대용량 비정형 데이터 자료 입력 및 출력

심규철* · 강병준* · 김경환* · 정회경***

*커뮤 S/W 기술 연구소 · ***배재대학교

Data Input and Output of Unstructured Data of Large Capacity

Kyu-Cheol Sim* · Byung-Jun Kang* · Kyung-Hwan Kim* · Hoe-Kyung Jung***

*Commu S/W Technology Institute · ***Paichai University

E-mail : {kcshim, junycom, KHKIM}@commu.co.kr, hkjung@pcu.ac.kr

요 약

최근 들어 워드 파일을 XML로 변환하여 서비스하기 위한 요구가 많아지고 있다. 이에 본 논문에서는 워드 파일(아래한글, MS-Office)로 입력된 데이터를 XML 파일로 변환하여 사용자가 XML 매핑 파일을 만들어 워드 프로세서에 입력된 데이터를 바로 추출하여 데이터베이스에 저장하는 시스템을 제안한다. 이는, 워드프로세스에 양식을 미리 작성하여 필요한 데이터를 데이터베이스에서 조회하여 워드프로세서 문서를 어플리케이션 프로그램에서 워드 파일을 생성 할 수 있다.

ABSTRACT

Request to provide a service to XML word file recently has been increasing. In this paper, it is converted to an XML file data input (HWP, MS-Office) a Word file, stored in a database by extracting data directly input to the word processor user creates an XML mapping file I to provide a system that. This can be retrieved from the database the required data to previously created forms word processor, to generate a Word file from the application program a word processing document.

키워드

대용량 비정형 데이터, XML, 매핑파일, 자료입력

Ⅰ. 서 론

일반적으로 웹상의 데이터 입력은 정형화된 입력 화면을 제공하고 사용자가 화면에 데이터를 입력하면 적합성 등을 검사하여 이를 데이터베이스에 저장한다. 데이터베이스에 저장된 데이터를 웹 화면에 출력하여 사용자에게 데이터를 제공하고 있다. 그러나 비정형화된 데이터를 웹을 이용하여 데이터를 입력할 수 있는 시스템의 구축이 매우 어렵고 많은 시간이 소요되어 일반적으로 사용자가 사용하는 아래한글이나 MS-Office(워드, 파워포인트, 엑셀 등)를 이용하여 다양한 양식에 데이터를 입력하고 입력된 파일은 웹을 통하여 업로드 한다. 비정형화된 데이터를 추출하여 데이터베이스에 데이터를 입력할 수 있는 시스템 개발이 필요하여 개발을 진행하고 있다. 본 논문에서는, 업로드 된 binary 파일을 바로 XML 파일로 변환할 수 있는 엔진을 개발하고 변환된 XML 파일과 Mapping XML 파일을 파싱하여 데이터베이스에 데이터를 입력하거나 또는 데이터베이스에 입력된 데이터를 양식(HWP, MS-Office) 파일로 출력할 수 있는 시스템을 설계 및 구현을 진행하고 있다. 그림 1은 전체 시스템의 흐름도이다.

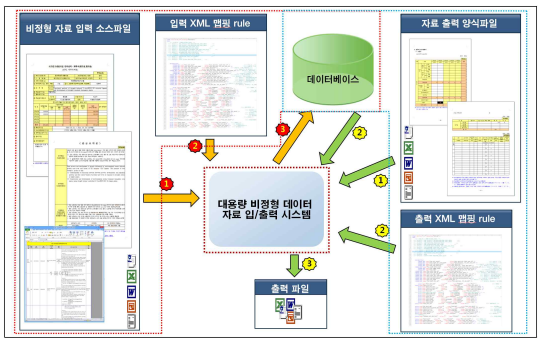

그림 1. 전체 시스템의 흐름도

Ⅱ. 관련연구

2.1 원본파일 변환 엔진

본 논문에서는 원본파일(HWP, MS-Office)를 binary 상태에서 바로 XML로 변환 할 수 있는 엔진을 개발하는데 HWP파일의 경우 한글과컴퓨터에서 제공하는 HWP Document File Formats[1]의 형식으로 원본을 읽어서 바로 XML로 변활 할 수 있는 엔진을 개발하고 MS-Office파일의 경우 다양한 Open Source[3]를 이용하여 원본 파일을 바로 XML파일로 변환하는 엔진을 개발한다.

