Dublin Core 메타데이타 형식에 관한 연구

A Study on Dublin Core Metadata Format

김선미, 최석두
이화여자대학교 문헌정보학과

Sun-Mi Kim, Suk-Doo Choi
Dept. of Lib. & Inf. Sci., Ewha Womans Univ.
sdchoi@mm.ewha.ac.kr

메타데이터 요소집합이란 다양한 형식의 전자문헌 중에서 필요한 정보원을 찾아내는데 도움이 되는 전자문헌의 가장 기본적인 특징을 기술하기 위한 것이며, 색인보다는 정보량이 많고 공식적인 목록레코드보다는 덜 완전한 레코드를 만들어 이 양극을 조정하려는 것이다. 이의 이점은 다음과 같다. 자동정보원탐색기가 데이터를 수집할 수 있는 형식으로 저자나 출판자가 메타데이터를 만들도록 장려할 수 있으며, 메타데이터레코드의 작성업무를 단순화함으로써 메타데이터요소의 템플릿을 갖는 네트워크출판도구의 개발을 장려할 것이다. 또한 만들어진 메타레코드는 필요하다면 보다 상세한 목록레코드의 바탕이 될 수 있을 것이며, 표준이 된다면 메타데이터레코드는 모든 이용자 집단에게 이해될 수 있을 것이다. 현재 이와 같은 관점에서 활발히 논의되고 있는 Dublin Core 메타데이터 형식을 중심으로 그 현황에 대하여 논한다.

1 서 론

메타데이터란 문자 그대로 '데이터의 데이터', 즉 어떤 오브젝트의 특성을 대체하는 레코드의 내용을 말한다. 문제는 단순한 메타데이터레코드가 어떻게 넓은 범위의 전자적 오브젝트를 충분히 기술할 수 있도록 정의하는가에 있다.

Dublin Core(*Dublin Metadata Core Element Set*: 이하 DC라 한다)의 기본적인 사상은 더 복잡한 레코드, 보다 상세한 레코드, USMARC과 같은 고도로 통제된 시스템으로 링크시킬 수 있는 단순구조레코드를 만들어, 색인보다는 정보량이 많고 공식적인 목록레코드보다는 덜 완전한 레코드를 만들어 이 양극을 조정하려는 생각이다. 이를 위한 메타데이터 요소집합이란 레코드를 단순하게 하여 저자나 정보제공자도 작성할 수 있도록 하고, 탐색시스템에 독립적이고 누구나 이해할 수 있는 요소집합을 만들어, 다양한 형식
의 전자문헌 중에서 필요한 정보원을 찾아내는 데 도움이 되는 전자문헌의 가장 기본적인 특징을 기술하기 위한 요소이다.

이를 위하여 1995년 3월 1-3일 Dublin, Ohio에서 사서, 보존문서가, 인문학자, 지리학자, Z39.50 표준학자, SGML위원회 등 52명이 모여 1차 워크샵을 열고 이를 논의하였다. 이어서 제2차회의는 영국의 Warwick(April, 1996), 제3차회의는 다시 Dublin (September 24-25, 1996), 제4차회의는 호주의 Canberra(March 3-5, 1997)에서 개최되었다. 특히 제3차회의의 내용은 이미지기술에 관한 것이었다.

현재 이와 같은 관점에서 활발히 논의되고 있는 Dublin Core 메타데이터 형식을 중심으로 그 현황에 대하여 논한다.

2 필요성

현재 네트워크 전자문헌에 대한 기술방법은 크게 두 가지 형식이 있다. 하나는 Lycos나 WebCrawler와 같이 locator를 이용하여 자동생성된 색인이며,

또 하나는 전문가에 의해서 만들어진 MARC과 같은 목록레코드이다. 자동생성된 레코드는 사용하기에 너무 적은 정보를 갖고 있으며, 수동으로 생성된 레코드는 인터넷상에서 이 방대한 전자문헌을 만들고 유지하기에는 돈이 너무 많이 든다. 또한 현재의 전자정보원을 기술하는 데 사용하고 있는 공식적인 표준(예를 들면, TEI 헤더 혹은 MARC)은 일반 정보원의 극히 일부에만 적용할 수 있다.

DC는 이를 조정하기 위한 것이며, 기본적으로 다음과 같은 장점을 가지고 있다.

