학술지 기사에 대한 메타데이터 품질의 계량화 방법에 관한 연구*

A Study on Quantitative Measurement of Metadata Quality for Journal Articles

이용구(Yong-Gu Lee)**
김병규(Byung-Kyu Kim)***

초 록

기존 메타데이터의 품질 측정 방법은 오류가 발생한 레코드를 단순히 계수하여 그 비율로 품질을 측정하였다. 이러한 한계를 극복하기 위해 메타데이터 요소별로 상대적 중요 정도를 나타내는 가중치를 적용함으로서, 메타데이터 품질을 체계적으로 계량화 하는 측정 방법을 제시하고자 하였다. 구체적인 가중치 부여 방법으로 엔트로피, 이용자 과업, 그리고 이용 통계를 활용하였다. 또한 이들을 결합하여 통합 가중치를 제시하고 실제 서비스 되고 있는 학술지 기사 메타데이터에 적용하였다. 실험 결과, 엔트로피 가중치 방법은 데이터 자체의 특성을 잘 반영하며, 이용자 과업을 적용한 방법은 이용자의 정보요구를 해결하는 필요한 메타데이터 요소를 제시하며, 통합 가중치는 특정 메타데이터 요소의 오류에 영향을 받지 않으면서 균형 잡힌 측정값을 제시하여 계량화 방법에 적합한 것으로 나타났다.

ABSTRACT

Most metadata quality measurement employ simple techniques by counting error records. This study presents a new quantitative measurement of metadata quality using advanced weighting schemes in order to overcome the limitations of exiting measurement techniques. Entropy, user tasks, and usage statistics were used to calculate the weights. Integrated weights were presented by combining these weights and were applied to actual journal article metadata. Entropy weights were found to reflect the characteristics of the data itself. User tasks presented the required metadata elements to solve user's information need. Integrated weights showed balanced measures without being affected by the influence of error elements, This finding indicates the new method being suitable for quantitative measurement of metadata quality.

키워드: 메타데이터 품질, 계량화 방법, 엔트로피, 이용자 과업, 이용통계
metadata quality, quantitative measurement, entropy, user task, usage statistics

---

* 이 논문은 2010년도 제17회 한국정보관리학회 추계학술발표회에서 발표한 것을 수정・보완한 것임.
** 계명대학교 문헌정보학과(yonggulee@kmu.ac.kr) (제1저자)
*** 한국과학기술정보연구원 지식기반실 선임연구원(yourovin@kisti.re.kr) (공동저자)
■ 논문접수일자 : 2011년 3월 18일 ■ 최초심사일자 : 2011년 3월 18일 ■ 게재확정일자 : 2011년 3월 21일
■ 정보관리학회지, 28(1): 309-326, 2011. 〔DOI:10.3743/KOSIM.2011.28.1.309〕

1. 서 론

오늘날과 같은 정보사회에서 직면한 중요한 문제 중에 하나는 다양한 정보원이나 정보 유형뿐만 아니라 이들이 담고 있는 방대한 양에 대한 검색이다. 이는 검색과 관련된 기술 자체도 중요하지만, 해당 정보를 검색하기 위한 접근점(access point)인 메타데이터 또한 중요하다.

메타데이터와 같은 색인은 정보원과 이용자를 연결시켜주는 매개체로서 특정 주제 분야의 문헌을 탐색하거나 연구의 성과를 조사하고자 하는 사람들에게 요구되는 문헌을 찾을 수 있는 정보, 특정 문헌을 내포하고 있는 문헌의 서지사항 또는 문헌의 소재 등을 알려주는 검색, 탐색 및 식별도구로 알려져 있다(윤구호 2001).

메타데이터는 대상 자원에 대한 접근, 식별, 선정에서 중요한 역할을 담당하므로 메타데이터의 품질은 매우 중요하다. 왜냐하면 좋은 품질의 메타데이터는 좋은 검색 성능을 가져오는데 일조하기 때문이다. 이러한 메타데이터의 품질이 연구자에게 미치는 영향에 대해, 이응봉 등(2001)은 우리나라의 과학기술 분야의 데이터베이스 평가와 이를 통한 지속적인 품질향상은 연구자가 필요로 하는 데이터의 탐색과 수집에 소요되는 노력과 시간을 단축하고, 실험과 분석에 보다 많은 시간을 투자하게 함으로써, 양과 질적으로 우수한 연구물의 생산과 직결된다고 하였다. 또한 효율적인 정보환경을 조성함으로써, 관련분야 연구 인력의 연구생산성을 증대하고, 연구자들 사이의 동일 연구과제에 대한 중복연구를 방지하여 국가 연구생산성을 향상시킬 수 있다고 하였다.

메타데이터 품질측정과 관련하여 품질측정의 목적과 평가대상 그리고 평가주체에 따라 다양한 평가기준과 절차 그리고 방법이 적용될 수 있다. 다만 메타데이터의 품질측정을 위한 대부분의 기존 연구들은 원본과 비교하여 메타데이터의 옳고 그름에 대해 판정하고 그 수가 어느 정도인지 계수하는, 즉 단순한 수치만 나열하는데 그쳤다. 예를 들어 100건의 서지레코드를 평가하여 5건의 오류를 발견했다면, 어떤 메타데이터 요소가 오류가 나든, 모두 5%의 오류율을 가지게 된다. 또한 표제나 저자보다 페이지 정보는 이용자에게는 상대적으로 덜 중요하지만, 기존 연구들은 표제나 저자가 잘못 기입된 경우와 시작 페이지나 끝 페이지가 잘못 기입된 경우를 동일하게 평가하였다. 이는 메타데이터 요소에 대한 상대적인 중요도 내지 가치를 반영하지 못하고 있음을 뜻한다.

따라서 이 연구에서는 메타데이터의 요소와 메타데이터를 이용하거나 관리하는 대상에 따른 상대적 중요도를 반영하는, 보다 체계화된 메타데이터 품질측정 방법을 제시하고자 하였다. 또한 더 나아가 이러한 개념에 기초하여 품질 측정에서 계량화 내지 수량화한 측정방법을 실제로 운용되는 학술지 기사 메타데이터 데이터베이스를 대상으로 적용하고자 하였다.

2. 이론적 배경

가장 간결한 메타데이터 정의는 '메타데이터는 데이터에 관한 데이터'란 것이다. Caplan(2003)의 이 정의를 보다 자세히 기술하면, 메타데이터는 어떤 매체의 유형에 저장되거나 어떤 포맷으로 이루어진 정보원에 관한 구조화된

정보를 의미한다. 또한 Haynes(2004)는 메타데이터를 데이터 레코드나 정보원의 내용, 포맷, 속성 등을 기술하는 데이터로 정의하였으며, 구조화된 자원이나 텍스트 문서와 같이 비구조화된 자원을 기술하는데 이용할 수 있다고 하였다. 즉, 메타데이터는 전자 자원, 디지털 데이터, 그리고 책이나 학술지, 보고서 등과 같은 인쇄된 문헌을 기술(description)하는데 적용되며, 웹 자원과 같이 정보자원 내에 포함될 수도 있고 데이터베이스처럼 따로 유지될 수 있다.