HWP, MS-Office 원본파일은 MICROSOFT사의 Compound Document File Format[2]를 사용하여 4기가바이트 이상의 대용량 파일을 저장할 수 있는 파일형식으로, 원본파일을 XML로 변환하기 위하여 Compound Document File 형태로 파일을 읽어서 다시 각각의 원본파일 Format으로 읽어 야 한다. 본 논문에서는 Compound Document File를 읽을 수 있도록 엔진에 포함하여 개발한다.

2.2 Mapping XML 파일

본 논문에서는 원본을 XML로 변환하여 이 변환된 XML파일에서 데이터 부분을 추출하기 위하여 Mapping XML 파일을 작성 한다. 원본 XML 파일과 Mapping XML 파일을 파싱하여 원본의 XML 파일에서 데이터를 추출한다. 추출하기 전에 먼전 원본 XML파일과 Mapping XML 파일의 Validation check하여 XML 파일의 정확성을 확인하여 다음 작업을 진행 하도록 한다. 설명은 아래 그림 2와 같다.

```
<?xml version="1.0" encoding="EUC-KR"?>
<문서>
    <!-- 양식A101 : 과제마스터(TNR_SUBJ) -->
        <!-- 양식A201 : 과제_[양식A201] 한글(영문)요약문(TNR_SUBJ_SMRY)       -->
        <!-- 양식A202 : 과제_[양식A201] 한글(영문)요약문(TNR_SUBJ_SMRY)       -->
        <!-- 양식A103 : 과제_연구분야(TNR_SUBJ_RSRC_FLD)                  -->
        <!-- 양식A301 : 과제_연구책임자정보(TNR_SUBJ_RPMN)                 -->
        <!-- 양식A501 : 최근5년이내에 종료된 사업수행현황 (TNR_SUBJ_OTH_PTCP)
        -->
        <!-- 양식A502 : 타연구사업수행현황 (TNR_SUBJ_OTH_PTCP)        -->
        <!-- 양식A611 : 연구비총괄표(TNR_SUBJ_BITM_RCST)   -->
    <!-- 양식A401 : 내부인건비(TNR_SUBJ_PTCP_MNPW)                 -->
    <!-- 양식A612 : 외부인건비(TNR_SUBJ_PTCP_MNPW")                -->
        <!-- 양식A613 : 학생인건비(TNR_SUBJ_PTCP_MNPW")                -->
        <!-- 양식A606 : 재료비                  -->
        <!-- 양식A604 : 신진연구 연구장비 지원금(TNR_RSRC_EQIP)                -->


    <properties>
        <TABLE_KEY name="TAC_ACP_SBJT"              value="ACP_SBJT_NO,
REPT_DVS_CD" />              <!-- 양식A101 : AC_보안등급 -->
        <TABLE_KEY name="TAC_SBJT_TPI_CORP_NUM"     value="ACP_SBJT_NO,
TPI_AGC_DVS_CD" />              <!-- 양식A101 : AC_과제참여기업수 -->
        <TABLE_KEY name="TAC_ITE_ASSO_RSCL"
value="ACP_SBJT_NO" />
        <!-- 양식A101 : 과제_국제공동연구 -->
        <TABLE_KEY name="TAC_ANU_RSRCCT"
value="ACP_SBJT_NO,  SBJT_ANU_NGR" />                          <!-- 양식
A101 : AC_연차연구비 -->
                <TABLE_KEY name="TAC_SBJT_RSCH_SPHE"
value="ACP_SBJT_NO, RSCH_SPHE_CLS_CD, RSCH_SPHE_CD" /><!-- 양식A103 : 과제_연
구분야 -->
        <TABLE_KEY name="TAC_SMMR_CNTN"             value="ACP_SBJT_NO,
KR_ENG_DVS_CD" />              <!-- 양식A201, 양식A202 :AC_요약내용 -->
        <TABLE_KEY name="TAC_KWD"
value="ACP_SBJT_NO, KR_ENG_DVS_CD, KWD_SNO" />        <!-- 양식A201, 양식
A202 :AC_키워드 -->
                <TABLE_KEY name="TAC_RSCH_RPB_PE"
value="ACP_SBJT_NO" />
            <!-- 양식A301 : 과제_연구책임자정보 -->
                <TABLE_KEY name="TAC_RSCH_RPB_PE_EDBC"
value="ACP_SBJT_NO,EDBC_SNO" />
<!-- 양식A302 : 과제_연구자학력-->
                <TABLE_KEY name="TAC_IOE_ANU_CTAM"
value="ACP_SBJT_NO, SBJT_ANU_NGR, IOE_CD" />              <!-- 양식A611 : 과
제_비목별연구비소요내역 -->
                <TABLE_KEY name="TAC_SBJT_TPI_CORP_SPPT_AMT"
value="ACP_SBJT_NO, TPI_CORP_SNO" />                          <!-- 양식
A105 : 참여기업별 부담금액 -->
                <TABLE_KEY name="TAC_OTH_RSCH_BIZ_RUN"
```