1) 자동정보원탐색기가 데이터를 수집할 수 있는 형식으로 저자나 출판자가 메타데이터를 만들도록 장려할 수 있다.
2) 메타데이터레코드의 작성업무를 단순화함으로써 메타데이터요소의 템플릿을 갖는 네트워크 출판도구의 개발을 장려할 것이다.
3) DC로 만들어진 레코드는 필요하다면 보다 상세한 목록레코드의 바탕이 될 수 있을 것이다.
4) DC가 표준이 된다면 메타데이터레코드는 모든 집단에게 이해될 수 있을 것이다.

3 Dublin Core의 기본요소

DC에서는 인터넷환경에서 기본적으로 DLO를 찾을 수 있는 최소의 메타데이터요소로 15개를 우선 선정하였다. DLO(*document-like objects*)란 신문기사나 사전 등의 전자판을 말한다. DLO는 기본이 텍스트이며 DLO의 메타데이터는 전통적인 인쇄텍스트를 기술하는 메타데이터와 매우 유사하다. 기본요소와 함께 *scheme, type, role*이라는 세 가지 한정어와 해당되는 다양한 값들을 이용하여 기술한다. 지금까지 논의되고 있는 15종류의 기본요소(레이블)와 한정어의 종류를 보면 다음과 같다(Knight & Hamilton, 1997). 설명없이 값만을 나열한다.

1) Subject and Keywords(SUBJECT)
   - Scheme - Internal; LCSH; UDC; DDC; NLM; MeSH; Colon; JEL; RCHME; AAT; ULAN; TGN; ICONCLASS; SHIC2; TGM1; MSC; YKL; SAB
   - Type - Keyword; Notes

2) Description(DESCRIPTION)
   - Scheme - Internal; URL; URN
   - Type - Freetext; Abstract

3) Title(TITLE)
   - Scheme - Internal; AACR2
   - Type - Main; Long; Short; Alternative; Subtitle; PartTitle; Spine; Translated

4) Creator or Author(CREATOR)
   - Scheme - Internal; USMARC
   - Type - Name; Email; Postal; Phone; Fax; Affiliation; HomeEmail; HomePostal; HomePhone; HomeFax; Homepage; Keywords; Order

5) Publisher(PUBLISHER)
   - Scheme - CREATOR의 내용과 동일.
   - Type - CREATOR의 내용과 동일.

6) Other Contributor(CONTRIBUTORS)
   - Scheme - CREATOR의 내용과 동일.
   - Type - CREATOR의 내용과 동일.
   - Role - Editor; Illustrator; Binder; Translator; MachineReadableCreator; Sponsor; Compiler; Funder; Composer; Cataloger; Contact; Reviewer; Proofreader

7) Date(DATE)
   - Scheme - *IETF.RFC-822; ANSI.X3.30-1985; ISO.31-1: 1992; FGDC; SSE*
   - Type - Creation; Current; Modified; ValidFrom; ValidTo

8) Resource Type(TYPE)
   - Scheme - DCObjects; Freetext
   - Type - *Advertisment; Article; Bibliography; Book;* Booklet; Collection; Course Material; Dataset; Honours Thesis; Image; In Book; In Collection; In Proceedings; Journal; Magazine; Manual; Masters Thesis; Message On Moderated Mailing List; Message On Unmoderated Mailing List; Misc; Music; Newspaper; Organisation Info; PhD Thesis; Personal Info; Poem; Posting To Moderated Newsgroup; Posting to Unmoderated Newsgroup; Preprint; Proceedings; Research Paper; Service; Tech Report; Unpublished; Unrefereed Article; Video

9) Format(FORMAT)
   - Scheme - IMT; MIME; Freetext
   - Type - text/html; text/plain; image/gif; application/x-gzip; application/x-compress; application/x-ns-proxy-autoconfig; application/x-javascript; application/x-tcl; application/x-csh; application/

x-postscript; application/octet- stream; application/x-cpio; application/x-gtar; application/x-tar; application/x-shar; application/x-zip-compressed; application/x-stuffit; application/mac-binhex40; video/x-msvideo; video/quicktime; video/mpeg; audio/x-wav; audio/x-aiff; audio/basic; application/fractals; image/ef; image/x-MS-bmp; image/x-rgb; image/x-portable-pixmap; image/x-portable-bitmap; image/xwindowdump; image/x-pixmap; image/x-bitmap; image/x-cmu-raster; image/tiff; application/x-texinfo; application/x-dvi; application/x-latex; application/x-tex; application/rtf

