정보관리연구, vol.43, no.3 2012, pp.161-180
http://dx.doi.org/10.1633/JIM.2012.43.3.161

# 출판사 생성 이용통계 데이터의 품질 평가에 대한 연구

## Evaluation on Quality of Publisher-Generated Usage Statistics

정 영 임*
**Young-Im Jung**

차 례



초 록

본 논문에서는 최근 연구가 활성화되고 있는 전자저널 이용통계의 개념과 필요성에 대해 알아보고 COUNTER 그룹에 의해 진행 중인 이용 지수 프로젝트(Journal Usage Factor Project)의 동향을 파악하였다. 또 생성 주체별 이용통계 데이터가 가지는 장점 및 한계점을 살펴보았고, 31개 출판사의 실제 이용통계 데이터를 분석하여 다양한 오류와 문제 유형을 발견함으로써 출판사 생성 이용통계 데이터의 품질이 완전히 신뢰할만한 수준이 아님을 지적하였다.

키 워 드

출판사 생성 이용통계, 데이터 품질 평가, 저널 이용 지수

* 한국과학기술정보연구원 정보서비스센터 해외정보실
(Department of Overseas Information, Information Service Center, KISTI, acorn@kisti.re.kr)
• 논문접수일자: 2012년 3월 19일
• 최종심사(수정)일자: 2012년 4월 4일
• 게재확정일자: 2012년 4월 13일

**ABSTRACT**

In this study, concept and importance of e-journal usage statistics has been examined and Journal Usage Factor project performed by Project COUNTER group has been investigated. Advantages and limits of usage statistics generated by library, link resolver and publisher have been clarified. By analyzing various errors and problems included in publisher-generated usage statistics, we conclude that the current usage statistics is not completely credible yet.

**KEYWORDS**

Publisher-Generated Usage Statistics, Data Quality Evaluation, Journal Usage Factor

## 1. 서 론

학술정보자원의 형태가 인쇄에서 디지털로 전환되면서 연구자들에게 전자정보의 이용은 큰 비중을 차지한다. 전자정보를 제공하고 지원하는 비용이 도서관 전체 예산의 대부분을 차지함에 따라(심원식 2005), 도서관 담당자가 자관의 전자정보 이용에 대해 파악하는 것은 많은 예산을 들여 구입한 전자정보의 활용을 촉진하고 나아가 저널의 구독 및 취소 등의 수서 정책 수립과 합리적인 예산 운용을 위해 필수적이다. 특히 급변하는 디지털 환경에서 정보 검색, 색인, 목록 등 전통적인 사서의 역할이 검색 엔진, 출판사 시스템을 통한 자동화로 축소되고 있는 가운데, 자관의 정보자원 수요 및 이용 현황을 파악하여 한정된 예산 안에서 가장 효율적이고 합리적으로 자관의 장서를 구성하는 일은 시스템이나 그 누구도 대신 해줄 수 없는 사서 본연의 임무라 하겠다.

과거 인쇄물의 이용을 직접 파악하기 어려워 도서관 방문자 수, 대출건수, 상호대차 건수 등 부분적이고 간접적인 통계를 이용해왔으나, 전자정보의 이용은 도서관에든 출판사 시스템에든 기록이 남기 때문에 자관에서 구독 중인 정보자원에 대한 이용통계를 확보할 수 있다.

이용통계 산출 항목에 대한 표준화 연구도 많이 진행되어 도서관에서는 보다 일관적인 이용통계 보고서를 제공받고 있다. 또 방대한 양의 이용통계를 사람이 일일이 수집하지 않고 출판사 서버에서 이용통계를 수집하는 시스템으로 전송할 수 있는 이용통계 데이터 전송 표준 규약인 SUSHI가 구현되었으며, 최근 영국의 JISC, 국내 KISTI 등 전자학술정보의 국가적 유통을 담당하고 있는 기관에서 다수 출판사의 이용통계 자동 수집 시스템을 구축하여 통합 이용통계 서비스를 제공하고 있다(MacIntyre 2011; 정영임 외 2012). 사서가 전자정보 이용통계를 활용할 수 있는 기반이

마련된 셈이다.

한편 전자정보 이용통계 활용의 기반이 마련되면서, 도서관 외에도 다양한 주체들이 이용통계 데이터 활용에 관심을 가지기 시작했다. 출판사, 정책 입안자, 개인 연구자 및 연구자 그룹, 그리고 저널의 가치 평가에 저널의 이용을 반영하고 이를 표준화하려는 표준 그룹이다.[1] 그러나 이용통계를 다양한 의사 결정 및 측정 도구의 기반 자료로 이용하기 위해서는 데이터를 자세히 분석하여 품질을 평가하는 작업이 선행되어야 한다.