그림 2. Mapping XML 파일의 예

Ⅲ. 대용량 비정형 데이터 자료입력 및 출력 시스템 설계

3.1 대용량 비정형 데이터 자료입력 및 출력 소프트웨어 구성도

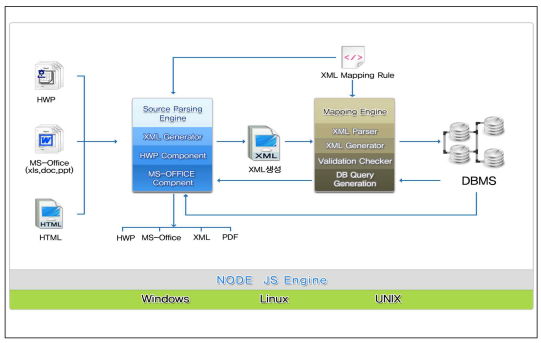

그림 3. 대용량 비정형 데이터 자료입력 및 출력 소프트웨어 구성도

그림 3의 대용량 비정형 데이터 자료입력 및 출력 소프트웨어 구성은 다양한 O/S 환경에서 설치가 가능하도록 개발하고 또한 Active-X 방식을 사용하지 않도록 설계하여 다양한 브라우저를 지원 할 수 있도록 설계하였다. 본 시스템의 엔진은 크게 2개로 분리 설계하여 다른 시스템 구축 시에도 엔진을 각각 활용 할 수 있도록 하였으며, 두 엔진간의 통신은 IPC를 사용하여 서로 데이터를 송수신 할 수 있도록 하였다. 웹서버를 지원하기 위하여 Open Source 엔진인 NODE JS 엔진을 활용한다. NODE JS 엔진은 기존 웹서버나 웹 어플리케이션 서버보다 가볍고 JavaScript를 사용하여 개발이 가능하며 다양한 어플리케이션에 활용이 가능하다 최근에는 계속 버전을 업그레이드하여 안전성 및 성능이 많이 향상되었다.

3.2 원본파일 파싱엔진 설계

원본파일은 NODE JS 엔진을 통해 HTTP 프로토콜을 사용하여 서버에 업로드 할 수 있는 기능을 제공하고 HWP Component와 MS-Office Component를 사용하여 원본파일을 XML 파일로 변환한다. 또한 원본파일을 binary파일로 출력해 사용하기 위하여 원본파일을 파싱하여 파일의 변수부분에 데이터를 치환하여 원본파일의 복사본을 binary 파일로 변환하는 기능을 함께 제공할 수 있도록 설계하였다. 변환된 XML파일을 서버의 임시영역에 저장하여 Mapping 엔진에 Mapping 작업을 의뢰하기 위하여 각종 정보를 통신 모듈을 통하여 전달한다. 그림 4는 원본파일 파싱 엔진의 구성도이다.

그림 4. Source Parsing Engine

3.3 Mapping 엔진 설계

Mapping 엔진은 XML 파서를 이용하여 원본 XML 파일과 Mapping XML 파일을 Validation check를 하여 이상이 없을 때 XML Generator를 이용하여 원본파일과 Mapping XML 파일을 분석하여 데이터베이스에 입력할 수 있는 SQL 문장을 생성한다. 만약 Validation문제나 두 XML 파일에 에러가 발생하면 사용자에게 그 에러코드를 전송한다. Mapping 엔진은 XML 파싱 및 데이터베이스에 데이터를 입출력할 수 있는 모듈을 모두 포함하여 입력 및 출력에 사용할 수 있도록 독립적인 형태로 설계하였다.