디지털화된 정보자원 뿐만 아니라 기존의 매체에 대해 정보검색과 관리 측면에서 메타데이터는 다음과 같이 중요한 위치를 차지한다(Haynes 2004).

첫째, 메타데이터는 검색 성능을 향상시킨다. 메타데이터는 자원의 내용에 대해 개별적인 기술 요소를 만들어서 검색을 향상시킬 수 있다. 둘째, 메타데이터는 전자화된 디지털 객체를 관리하는 방법을 제공한다. 셋째, 메타데이터는 데이터의 진정성(authenticity)을 결정하는데 도움을 줄 수 있다. 넷째, 메타데이터는 상호운용성(interoperability)에서 중요한 역할을 한다. 다섯째, 인터넷 포털에서 사용되는 소프트웨어는 메타데이터 표준을 따르며 이들을 이용하여 웹사이트의 콘텐츠에 대한 접근점을 제공한다.

여러 문헌들을 살펴보면, 메타데이터는 정보자원의 기술 관점과 목적에 따라 크게 다섯 가지 범주로 구분되어진다(오동근 2004; Intner, Lazinger and Weihs 2006; 김태수 2008).

① 기술 메타데이터(descriptive metadata): 정보자원과 그 내용에 관련된 사항이면서 동시에 정보자원을 탐색하기 위한 메타데이터이다. 발견(discovery, 어떤 자원을 어떻게 찾아내는가), 식별(어떤 자원을 다른 유사한 자원과 어떻게 구별할 수 있는가), 선정(어떤 자원이 특정의 요구를 어떻게 결정하는가)이라는 목적을 충족시켜 주는 것으로 이해되고 있다. 기술 메타데이터는 또한 병치(collocation, 어떤 저작의 모든 버전들을 함께 모으는 것)와 수집(aquisition, 특정 자원의 한 카피를 입수하거나 그에 접근하는 것)을 위해서도 사용할 수 있다.

② 관리 메타데이터(administrative metadata): 정보자원의 보존이나 접근 제어용 메타데이터이다. 자원의 관리를 용이하게 하기 위한 정보이며, 컨텐트(내용)에 대한 접근을 관리하거나 컨텐트를 아카이빙(archiving)하는 책임을 누가 가지는지에 대한 정보이다. 또한 그와 관련하여 어떤 통제나 처리가 수행되고 있고 접근이나 이용에 대한 어떤 제약이 적용되고 있는지와 같은 정보도 포함된다.

③ 구조 메타데이터(structural metadata): 정보자원의 물리적, 논리적 내부구조에 관한 메타데이터이며, 복합적인 디지털 객체들을 함께 묶어주는 것으로 생각할 수 있다.

④ 기술 메타데이터(technical metadata): 정보자원의 이용에 필요한 기술(technical) 요건에 관한 메타데이터이다.

⑤ 보존 메타데이터(preservation metadata): 정보자원의 보존 관리에 관한 메타데이터이다.

앞서 살펴본 바와 같이 메타데이터는 정보자원에 대한 식별하고 기술하기 위해 필요할 뿐만 아니라 그 자원이 어떻게 다루어져야 하는지를 나타내야 하며, 그 자원의 기능이나 이용 그리고 다른 자원과의 관계에서 어떻게 관리되어야 하는지를 나타내어야 한다. 이러한 측면들을 포함하기 위해 이러한 다섯 가지의 메타데이터 유형이 필요하다.

국내의 메타데이터 품질측정에 관한 연구들을 살펴보면, 대부분은 다양한 평가지표를 이용하여 오류가 발생한 서지 레코드의 건수나 비율을 나열하였다. 이들 연구를 간략히 기술하면 다음과 같다.

이제환(2002)은 한국교육학술정보원이 구축한 종합목록 DB의 품질을 검증하고 이를 개선하기 위한 방안을 제시하기 위한 실행 모델을 제시하고 품질평가 기준인 다양한 지표를 제시하였다. 또한 이응봉 등(2001)은 과학기술분야의 데이터베이스에 대한 품질평가 기준 및 방법을 개발하고, 이를 실제 데이터베이스에 실증적으로 적용하였다. 평가 기준으로 데이터베이스가 수록한 데이터 자체와 데이터가 사용자에게 서비스되는 측면으로 평가 차원으로 나누어 10가지의 품질 기준을 제시하였다.

최근 외국의 선행연구(Ochoa and Duval 2006; Stvilia et al. 2007)는 정보품질(information quality)을 평가하기 위한 일반적인 프레임웍을 제안하였는데, 이는 정보품질의 문제에 대한 근원을 포괄적으로 밝히고 이를 적합한 이론과 실제에 따라 체계적으로 나눈 여러 차원(dimensions)에 해당하는 평가기준을 정의하고 지표를 제시한 후, 이를 실제 더블린 코어 레코드와 온라인 백과사전의 기사에 대해 적용하여 유효한지 확인하였다.

또한 Stvilia와 Gasser(2007)는 메타데이터 품질에 대한 가치 중심 평가(value-based assessment) 방법을 제시하고, 이를 더블린 코어 메타데이터 레코드에 적용하여 메타데이터 제공자와 최종 이용자가 같은 메타데이터에 대해 서로 다른 가치를 갖고 있음을 제시하였으며, 메타데이터의 더 좋은 이용을 촉진하기 위해 평가 모형을 최종 이용자에게 명백하게 하고 그들의 참여를 유도해야 한다고 주장하였다.

3. 품질측정 계량화 방법

3.1 실험 데이터 및 실험과정

이 연구의 전반적인 과정을 설명하면, 가장 먼저 메타데이터의 품질을 측정하기 위한 측정지표의 선정 및 정의를 내리고 이들 측정지표에 적용할 가중치를 부여하는 방법을 제시하였다. 이 연구에서 제시된 측정지표로는 완결성과 정확성을 선정하였다. 이들에게 적용할 가중치 부여 방법으로는, 첫째 메타데이터 자체가 가지는 중요도를 계산하기 위한 엔트로피(평균 정보량)를 이용하였으며, 둘째 이용자가 메타데이터를 통해 수행하고자 하는 과업(user task)을 계량화하여 이를 가중치로 이용하였다. 셋째 실제 웹서버와 검색엔진의 트랜잭션 로그 데이터로부터 메타데이터 요소별 이용 비율을 추출하여 가중치를 산출하였다. 세 가지 부여 방법을 조합하여 통합된 가중치 부여 방법을 통해 최종적인 평가를 수행하였다.

