맵핑파일과 HWPML을 이용한 아래한글(HWP) 데이터 추출 방법

심규철* · 강병준* · 하의륜** · 저순** · 박성현** · 김아용** · 정회경**

*㈜커뮤 · **배재대학교 컴퓨터공학과

Unstructured Big Data Input and Output

Kyou-Chul Shim* · Byeong-Jun Kang* · YiLun-He** · Xun-Chu** · Sung-Hyun Park**

· A-Yong Kim** · Hoe-Kyung Jeong**

*COMMU CO.,LTD · **Department of Computer Engineering, PaiChai University

E-mail : {kcshim, junycom}@commu.co.kr, mrheyilun@gmail.com, 846045855@qq.com,

enoid00@gmail.com, janlssary@naver.com, hkjung@pcu.ac.kr

요 약

사용자가 작성한 아래한글(HWP)의 양식에서 사용자가 입력한 데이터를 추출하기 위하여 데이터 추출용 맵핑파일을 XML로 작성하여 아래한글의 HWPML형태의 XML 파일을 맵핑파일과 데이터를 맵핑하여 사용자가 입력한 데이터를 추출하여 대량의 양식 데이터를 입력 및 출력에 활용가능하게 할 것이다.

ABSTRACT

User-written in a form of Hangul (HWP), under the data entered by the user in order to extract an XML mapping files for extraction of data by filling in the below map file and an XML file in the form of a HWPML Korean data, mapping them to extract data entered by the user by taking advantage of the large amounts of form data input and output, will make it possible.

키워드

아래한글, 맵핑파일, XML, HWPML, 비정형데이터

Ⅰ. 서 론

사용자에게 일정한 형식의 데이터를 수집하기 위하여 웹을 이용하여 사용자가 데이터를 입력하도록 입력화면을 작성하여 제공하고 사용자는 웹을 사용하여 데이터를 입력하여 저장한다. 이 방법은 다양한 양식의 입력을 받기 위해서는 서로 다른 화면을 제공하기 위하여 입력 화면을 프로그래밍하여 제공하고 관리하여야 한다. 이런 방법은 계속 변경 되어지는 양식을 통하여 데이터를 수집하기 위해서는 매우 많은 비용이 소요되고 있다. 또한 사용자 입력 데이터 용량이 많은 경우는 웹에서 데이터 입력을 받을 경우 모든 데이터 입력이 완료 될 때까지 입력 작업을 멈출 수 없다. 이런 방법을 해결하기 위해서 아래한글 양식을 제공하고 제공된 양식의 서식에 맞춰서 사용자가 데이터를 입력하고 수정해서 최종 완료된 양식 문서를 웹을 통하여 업로드하고 이 양식에서 사용자가 입력한 데이터를 추출하여 데이터베이스에 저장하면 사용자는 익숙한 아래한글을 통하여 데이터를 입력하고 또한 대량의 데이터를 쉽게 입력 받을 수 있어서 빠른 데이터 입력이 가능 할 것이다. 또한 사용자에게 제공하는 양식이 변경되어도 프로그램은 변경이 되지 않고 맵핑파일을 작성해서 서버에 저장하면 그 맵핑파일을 통해서 데이터를 추출 할 수 있으며 빠른 시

간에 다양한 양식을 제공하여 데이터를 수집할 수 있다.

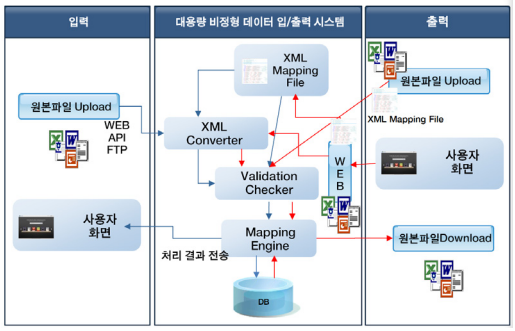


그림 1. XML 맵핑 다이어그램

Ⅱ. 관련연구

2.1 HWPML

한글 워드 프로세서 문서를 기술하기 위한 W3C XML 기반의 개방형 마크업 언어이다. HWPML 엘리먼트에 대한 설명은 표 1과 같다.[1]

한글 워드프로세서의 원본 문서를 HWPML로 변경 저장하여 Mapping XML 파일과 매핑하여 필요한 부분의 데이터를 추출한다.

표 1. HWPML 엘리먼트

엘리먼트 명				
설명		엘리먼트에 대한 설명		
부모 엘리먼트				
자식 엘리먼/엘리먼트 값				
속성	속성명1	속성1에 대한 설명	값의 범위	기본값
	속성명2	속성2에 대한 설명	값의 범위	기본값