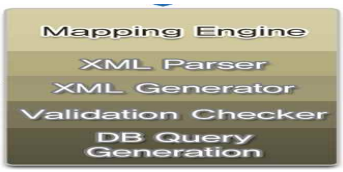

그림 5. Mapping Engine

Ⅳ. 대용량 비정형 데이터 자료입력 및 출력 시스템 구현

4.1 시스템 모듈화

본 논문의 '대용량 비정형 데이터 자료입력 및 출력 시스템'은 그림 3과 같이 엔진을 분리하여 개발하였다. 분리된 엔진들은 독립적으로 실행이 가능하도록 하여 향후 다른 시스템에서 각각의 엔진을 사용할 수 있도록 구현하였다. 또한 엔진에 구현된 Component도 독립적인 모듈로 구현하여 다른 어플리케이션에서도 라이브러리 또는 단위 어플리케이션으로 활용이 가능하도록 구현하였다.

4.2 대용량 비정형 데이터 입력 시스템

대용량 비정형 데이터 입력시스템은 관리자가 원본파일을 배포하여 사용자가 원본파일에 데이터를 작성하고 시스템에 업로드하면 원본파일과 Mapping XML 파일을 파싱하여 데이터베이스에 데이터를 입력할 수 있도록 구현하였다. 입력된 데이터는 바로 화면에 HTML 형태로 출력하여 필요시 웹 화면에서도 수정이 가능하도록 하였고, 원본파일을 수정하고 다시 업로드하면 데이터의 내용을 수정 할 수 있도록 구현하였다.

4.3 대용량 비정형 데이터 출력 시스템

대용량 비정형 데이터 출력시스템은 기존 Active-X 방식의 WEB Reporting Tool의 처리방식과는 다르게 HWP, MS-Office의 양식파일을 작성하여 데이터출력 부분에 변수명을 기입하고 데이터베이스의 데이터를 조회하여 원본파일의 변수명과 치환하여 다시 원본파일을 작성한다. 작성된 원본파일은 HTTP 프로토콜을 사용하여 사용자에게 다운로드할 수 있는 기능을 구현하여 제공한다. 4.2대용량 비정형 데이터 입력 시스템과 4.3 대용량 비정형 데이터 출력 시스템의 설명은 아래 그림 6과 같다.

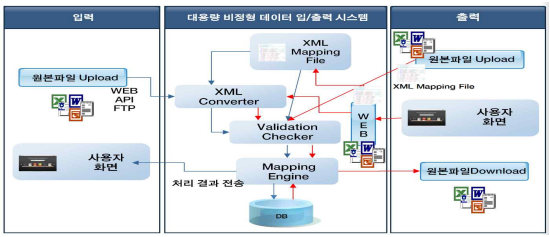

그림 6. 구현 시스템 상세 구성도

Ⅴ. 결 론

웹상에서 비정형 데이터를 다양한 양식을 통하여 시스템에 빠른 입출력을 제공할 수 있는 프로그램으로 사용자나 관리자 모두에게 편리한 방식의 데이터 입력 시스템으로 사용할 수 있어서 매우 편리하게 사용이 가능 할 것으로 예상된다. 개발자는 다양한 양식을 입력 받기 위하여 프로그램의 개발이나 변경이 거의 없으므로 시스템 구축시 많은 시간을 단축시킬 수 있다. 본 시스템은 하나의 독립된 솔루션으로 구축이 가능하여 향후 어플라이언스 서버에 소프트웨어를 탑재하여 단독으로 다양한 포맷의 데이터를 고속으로 처리할 수 있도록 개발이 가능할 것이다. 하지만 이 시스템은 사용자가 Mapping XML File을 규칙에 맞도록 작성하여야 하므로 Mapping XML File 작성 방법을 습득하여야하는 단점이 있다. 이를 해결하기 위하여 사용자가 Mapping XML File를 쉽게 작성할 수 있도록 GUI 어플리케이션을 작성하여 제공한다면 이 단점도 해결이 가능하다.

참고문헌

[1] (주) 한글과컴퓨터, 한글 문서 파일 구조 (Hwp Document File Formats)
[2] Daniel Rentz, Microsoft Compound Document File Format
[3] http://www.apache.org