일반적으로 메타데이터의 품질을 측정하기

위해서는 측정 대상 데이터로부터 무작위로 표본을 추출하고, 이 표본에 대해 원본과의 비교를 통해 이미 작성된 메타데이터의 오류 여부를 체크하여 품질을 측정한다. 또는 특정 키워드를 이용하여 메타데이터를 검색한 후, 검색 결과에 대해 이와 같은 과정을 거치기도 한다.

이 연구에서는 실험 데이터로 한국과학기술정보연구원(KISTI)의 과학기술학회마을(http://society.kisti.re.kr/)에서 제공하는 학술지 기사 메타데이터를 선정하고, 품질을 측정하였다. 이 연구는 2005년부터 2009년까지 5년 동안 구축된 601종의 학술지와 151,908건의 기사 메타데이터를 이용하였다. 또한 메타데이터의 오류를 체크하기 위해 검증할 표본은 2009년에 구축된 데이터 중에 3%를 임의추출 하였다. 추출된 기사 레코드는 974건이며, 이 기사를 수록한 학술지는 372종이다.

추출된 학술지 기사 표본에 대해 KISTI의 표준 업무 절차에 따라 기 구축된 메타데이터를 일차적으로 교차검증을 하고 최종 검토를 하는 방식으로 2차에 걸친 오류체크를 수행하였다. 이에 따른 결과는 〈표 6〉과 같다.

3.2 품질 측정 지표

일반적으로 데이터베이스 내에서 데이터의 품질은 데이터의 정확성(accuracy), 완결성 또는 완전성(completeness), 최신성, 일관성 등을 평가지표로 이용하여 측정한다. 또한 서비스의 품질은 주로 서비스의 이용관점에서 검색성, 사용용이성, 비용 등을 이용하여 측정한다. 이 연구에서는 학술지 기사의 메타데이터 품질을 측정하기 위하여 다양한 평가지표 중에 가장 자주 사용되는 정확성과 완결성을 중심으로 측정하고자 한다.

3.2.1 완결성

완결성은 메타데이터를 이용하여 자원을 완전히 기술하는 정도로 정의할 수 있는데, 좀 더 구체화 하면 대상 자원에 대한 완전한 표현으로 정의할 수 있다. 즉 완결성은 자원을 보다 상세하게 기술함을 정의하고 그 정도를 측정하는 방법으로, 대상 자원인 원문에 기반을 두어 메타데이터 요소를 누락 없이 얼마나 충실하게 기술하는가를 측정하는 방법으로 다음과 같은 두 가지로 계산할 수 있다(Ochoa and Duval 2006). 각각의 방법을 구체적인 공식으로 표현하면, 〈식 1〉과 〈식 2〉와 같다. 먼저 〈식 1〉은 메타데이터 스키마 요소 중에 누락(Null)을 제외한 비율로, 해당 정보자원에 대한 메타데이터를 보다 완전하게 표현한 정도를 나타낸다. 이 식은 메타데이터 요소에서 값의 유무를 나타내는 이진방식이라고 볼 수 있다.[1)]

$$Q_{binary} = \frac{\sum_{i=1}^{N} P(i)}{N} \qquad \langle 식\ 1 \rangle$$

〈식 1〉에서 N은 메타데이터 스키마에서 반드시 기술해야 하는 요소의 수이다. 그리고 $P(i)$는 해당 요소가 값이 널(누락)이 아니면 1, 널이면

1) 이 연구에서는 메타데이터 요소별 기술수준을 필수 요소와 해당시 필수 요소로 나누어 적용하였다. 예를 들어, 학술지의 대등서명이 존재하면 해당시 필수로 지정하였다. 다만, 메타데이터 요소에 대한 정확한 기술 수준을 정하는 것은 이 연구의 범위를 벗어나므로 생략하였다.

0으로 이진 값으로 자원을 얼마나 상세히 기술하였는지 체크한다. 예를 들면, 10개의 필수에 해당하는 요소 중에 2개의 요소를 누락하였다면, 해당 자원에 대한 완결성 품질 측정값은 0.8에 해당한다.

아래 〈식 2〉는 메타데이터 요소의 상대적 중요도에 따른 가중치 비율을 이용하여 완결성을 측정하는 방법이다. 즉 메타데이터의 전체 요소들이 모두 동일하게 중요한 의미를 가지지 않기 때문에, 보다 중요한 의미를 갖는 요소들의 값이 제대로 기술되어졌는지 반영하는 가중치 방법의 공식이다.

$$Q_{weight} = \frac{\sum_{i=1}^{N} \alpha_i * P(i)}{\sum_{i=1}^{N} \alpha_i} \qquad \langle 식\ 2 \rangle$$

〈식 2〉에서 a_i는 i번째 요소의 상대적 중요도를 나타내는 가중치로, 메타데이터의 전체 요소 N개마다 서로 다른 가중치를 가진다. 이 연구에서는 이러한 가중치를 구하기 위하여 3.3절에서 세 가지 방법을 제시하였다. 따라서 〈식 2〉는 메타데이터의 모든 요소 중에 그 값이 누락된 요소를 제외한 나머지 요소들의 가중치 합(식의 분자)을 전체 요소의 가중치 합(식의 분모)으로 나눈 비율을 뜻한다. 예를 들어 특정 요소의 가중치가 0.5인 요소가 누락된 경우, 가중치 0.1을 갖는 요소가 누락된 경우보다 품질 측정에서 더 작은 값을 가지므로 그만큼 중요함을 의미한다.

3.2.2 정확성

정확성은 메타데이터가 자원을 올바르게 기술하는 정도로 정의할 수 있다. 좀 더 구체적으로 이 정의는 메타데이터가 입력지침(문법 규칙과 구문 규칙)을 준수하는 정도를 의미하며, 이는 철자 오류, 띄어쓰기, 철자 누락, 일관되지 않은 형식, 부적합한 값의 입력 등을 말한다. 따라서 철자 오류, 띄어쓰기, 누락 등과 같은 오류로 인해 입력된 데이터의 부정확 정도를 측정하는 평가지표이다.

이러한 정확성을 측정하기 위한 방법으로, 앞서 완결성에서 제시되었던 측정 공식인 〈식 1〉과 〈식 2〉를 그대로 적용하였다. 하나의 메타데이터 레코드에 대해 철자나 띄어쓰기 등과 같은 오류가 발생한 요소에 해당하는 값만큼 제외함으로써 품질 측정값을 떨어뜨려 정확성을 낮추는 결과를 가져오도록 하였다.