10) Resource Identifier(IDENTIFIER)
- Scheme - URL; URN; ISBN; ISSN; FPI; SICI; Version
- Type - Primary; Copy

11) Relation(RELATION)
- Scheme - URL; URN; ISBN; ISSN; FPI;
- Type - IsParentOf; IsChildOf; IsMemberOf; IsDrivedFrom; HasBibliographicInfoIn; IsRevisionHistoryFor; IsCriticalReviewOf; IsOverviewOf; IsContentRatingFor; IsTermsAndConditionsFor; IsDataFor

12) Source(SOURCE)
- Scheme - Freetext; URL; URN; ISBN; ISSN; FPI
- Type

13) Language(LANGUAGE)
- Scheme - Freetext; Z3953 ;ISO.639; Computer
- Type

14) Coverage(COVERAGE)
- Scheme - LatLong; OSGS; ANSI.X3.30-1875; Freetext
- Type - Spatial; Temporal

15) Right Management(RIGHTS)
- Scheme - Freetext; URL; URN
- Type

4 제정원칙

DC는 고유성, 확장성 구문독립성, 선택성, 반복성, 수정가능성이라는 여섯 가지 기본원칙을 정하고 이를 만족시킬 수 있도록 기본요소를 개발하였다(Weibel, Godby & Miller, 1995).

고유성이란 오브젝트의 고유한 특성만을 기술하게 하는 것이다. 즉, 주제요소는 고유데이터이지만 가격이나 액세스 관련사항과 같은 변동정보는 비고유한 데이터가 된다. 확장성이란 작은 기본요소집합만으로는 적절하게 기술할 수 없는 데이터를 고유데이터에 추가할 수 있는 확장기구의 추가를 말한다. 특수목적이나 특정분야를 위해 나름대로의 추가해야 할 내용도 있을 것이며, DC 자체가 바뀔 수도 있다. 구문독립성이란 영역이나 응용프로그램에 따라 융통성 있게 적용할 수 있도록 구문구조를 고착시키지 않는 것을 말한다. 선택성이란 각 요소에 대하여 선택적이라는 것이다. 어떤 요소가 특정 오브젝트에서는 의미가 없을 수도 있기 때문이다. 반복성이란 공저자일 때와 같이 모든 동일한 요소는 반복해서 사용할 수 있음을 말한다. 수정가능성이란 모든 요소는 자기설명력을 가지고 있지만 상이한 분야의 요구를 만족시키기 위하여 한정어로서 각 요소를 수정할 수 있음을 말한다. 한정어가 없으면 그 요소는 보편타당한 의미를 갖는다.

5 기본요소의 확장

확장성은 어떤 메타데이터시스템에서도 기본적인 것이다. 왜냐하면 아무리 큰 메타데이터요소를 갖는다 하더라도 모든 종류의 정보원에 적합할 수는 없기 때문이다. 그러나 코어요소집합이 커지면 문제를 복잡하게 한다. 왜냐하면 요소집합이 커지면 다양한 이용자집단이 이해하기 어려워지기 때문이다. 이를 위하여 DC는 작지만 다음과 같이 세 가지 측면에서 확장성이 좋도록 설계되었다.

첫째, 르컬이 레코드에 필드를 추가할 수 있다. 이 추가필드는 제안한 그룹의 바깥에서 이해하리라 보증할 수는 없지만 그 코어요소를 이해하는 시스템에서는 에러를 일으키지 않는다. 이 확장은 텍스트의 비구조화 문자열 혹은 다른 표준레코드나 자신을 가리키는 포인터일 수도 있다.

둘째, 전술한 "*scheme*"이라는 하부요소가 확장기구의 역할을 한다. "*scheme*"에 대한 값의 집합은 자유로우며 이는 이용자그룹에서 만들어 쓸 수 있기 때문에 DC에서는 정의하지 않는다. Subject 등과 같은 몇 요소에서는 주제어가 생성된 분류계층이 한정되기 때문에 값의 집합이 작다. 어떤 요소에서는 기술할 정보원이 복잡하기 때문에 *scheme*의 값이 매우 클 수가 있다. 예를 들면, Identifier요소는 FTP, URL, URN, ISBN 등과 로컬에서 부여하는 여러 *scheme*이 있을 수 있다.