2.2 Mapping XML 파일

Mapping XML 파일은 그림 2와 같이 HWPML XML파일을 근간으로 데이터를 추출할 수 있는 엘리먼트를 정의하고 데이터의 속성을 정의할 수 있도록 Mapping 파일을 정의하였다.

```
<?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 예시

Ⅲ. Mapping 파일과 HWPML 맵핑 방법 설계

3.1 XML 데이터 매핑

매핑은 Mapping XML 파일과 HWPML 두가지의 다른 구조를 지닌 명세 형식에서, 동일하거나 비슷한 의미를 지닌 데이터 요소를 연결하고, 필요에 따라 연결정의에 추가적인 의미를 부여하여, 두 명세 형시간의 관계를 정의하는 것이다.

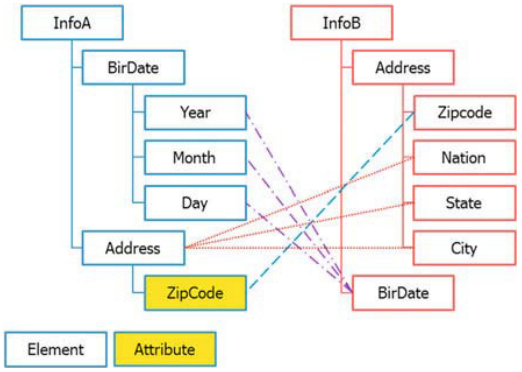


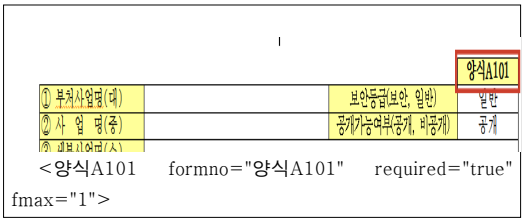
그림 3. XML 데이터 매핑

그림 3은 두 문서 구조 정보에 대해 데이터 매핑을 통해 정의한 예로 써, 보는 바와 같이 두 구조 정보는 같은 내용을 담고 있으면서도 다른 구조를 지니고 있다.

3.2 매핑 XML의 정의

Mapping XML의 내용은 여러 종류의 양식을 구분하기 위하여 formno와 테이블의 데이터를 추출하기 위하여 datapos와 같은 테그를 정의하여 Mapping XML를 정위하였다. 그림 4와 같은 양식을 구분하기 위하여 워드프로세서 파일로 최상의에 1칸의 테이블을 만들고 '양식'이라 표기하여

양식을 구분 할 수 있도록 하였다.


양식A101

① 부처사업명(대)		보안등급(보안, 일반)	일반
② 사 업 명(중)		공개가능여부(공개, 비공개)	공개

<양식A101 formno="양식A101" required="true" fmax="1">

그림 4. 예제 양식 및 Mapping XML 정의

원본 파일에서 데이터를 추출하기 위하여 테이블의 위치를 지정할 수 있도록 datapos를 정의하였다. datapos의 위치는 X(수평방향), Y(수직방향)로 지정할 수 있으며 X, Y와 같이 +(간째), *(건너띄기)의 기호를 함께 사용하여 작성 할 수 있다. datapos 태그작성 유형은 다음 표 2와 같다.

표 2. datapos테그의 유형

[유형1] - X+1 : 수평방향 우측으로 1칸 옆의 Cell에서 데이터를 추출함 [유형2] - Y+1 : 수직방향 아래로 1칸째 Cell에서 데이터를 추출함 [유형3] - Y*2 : 데이터가 반복되는 행 같은 경우 해당 헤더의 위치에서 2칸씩 건너뛰어 서 데이터를 추출 함 (헤더영역의 Cell이 2줄로 나뉘어 있음) [유형4] - Y+1 : 데이터가 반복되는 행 같은 경우 해당 헤더의 위치에서 1칸씩 건너뛰어 서 데이터를 추출 함 (헤더영역의 Cell이 1줄로 되어 있음) (칸수가 1인 경우는 +,* 둘다 같은 결과를 가져옴)

아래 그림 5의 예제는 datapos=Y+1의 Mapping XML정의한 예제 부분이다.

연구참여인력	연구책임자		연구원		연구조원			기타	합계
	총괄(주관)	제부	교수급	Post-doc.	박사과정	석사과정	학부생		
	1명	0명	0명	0명	2명	4명	0명	0명	1명

<연구참여인력>
<연구책임자>
<총괄주관 datapos="Y+1"/>
</연구책임자>
</연구참여인력>

그림 5. Mapping XML datapos=Y+1 예제

Ⅳ. 결 론

본 논문에서는 사용자 다양한 양식의 사용자 입력을 하나의 시스템에서 양식과 Mapping 파일을 변경하여 웹상에서 대량의 입출력이 가능하도록 구성할 수 있는 XML Mapping 기술을 정의하여 사용자의 입력 시간을 단축하고 시스템을 제공하는 제공자는 사용자 입력 시스템 개발에 시간을 단축할 수 있다. 또한 사용자는 웹 화면이 아닌 익숙한 아래한글을 통하여 다양한 정보를 입력이 가능하기 때문에 자료 입력을 오류를 줄이고 빠른 시간에 데이터 작성이 가능하다는 장점이 있다. 개발자는 Mapping 파일을 작성하기 위하여 Editor를 이용할 수 있으나 Mapping 파일을 작성할 수 있는 GUI 어플리케이션을 개발하여 제공하면 쉽게 Mapping 파일을 작성할 수 있는 기능을 제공 할 수 있다.

참고문헌

[1] (주) 한글과컴퓨터, “한글 문서 파일 구조 (Hwp Document File Formats)”, 2014.04
[2] Daniel Rentz, “Microsoft Compound Document File Format”, 2014.04
[3] http://www.apache.org, 2014.04