지금까지 하나의 메타데이터 레코드에 대해 〈식 1〉과 〈식 2〉를 이용하여 완결성과 정확성을 측정하였다면, 이를 추출된 표본 전체 메타데이터 레코드에 적용하고 그 값이 더하여 최종적인 완결성과 정확성을 측정한다. 이를 식으로 나타내면 〈식 3〉과 같다.

$$TQ = \frac{\sum_{i=1}^{R} Q_i}{R} \qquad \langle 식\ 3 \rangle$$

〈식 3〉에서 R은 검증 대상인 전체 메타데이터 레코드의 수이다. 또한 Q_i는 i번째 메타데이터 레코드의 완결성 또는 정확성 측정값으로, 각각의 측정지표에 대해 단순 이진 공식과 가중치 공식인 〈식 1〉과 〈식 2〉의 두 가지 값을 갖는다.

3.3 가중치 산출 방법

메타데이터는 정보자원의 내용을 기술하고 탐색하기 위한 목적을 갖는다. 이를 관련 주체별로 살펴보면, 하나는 메타데이터의 생성과 관리를 책임지고 있는 정보전문가 또는 사서이고, 다른 하나는 이렇게 구축된 메타데이터를 이용하여 정보를 획득하는 최종 이용자이다. 또한 이들이 이용하는 대상인 메타데이터 자체가 가지는 고유한 특성 내지 성질도 이용패턴에 영향을 미친다. 따라서 메타데이터 품질을 논의할 때 메타데이터 자체, 관리자 입장인 정보전문가(사서), 그리고 이용자 측면에서 메타데이터 요소별 가중치 부여를 고려하여야 한다. 따라서 이 연구에서는 가중치를 산출하거나 부여하기 위해 메타데이터 자체의 엔트로피, 이용자 과업의 계량화, 메타데이터 요소별 이용통계 비율을 활용하였다.

3.3.1 데이터 기반 가중치

정보원에서 발생하는 메시지의 평균 정보량을 엔트로피라고 한다. 엔트로피는 정보가 전달되기 이전의 불확실성의 정도를 나타내며, 특정 지식상태에서 입수된 정보에 의해 변화된 지식상태에 따라 결정된다. 이에 따라서 드물거나 뜻밖인 사건일수록 지식상태의 변화를 크게 유발하므로 정보량도 크다고 할 수 있다(사공철 외 2001. 281).

정보를 부호화할 때에는 가급적 빈번히 발생하는 메시지는 정보량이 작다. 이를 더블린 코어의 메타데이터 요소에 적용시킨 연구(Stvilia et al. 2004. 121)에 의하면, 표제와 생성자(creator) 요소는 파일형식(format)과 언어 요소에 비해 보다 많은 엔트로피를 갖고 있어 전달하는 정보량이 크다. 즉 표제나 생성자 요소로 검색하는 것이 언어 요소로 검색하는 것보다 더 효과적임을 뜻한다. 왜냐하면, 높은 엔트로피를 갖는 요소가 특정성이 높아 검색 결과를 줄여주기 때문이다. 이러한 개념을 이용하여 메타데이터의 요소별 상대적 중요도를 계산할 수 있다.

또한 정보전문가나 최종 이용자의 이용과 무관하게 메타데이터 자체가 서로 다른 엔트로피를 갖는다. 예를 들어 특정 기관에서 한 언어로 된 하나의 학술지만 메타데이터를 구축하여 제공한다면 해당 학술지의 학술지명, 언어, 발행기관명은 모두 동일한 값을 가지게 때문에 이용자는 자신의 정보요구나 검색 패턴에 상관없이 이들 요소를 검색할 필요가 없어진다. 이와 같이 구축된 메타데이터 자체로부터 요소별 엔트로피를 계산하여, 낮은 엔트로피를 갖는 요소는 큰 엔트로피를 갖는 요소에 비해 상대적으로 낮은 중요도(가중치)를 부여할 수 있다.

이 연구에서는 요소별 엔트로피를 구하기 위해 〈식 4〉와 같은 표준화된 엔트로피 공식을 적용하였다. 이 식에서 p_i는 주어진 요소에서 i번째 고유 값의 확률이며, n은 그 요소에서 고유 값의 수이다. 이 식을 5년 동안의 기사용 전체 메타데이터에 적용하여 요소별로 엔트로피를 계산하였으며, 그 결과를 〈표 1〉에 제시하였다. 표의 엔트로피를 메타데이터 요소별 데이터 기반 가중치로 간주하였다. 즉, 엔트로피의 전체 합에서 요소별 엔트로피로 나누어 최종 표준화된 가중치를 구하였다.

$$entropy = -\sum_{i=1}^{n} p_i Log(p_i)/Log(n) \qquad \langle 식\ 4 \rangle$$

〈표 1〉 학술지 기사 메타데이터 요소별 엔트로피 및 가중치

수준	요소명	총수	고유수	엔트로피	가중치
학술지	학술지식별자	151,908	601	0.503	0.0350
	본서명	151,908	601	0.503	0.0350
	대등서명	121,463	455	0.401	0.0279
	부서명	0	0	-	0.0000
	축약서명	415	2	0.001	0.0001
	진행서명/후속서명	44,125	177	0.121	0.0084
	발행자명	151,908	421	0.464	0.0323
	발행국	151,908	1	0.000	0.0000
	URL	151,908	601	0.503	0.0350
	ISSN	151,053	584	0.499	0.0347
	학술지형식	151,908	5	0.079	0.0055
	학술지분류	64,881	110	0.177	0.0123
	간기	151,908	12	0.129	0.0090
	기본 언어	151,908	2	0.037	0.0026
	권/호 정보	151,215	203	0.192	0.0134
	발행일자	151,908	48	0.153	0.0106
	총 기사수	151,908	244	0.440	0.0306
기사	기사식별자	151,908	151,908	1.000	0.0696
	본서명(한글)	151,908	151,580	1.000	0.0696
	대등서명(영문)	110,336	110,188	0.726	0.0505
	DOI/URI	151,908	151,808	1.000	0.0696
	페이지-시작/끝	151,908	6,591	0.557	0.0388
	키워드	138,377	137,409	0.910	0.0633
	초록	139,559	139,424	0.919	0.0639
	원문정보	151,908	151,808	1.000	0.0696
	원문형식	151,908	1	0.000	0.0000
	참고문헌수	151,429	197	0.310	0.0216
	언어	151,908	5	0.047	0.0033
	저자명	151,900	129,664	0.978	0.0680
	소속	11,900	115,077	0.956	0.0665
	이메일	151,810	55,111	0.553	0.0385
	역할	151,900	235	0.215	0.0150

〈표 1〉에서 총 수는 실험 대상인 학술지 전체 기사 151,908건에 대해 해당 요소가 기술된 횟수를 나타내며, 고유수는 해당 요소가 가지는 고유한 데이터의 빈도를 나타낸다. 이들을 〈식 4〉에 적용하여 엔트로피 값을 구하였다. 표에서 부서명은 입력된 메타데이터가 전무하여 해당 요소의 엔트로피를 구하지 못하였다.