셋째, DC 자체를 레이블링할 수 있는 확장기구를 갖는다. 이것은 버전번호라고 이해할 수 있다. 기본집합에 새로운 요소가 추가되거나 기존의 요소가 의미를 바꾸게 되면 버전번호를 바꿀 수 있다. DC의 최근 버전은 0.1이다.

6 기타 메타데이터 형식

현재 사용되고 있는 메타데이터형식은 매우 다양하다. 예를 들면, MARC, TEI(Text Encoding Initiative) 이외에도 EAD(Encoded Archival Description), Geospatial Metadata, GILS(Government Information Locator Service), Handles, IAFA(Internet Anonymous FTP Archives), MCF (Meta Content Format), PICS(Platform for Internet Content Selection), RDM(Resource Description Messages), RFC1807(A Format for Bibliographic Record), ROADS (Resource Organization and Discovery in Subject-based Services), SHOE(Simple HTML Ontology Extensions), SOIF(Summary Object Interchange Format), X3L8(ANSI Standard for Data Representation), URC, Nordic 등의 수 많은 메타데이터형식이 있다.

이 중 MARC, TEI 등을 제외하고는 특정 분야의 특수한 목적으로 사용되고 있는 것이 대부분이다. 그 중 일반적인 문헌처리에 가까운 형식중의 하나인 SOIF형식을 DC와 비교하기 위하여 좀 더 상세히 보면 다음과 같다. SOIF의 형식은 속성-값의 쌍으로 구성된다. Netscape에서 사용하고 있으며, 일반 텍스트, SGML (및 HTML), PostScript, MIF, RTF 등의 형식에서 자동으로도 생성할 수 있는 프로그램을 준비하고 있다. SOIF의 속성명을 설명없이 나열하여 보면 다음과 같다. 기본기술요소는 Abstract, Author, Description, Keyword, Title이며, 관리용 메타데이터 요소로 Gatherer-Host, Gatherer-Name, Gatherer-Port, Gatherer-Version, Refresh-Rate, Update-Time, Last-Modification-Time, MD5, Time-to-Live 등을 갖고 있다. 특히 주제요소는 갖지 않으며 Type속성(다시 세분)에서 기술할 수 있다. 이외에 URL-References가, Type과 연계하여 File-Size, Full-Text가, 기타로 MaintEmail, Version, CopyPolicy 등이 있다.

7 결론

전 장에서 SOIF형식의 예를 들었으나 이와 같이 대부분의 형식이 서로 상이하여 호환성의 문제가 있으므로 이들의 고려사항을 통합하여 하나의 메타데이터형식을 도출해내는 것이 중요한 일이다. 이를 위하여 DC가 출현했지만 아직 완전한 해결책은 아니다. 그러나 DC는 벌써 미국을 위시하여 많은 실제의 시스템에서 그대로 혹은 일부를 변형하여 사용하고 있다는 데 주목할 필요가 있다.

메타데이터는 기본적으로 만들기 쉽고, 색인이 쉽고, 全文보다 정확률이 높으며, 호환성이 좋아야 한다는 것이 중요하다. DC로 충분한가? 우리나라 자료용으로는 필요없는 요소도 있을 것이며, 추가되어야 할 특성도 있을 것다. 예를 들면, 각 자료에 대한 페이지, 판차, 권호 등에 대한 기술은 어떻게 할 것인가? 파일이나 각 요소의 크기와 같이 일부 정보는 명시적으로 기술할 것이 아니라 해당 시스템이 암묵적으로 가질 수도 있을 것이다.

국내에서도 인터넷 정보자원의 메타데이터의 관리를 위하여 SeriCore(이원석 등, 1997)를 설계한 바 있다. DC의 경향 등을 주시하면서 보다 일반적인 우리의 메타데이터 표준형식을 갖는 것이 시급한 일이라 사료된다.

참고문헌

이원석 등(1997). 인터넷 메타데이터 검색 및 관리시스템의 설계 및 구현. 「한국문헌정보학회지」, 13(2): 199-216.

The Harvest Information Discovery and Access System. (http://harvest.transarc.com/).

Knight, Jon, Martin Hamilton(1997). *Dublin Core Qualifiers.* (http://www.roads.lut.ac.uk/Metadata/DC-SubElements.html).

Weibel, Stuart, Jean Godby & Eric Miller(1995). *OCLC/NCSA Metadata Workshop Report.* (http://www.org:5046/oclc/research...ences/metadata/dublin_core_report. html).