогрин І. М. – Львів, 2008. – 20 с.
2. Extensible Markup Language (XML) 1.1 (Second Edition) [Електронний ресурс] / Режим доступу: http://www.w3.org/TR/xml11.
3. Касилов О. В. Моделирование лексикографических систем [Текст] / О. В. Касилов // Вісник Міжнародного Слов'янського університету. Сер. «Технічні науки». – Харків.:– 2004. – Т. 7. – № 1. – С. 13–15.