〈표 1〉에서 학술지 수준에서는 가장 큰 엔트로피를 갖는 요소는 학술지 식별자, 본서명, URL, ISSN, 발행자명 순으로 나타났으며, 기사 수준에

서는 기사식별자, 본서명, URI, 원문정보, 저자명, 초록, 키워드 순으로 나타났다. 기사 수준에서 본서명, 식별자, 초록과 키워드는 모두 고유수가 매우 크기 때문이다. 반면 언어, 발행년, 발행국 등의 요소는 매우 낮은 엔트로피를 갖는 것으로 나타났다. 이는 실험 대상이 국내 학술지이므로 언어 요소에서는 한국어와 영어가 대부분이며, 발행년은 모두 2005년에서 2009년까지 총 5년이며, 발행국으로 한국이라는 값을 갖기 때문이다. 이 가중치 기법은 이러한 작은 고유수를 가지는 메타데이터 자체의 특성을 잘 반영하는 것으로 보인다.

3.3.2 이용자 과업 기반 가중치

1997년 IFLA의 목록분과 위원회에 의해 승인된 FRBR(Functional Requirements for Bibliographic Records)은 개체-관계(Entity-Relationship) 모형으로써 오늘날 국제목록원칙규범과 RDA(Resource Description and Access)의 기반이 된다. 즉, FRBR은 데이터베이스에서 사용되는 '개체-관계' 분석 기법을 사용하여 서지데이터의 축적과 제시, 유통에 사용되는 구조를 분석하는 유용한 틀을 제시하였다.

FRBR 모형의 개체-관계 모형은 데이터 요건을 보다 체계적인 방식으로 분석하기 위한 구조를 제공하고 있다. 이 모형에서 제시한 구조는, 서지 레코드를 대상으로 이용자가 수행하는 과업과 관련된 속성과 관계를 중심으로, 서지데이터의 이용을 분석하기 위한 틀을 제시하였다(김태수 2003, 93). 이용자가 메타데이터를 이용할 때 그들이 어떠한 과업을 갖는지 그리고 그 과업에 따라 어떠한 메타데이터 요소가 필요로 하는지를 메타데이터 관리자 측면에서 분석하였다.

또한 가중치와 관련하여 이용자가 서지데이터의 이용과 관련된 4가지 과업(탐색, 식별, 선정, 획득)에 대해 특정 개체의 속성(메타데이터 요소)이나 관계에 대해 상대적인 값으로 표현하였다(김태수 2003, 94). 일례로 이용자 과업에서 식별기호는 탐색과 식별에서 발행지 요소보다 이용자에게 상대적으로 중요하다는 것을 뜻한다. 이는 중요도에 있어서 메타데이터의 요소에 따른 상대적인 차이가 존재함을 보여준다고 할 수 있다. 다만 학술지와 같은 연속간행물에 대해 FRBR에서 제시한 이용자 과업을 적용하기에는 다소 무리가 따른다. 왜냐하면 FRBR의 개체-관계 모형을 이용하여 연속간행물을 표현하고자 하는 노력이 계속되고 있어, 현재 최종적인 합의를 이끌어 내지 못했기 때문이다.

Delsey(2002)는 FRBR과 달리 MARC 태그별로 이용자 과업의 해당여부를 제시하였다. 이 연구에서는 이용자 과업에 대해 세 가지 측면(자원 발견, 자원 이용, 데이터 관리)으로 나누고 각각의 측면에 대해 총 12개의 이용자 과업을 세분화하였다. 즉 자원 발견은 탐색/식별/선정/획득으로, 자원 이용은 제한/관리/운영/해석/식별로, 데이터 관리는 과정/정렬/표현으로 세분화하였다. 일례로 표제저작자사항에 해당하는 245 태그의 데이터요소식별기호인 'a'에 해당하는 본표제는 자원 발견 측면에서 탐색, 식별, 획득 항목에 대해 체크되었다. 즉 본표제는 자원을 발견하기 위한 탐색, 식별, 획득이라는 이용자 과업에 사용되거나 중요함으로 해석할 수 있다. 또한 학술지의 간기에 해당하는 310 태그의 데이터요소식별기호인 'a'에 해당하는 현재 간행빈도는 자원 발견 측면에서

식별과 선정, 자원 이용 측면에서 관리 항목에 체크되어 각각의 이용자 과업에 대해 사용될 수 있다고 표시하였다.

이 연구에서는 학술지와 기사 수준에서 추출된 메타데이터 요소에 대해 MARC 태그로 매핑을 〈표 2〉와 같이 수행하였다. 다만 학술지 수준의 메타데이터 요소의 매핑은 MARC이 연속간행물용에 맞게 기술할 수 있으므로 크게 문제가 되지 않았지만, 기사를 기술하는 MARC 형식이 없으므로 단행본용에 맞게 매핑하였다.

〈표 2〉 학술지 메타데이터 요소와 MARC 태그의 매핑

수준	요소	MARC 태그	지시기호/ 식별기호
학술지	학술지식별자	001	
	본서명	245	a
	대등서명	246	01
	부서명	246	b
	축약서명	210	a
	진행/후속 서명	247	a
	발행자명	260	b
	발행국	260	a
	URL	856	u
	ISSN	022	a
	학술지 형식	008	21
	학술지분류	082	a
	간기	310	a
	기본 언어	041	a
	권/호 정보	362	a
	발행년	260	e
	총 기사수	300	a
기사	기사식별자	001	
	본서명	245	a
	대등서명	246	01
	DOI/URI	856	u
	페이지-시작/끝	300	a
	키워드	650	a
	초록	520	b
	원문정보	533	b
	원문형식	533	a
	참고문헌수	504	b
	언어	041	a
	저자명	245	c
	소속	100	u
	이메일	100	g
	역할	100	e

이 연구에서 사용된 메타데이터의 요소를 〈표 2〉와 같이 MARC 태그로 매핑한 후, Delsey (2002) 연구에서 해당 MARC 태그에 주어진 이용자과업의 수를 더하여 〈표 3〉과 같은 결과를 얻었다. 이 표에서 가중치는 전체 합계에서 요소별 합으로 나눈 비율로 계산하였다.

〈표 3〉에서 가장 높은 가중치를 갖는 요소는 학술지 수준의 기본 언어이며, 기사 수준의 언

〈표 3〉 이용자 과업에 의한 가중치

수준	요소	자원 발견	자원 이용	데이터관리	합계	가중치
학술지	학술지식별자			1	1	0.0147
	본서명	3			3	0.0441
	대등서명			2	2	0.0294
	부서명				0	0.0000
	축약서명	3			3	0.0441
	진행/후속 서명	3			3	0.0441
	발행자명	2			2	0.0294
	발행국	3			3	0.0441
	URL	2			2	0.0294
	ISSN	3			3	0.0441
	학술지 형식		1		1	0.0147
	학술지분류	3			3	0.0441
	간기	2	1		3	0.0441
	기본 언어	3	1		4	0.0588
	권/호 정보	2	1		3	0.0441
	발행년	3			3	0.0441
	총 기사수	2			2	0.0294
기사	기사식별자			1	1	0.0147
	본서명	3			3	0.0441
	대등서명			2	2	0.0294
	DOI/URI	2			2	0.0294
	페이지-시작/끝	2			2	0.0294
	키워드	2			2	0.0294
	초록	1			1	0.0147
	원문정보	3			3	0.0441
	원문형식	1			1	0.0147
	참고문헌수	1			1	0.0147
	언어	3	1		4	0.0588
	저자명	3			3	0.0441
	소속	1			1	0.0147
	이메일				0	0.0000
	역할	1			1	0.0147
합계		57	5	6	68	1.0000

어, 본서명, 축약서명, 진행/후속 서명, 발행국, ISSN 순으로 높게 나타났다. 학술지에 사용되는 기본언어와 기사의 언어가 가장 높은 가중치를 갖게 된 것은 자원 발견 측면에서 식별, 선정, 획득과 자원 이용 측면에서 해석 등의 이용자 과업에서 사용될 수 있다고 체크되어 요소별 합계가 가장 큰 값인 4가 되었기 때문이다. 반면 가장 낮은 가중치를 갖는 요소는 기사 수준의 이메일과 학술지 수준의 부서명이며 역할, 소속, 건수, 원문형식, 초록, 식별자 순으로 낮게 나타났다.

3.3.3 이용 통계 기반 가중치

이용자 입장에서는 메타데이터를 검색할 때 몇 가지 요소만을 대상으로 한다. 즉 이는 관리자 입장과 차이가 있다. 예를 들면 이용자의 경우 학술지의 기사를 검색할 때 대부분 서명이나 저자 요소를 주로 검색한다. 이는 메타데이터 스키마에 포함된 모든 요소들이 이용자의 이용에서 차이가 나며, 당연히 자주 이용하는 요소는 그만큼 이용자에게 중요하게 된다.

Hufford(1991)의 연구를 보면, 메타데이터 요소 중에 가장 많은 이용을 보이는 요소는 단연 서명(33%)이다. 나머지는 이용 정도에 따른 순위를 보면 저자(19.3%), 소장정보(13.7), 청구기호(13.2%)로 나타났다. 이는 메타데이터를 검색하는 시스템이 메타데이터 요소에 대해 어떠한 접근점을 주는냐 그리고 실제 그 기관의 이용자들이 메타데이터 요소에 대해 어떠한 이용 패턴을 보이느냐에 따라 이용률 값이 달라지므로 이용자 중심의 서비스에서는 매우 중요한 의미를 갖는다고 생각된다. 즉 더 많이 자주 이용되는 요소일수록 보다 더 정확하고 완전하게 기술해야 한다. 따라서 이를 메타데이터 스키마의 요소별 중요도인 가중치로 대체시킬 수 있다.

KISTI의 과학기술학회마을 홈페이지에서 보면 검색은 크게 논문검색과 학술지 검색으로 나뉜다. 논문검색은 기본검색과 고급검색으로 나뉘며, 학술지 검색은 학술지명, 학회별, 주제별로 브라우징 할 수 있게 되어 있다. 이 서비스에서 검색을 할 수 있는 요소는 전체, 제목, 저자, 키워드, 초록, 학회명, 학술지명이다. 2010년 9월부터 11월 1일까지 약 두 달 동안 이 서비스의 검색로그를 분석한 결과, 전체 검색건수 143,560번에서 기사명이 82,380건(57.4%), 전체가 31,091건(21.7%), 저자명이 15,757건(11%), 키워드가 13,705건(9.5%), 초록이 631건(0.4%) 순으로 나타났다. 이 중 전체는 나머지 요소를 통합하여 검색하는 것이므로 모두 같은 값을 가지므로 가중치를 계산하는데 제외하였다. 따라서 이용자의 이용 비율에 따른 가중치는 〈표 4〉와 같다. 검색에 이용되지 않는 요소들의 가중치는 이용건수를 구할 수 없으므로 모두 0으로 처리하였다.

〈표 4〉 이용 비율에 따른 이용건수와 가중치

수준	요소	이용건수	가중치
학술지	본서명	1,319	0.0115
	발행자명	480	0.0042
기사	본서명	82,380	0.7209
	키워드	13,705	0.1199
	초록	631	0.0055
	저자명	15,747	0.1378
합계		114,262	1.0000

3.3.4 가중치 비교

서로 다른 측면에서 산출된 가중치를 각각

비교해 보면 일반적으로 메타데이터를 이용하면서 느꼈던 중요도와는 매우 다른 패턴을 보이는 것을 알 수 있다. 이는 각각의 측면을 반영하여 가중치가 달리 제시된 것으로 볼 수 있다. 따라서 이들 세 가지 방법에 대한 통합 가중치를 구하기 위해 먼저 표준화된 각각의 방법별 가중치 모두를 더한 후 전체 합에 따른 비율로 계산하여 〈표 5〉를 얻었다.

〈표 5〉 가중치의 비교

수준	요소	엔트로피 가중치	이용자과업 가중치	이용비율 가중치	통합 가중치
학술지	학술지식별자	0.0350	0.0147	0.0000	0.0166
	본서명	0.0350	0.0441	0.0115	0.0302
	대등서명	0.0279	0.0294	0.0000	0.0191
	부서명	0.0000	0.0000	0.0000	0.0000
	축약서명	0.0001	0.0441	0.0000	0.0147
	진행/후속 서명	0.0084	0.0441	0.0000	0.0175
	발행자명	0.0323	0.0294	0.0042	0.0220
	발행국	0.0000	0.0441	0.0000	0.0147
	URL	0.0350	0.0294	0.0000	0.0215
	ISSN	0.0347	0.0441	0.0000	0.0263
	학술지 형식	0.0055	0.0147	0.0000	0.0067
	학술지분류	0.0123	0.0441	0.0000	0.0188
	간기	0.0090	0.0441	0.0000	0.0177
	기본 언어	0.0026	0.0588	0.0000	0.0205
	권/호 정보	0.0134	0.0441	0.0000	0.0192
	발행년	0.0106	0.0441	0.0000	0.0182
	총 기사수	0.0306	0.0294	0.0000	0.0200
기사	기사식별자	0.0696	0.0147	0.0000	0.0281
	본서명	0.0696	0.0441	0.7209	0.2782
	대등서명	0.0505	0.0294	0.0000	0.0266
	DOI/URI	0.0696	0.0294	0.0000	0.0330
	페이지-시작/끝	0.0388	0.0294	0.0000	0.0227
	키워드	0.0633	0.0294	0.1199	0.0709
	초록	0.0639	0.0147	0.0055	0.0262
	원문정보	0.0696	0.0441	0.0000	0.0379
	원문형식	0.0000	0.0147	0.0000	0.0049
	참고문헌수	0.0216	0.0147	0.0000	0.0121
	언어	0.0033	0.0588	0.0000	0.0207
	저자명	0.0680	0.0441	0.1378	0.0833
	소속	0.0665	0.0147	0.0000	0.0271
	이메일	0.0385	0.0000	0.0000	0.0128
	역할	0.0150	0.0147	0.0000	0.0099
합계		1.0000	1.0000	1.0000	1.0000

〈표 5〉에서 가장 높은 통합 가중치를 갖는 요소는 기사의 본서명(0.2782)이며, 완전명인 저자명(0.0833), 키워드(0.0709), 원문정보(0.0379) 순으로 높은 통합 가중치를 보였다. 기사의 본서명이 다른 요소보다 큰 가중치를 갖는 것은 이용 비율에서 극단적으로 큰 가중치에서 기인한다. 다만 Hufford(1991)의 연구처럼 요소별 편차가 적게 하여 통합 가중치를 만든다면, 각각의 가중치 산출 방법에 의한 결과보다 메타데이터의 세 가지 측면을 모두 반영한 합리적인 가중치 체계를 갖출 것으로 보인다.

4. 실험 데이터의 품질 측정

4.1 요소별 오류 유형

메타데이터를 생성하면서 생긴 오류의 유형을 완결성과 정확성으로 나누어 구분하였다. 누락은 완결성의 오류 유형으로, 띄어쓰기, 오탈자, LATEX, 부적합, 기타 등은 정확성의 오류 유형으로 구분하였다. 표본으로 추출된 기사 974건에 대해 오류를 검증한 결과는 〈표 6〉과 같다.

완결성의 오류 유형인 누락은 전체 51건으로 요소별로 이메일(29), DOI/URI(5), 초록(3)

〈표 6〉 요소별 오류 유형 및 통계

요소 \ 오류유형		누락	띄어쓰기	오탈자	LATEX	부적합	기타	합계
학술지	권정보	2						2
	호정보	1						1
	발행년						1	1
	발행월			4				4
	발행일			3				3
기사	본서명		5	4	4			13
	대등서명	2	4	7	14			27
	DOI/URI	5		1				6
	페이지-시작			3				3
	페이지-끝			4				4
	키워드(한글)	2		6			1	9
	키워드(타언어)	1	6	38	23		4	72
	초록	3	13	32	6			54
	참고문헌수					1		
	저자명(한글)	2					7	9
	저자명(타언어)	1		7			1	8
	소속(한글)	2		4	1		7	14
	소속(타언어)	1		4			9	15
	이메일	29		2			16	47
합계		51	28	119	48	1	46	293

순으로 나타났다. 이메일 이외는 누락 건수가 작아, 오류가 많지 않은 것을 알 수 있다. 또한 정확성을 측정하기 위해 필요한 나머지 오류 중에 가장 많은 유형으로는 오탈자가 119건이며, LATEX(48), 기타(46) 순으로 나타났다. 특히 오탈자의 경우 한글과 나머지 언어로 구분하여 각각에 대해 어느 정도 오류를 내는지 확인하였는데 한글을 제외한 외국어 키워드에서 38건이 발생했으며, 초록의 경우에도 32건이나 발생하였다. 또한 오류에 대한 메타데이터 요소별로 보면, 언어 구분없이 키워드가 81건으로 가장 많고, 초록(54), 이메일(47) 순으로 나타났다. 정확성의 오류유형이 발생한 유일한 레코드 수, 즉 중복 오류를 제거한 레코드는 164개 이며, 누락에서 중복을 제거한 레코드 수는 45건이다. 이 표에서 부적합은 값이 있긴 하나 잘못된 값이 들어간 오류를 말한다.

이 연구에서 쓰인 메타데이터는 구축과정에서 3단계의 걸쳐 사람에 의해 수작으로 검증되며, 데이터베이스에 입력시 무결성을 유지하도록 입력 인터페이스나 자체적인 규칙을 갖고 있다. 무엇보다 지금 이 메타데이터를 통해 국내에 주요 도서관 및 웹에 서비스되고 있기에 해당 기관에서도 메타데이터 오류에 대해 많은 관심과 교차 검증을 하고 있기에 전체적으로 오류가 크지 않다고 볼 수 있다.

4.2 품질 측정 결과

이 연구에서 최종적인 품질 측정 결과를 〈표 7〉에 제시하였다. 이 표에서 계수 방법은 기존의 연구에서 품질측정 방법으로 단순히 오류가 발생한 레코드를 세고 이를 전체 검증 레코드 수로 나눈 비율을 말한다. 두 번째 이진 방법은 표본의 한 레코드내의 메타데이터 요소에 〈식 1〉을 이용하여 오류가 발생하지 않은 요소만을 가중치 없이 계산하였고, 이를 〈식 3〉을 적용하여 전체 표본 레코드의 값을 더하여 측정값을 구하였다. 세 번째로 가중치 방법은 〈식 2〉를 이용하여 메타데이터 요소별 가중치를 반영하였으며, 〈식 3〉을 이용하여 표본 전체 레코드의 합계를 계산하여 측정값을 구하였다. 〈식 2〉의 가중치는 앞서 설명한 〈표 2〉의 엔트로피, 〈표 3〉의 이용자 과업의 계량화, 〈표 4〉의 이용 통계 비율에서 계산된 가중치를 적용하였다. 마지막으로 가중치 방법 간의 상호작용을 이끌어내기 위해 통합 가중치를 적용하였다. 이 방법 또한 가중치 방법과 동일하게 계산하였으며 〈표 5〉의 가중치를 적용하였다. 각각의 품질 측정 방법에 따라 두 가지 측정지표인 완결성과 정확성을 이용하였다.

먼저 계수방법에 따른 지표를 살펴보면, 전체 검증 건수 974개에 대해 누락된 요소를 갖고 있는 레코드 건수가 45개로 오류율 4.62%를

〈표 7〉 품질 측정 결과

측정방법 / 측정지표	계수방법	이진방법	가중치 방법			통합 가중치
			엔트로피	이용자과업	이용 비율	
완결성	95.38%	94.76%	99.75%	99.93%	99.92%	99.86%
정확성	83.16%	75.15%	98.54%	99.38%	97.83%	98.54%

보여 최종적인 완결성은 95.38%를 가져왔다. 반면 이 방법의 정확성은 오류율인 16.84%로, 83.16%를 가져왔다. 이는 하나 이상의 오류를 포함하는 레코드를 1 건으로 계수하는 방식의 영향에 따라 더 작은 값을 갖지 않은 것으로 생각된다. 즉 오류 수는 293개이지만 오류가 발생한 레코드 수는 164이다.

이진 방법은 완결성 94.76%, 정확성 75.15%로 계수방법보다 낮게 나타났다. 특히 정확성이 매우 낮은 것을 알 수 있다. 계수 방법의 경우 측정 단위가 레코드이므로 레코드에 여러 오류가 중복해서 나타나면 오류 수에 관계없이 하나로 계수되지만, 정확성의 이진방법은 오류난 모든 요소를 직접 계산하므로 시도된 다양한 방법 중에 가장 낮은 값을 보였다.

가중치 방법 중에서는 완결성은 모두 비슷하게 높은 결과를 가져왔다. 이는 누락 오류가 몇몇 요소에 한정해서 발생하기 때문인 것으로 보인다. 또한 정확성의 경우 가장 많은 오류가 발생한 키워드, 초록, 소속기관 등이 낮은 가중치를 갖는 이용자과업이 상대적으로 높은 값을 가지는 것으로 나타났다. 최종적인 통합 가중치의 경우, 완결성이 99.86%이고, 정확성이 98.54%로 특별히 한 메타데이터 요소의 오류에 영향을 받지 않으면서 균형 잡힌 측정값을 보인 것으로 나타났다.

5. 결 론

이 연구에서는 메타데이터의 품질을 측정할 때 요소별 상대적 중요도를 나타내는 가중치를 보다 체계적으로 계량화할 수 있는 여러 가지 방법론을 제시하였으며, 이를 메타데이터가 가지는 그 자체의 특성 그리고 메타데이터 이용 및 관리 측면에서 살펴보았다. 또한 KISTI 과학기술학회마을의 메타데이터를 이용하여 이러한 품질측정 방법을 적용하고 그 결과를 제시하였다. 연구에 사용된 실험 데이터는 총 974건의 학술지와 기사 관련 메타데이터로 2번의 검증을 거쳐 오류를 찾아내고 이를 품질 측정 지표인 완결성과 정확성을 이용하여 평가하였다. 실험 결과 다음과 같은 결과를 얻었다.

첫째, 메타데이터 요소의 가중치로 엔트로피는 메타데이터 자체 내에 존재하는 특성을 반영하는 방법임을 보여 주었다. 이는 기관의 성격에 따른 메타데이터의 특징이나 구축상황에 따른 메타데이터의 특성을 측정값으로 표현하여 가중치로 반영할 수 있음을 뜻한다.

둘째, 이용자 과업에 근거한 가중치 산출은 이용자가 자신의 정보요구를 해결하기 위한 과업을 달성하기 위해 어떠한 메타데이터 요소가 필요로 하는지를 이론적으로 뒷받침하고, 더 나아가 체계적으로 계량화할 수 있는 방법을 제시하였다. 언어 요소를 제외하면 본서명, 축약서명, 진행/후속 서명 순으로 가중치가 높게 나타나 서명 위주의 요소가 중요함을 보여주었다.

셋째, 이용자의 실제 이용 통계를 추출하여 메타데이터 요소별 이용비율을 가중치로 간주하는 방법은 그 기관의 이용자들이 어느 메타데이터 요소가 이용되느냐에 따라 이용률 값이 달라지므로, 이용자 지향의 서비스를 평가할 수 있는 기초를 제시하였다는 측면에서 중요한 의미를 갖는다.

넷째, 통합 가중치의 경우 기사의 본서명, 저자명, 키워드, 원문정보 순으로 높은 가중치를

보였을 뿐만 아니라, 메타데이터 특정 요소의 오류에 영향을 받지 않는 균형 잡힌 측정값을 제시하였다. 이러한 결과는 통합 가중치가 메타데이터의 세 가지 측면을 모두 반영하여 품질 측정의 계량화에 적합한 것으로 보인다.

이 연구에서는 메타데이터의 품질 측정을 위해 다양한 계량화 방법을 제시하였다. 이연구의 제한점으로 이용 통계를 이용한 가중치 산출 방법에서 너무 한정된 요소만을 제시되는 한계를 보였다. 또한 학술지 기사 메타데이터만을 대상으로 실험을 하였으므로 다양한 종류의 메타데이터에 대해 적용해 볼 가치가 있다. 실험 대상 문헌의 규모도 좀 더 크게 할 필요가 있다. 또한 추후에 기존의 다른 품질 측정 기법을 결합하여 보다 최적화된 성능을 가져오는지 알아볼 필요가 있다.

참 고 문 헌

김태수. 2008.『목록의 이해』. 개정증보판. 서울: 한국도서관협회.

사공철, 김태수, 정영미, 최석두. 2001.『정보학사전』. 서울: 문헌정보처리연구회.

윤구호. 2001.『색인 초록』. 개정증보판. 서울: 한국도서관협회.

이응봉, 조현양, 류범종, 최재황. 2001. 과학기술분야 데이터베이스의 품질향상을 위한 품질평가 연구.『한국문헌정보학회지』, 35(2): 110-132.

이제환. 2002. 공동목록 DB의 품질평가와 품질관리: KERIS의 종합목록 DB를 중심으로.『한국문헌정보학회지』, 36(1): 61-89.

Caplan, P. 2003. *Metadata Fundamentals for all Librarians*. Chicago, IL: ALA Editions.

Caplan, P. 2004.『메타데이터의 이해』. 오동근 역. 대구: 태일사.

Delsey, T. 2002. *Functional Analysis of the MARC 21 Bibliographic and Holdings Formats*. [cited 2011.3.1]. 〈http://www.loc.gov/marc/marc-functional-analysis/functional-analysis.html〉.

Haynes, D. 2004. *Metadata for information management and retrieval*. London: Facet Publishing.

Hufford, J. 1991. "Elements of the bibliographic record used by reference staff members at three ARL academic libraries." *College and Research Libraries*, 52(1): 54-64.

IFLA Study Group on the Functional Requirement for Bibliographic Records. 2003.『서지레코드의 기능상의 요건: 최종보고서』. 김태수 역. 서울: 국립중앙도서관.

Intner, S., S. Lazinger, and J. Weihs. 2006. *Metadata and its impact on libraries*. Westport, Connecticut: Libraries Unlimited.

Ochoa, X. and E. Duval. 2006. "Quality Mertrics

for Learning Ojbejct Metadata." *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 1004-1011.

Stvilia, B. and L. Gasser. 2008. "Value-based metadata quality assessment." *Library and Information Science Research*, 30(1): 67-74.

Stvilia, B., L. Gasser, M. B. Twidale, and L. Smith. C. 2007. "A framework for information quality assessment." *Journal of American Society of Information Science and Technolgy*, 58(12): 1720-1733.

Stvilia, B., L. Gasser, M. B. Twidale, S. L. Shreeves, and T. W. Cole. 2004. "Metadata quality for federated collections." *Proceedings of ICIQ04 - 9th International Conference on Information Quality*, 111-125.