따라서 본 논문에서는 이용에 기반한 다양한 의사 결정을 지원하고, 보다 정확하고 신뢰할만한 저널 가치 측정 도구의 개발을 지원하기 위해, 선행 연구 측면에서 현재 보편적으로 활용되고 있는 출판사 생성 이용통계의 품질을 검증해보고자 한다. 본 논문의 구성은 다음과 같다. 먼저 제2절에서는 전자정보 이용통계의 개념과 중요성 및 저널 평가에 인용기반 지수를 보완하여 이용통계를 기반으로 하는 측정도구를 개발하는 이용 지수 프로젝트(Usage Factor Project)[2]에 대해서 소개한다. 제3절에서는 이용통계를 생성하는 주체 및 방법에 따라 분류하고 각각에 대해 세부적으로 기술한다. 4절에서는 현재 가장 보편적으로 활용되고 있는 출판사 생성 이용통계 데이터의 다양한 오류와 품질 문제에 대해서 분석한다. 마지막으로 제5절에서는 본고의 연구 결과를 요약하고 출판사 생성 이용통계의 품질을 향상시키기 위한 다양한 주체들의 역할에 대해 제언하며 결론을 맺는다.

## 2. 이론적 배경

2장에서는 전자정보 이용통계의 개념 및 필요성에 대한 선행연구를 알아보고, 이용량에 기반하여 저널 가치를 측정하는 새로운 도구의 개발을 목표로 하는 이용 지수 프로젝트에 대해 기술한다.

### 2.1 전자정보 이용통계의 필요성

전자정보 이용통계는 도서관 이용자들이 외부 정보 공급사가 제공하는 전자 자료나 서비스를 사용한 양과 범위를 수치화한 데이터이다. 자주 사용되는 이용통계의 예로는 다음과 같은 것이 있다(심원식 2005).

- 특정 데이터베이스 안에서의 세션 수(count of sessions in a specific database)
- 실행된 검색 수(count of searches performed)
- 전문 다운로드 수(number of times full-text documents downloaded)

1) Shepherd(2012)는 출판사로부터 COUNTER 표준을 따르는 신뢰할만한 이용통계가 제공되고 있으므로 이용 기반 저널 평가가 가능하다고 천명했다.

2) COUNTER 그룹에 의해 진행 중인 Usage Factor Project가 현재까지 국내 선행연구에서 적절하게 명명된 바가 없어 본 고에서는 '이용 지수 프로젝트'라고 번역하였다.

- 세션 당 평균 사용 시간(times per session)

국내외 많은 선행 연구에서 전자저널의 이용통계의 필요성에 대해 역설해 왔다. 이용통계는 첫째, 개별 도서관과 도서관 컨소시엄의 자원 정책을 결정하는 근거 자료로 활용되고, 둘째, 이용통계가 이용자와 이용자의 정보추구 행태에 대한 정보를 제공하는 등 도서관에서 전자화된 저널의 이용통계는 여러 가지 측면에서 필수적이다. 또, 전자저널의 가치에 대한 측정 지수로서의 역할을 한다는 것과 나아가 정확한 장서 평가를 위한 방법론으로 이용중심의 평가 방법의 기반이 된다는 것이다.

2000년에서 2001년 사이 미국 연구도서관협회(ARL)의 주도로 수행한 E-metrics 프로젝트는 전자화된 정보자원과 관련된 다양한 측면을 계량화하여 평가하려는 시도였다. 도서관의 전자 네트워크 및 자원의 측정 가능한 요소로는 '기술적 인프라', '정보 콘텐츠', '정보서비스', '지원' 및 '운영'의 5개 분야가 있고, 각 구성 요소마다 이를 평가할 수 있는 측정 지수가 제안되었다. 그 중 전자정보의 이용과 관련된 통계지수는 (D2) 디지털 장서의 이용(Use of digital collection), (U1) 전자 참조서비스 시행 건수(# e-ref transactions), (U2) 전자 데이터베이스로의 로긴(세션) 건수(# of logins (sessions) to e-databases), (U3) 전자 데이터베이스 내에서의 질의 수(# of queries in e-databases), (U4) 전자 데이터베이스에 요청된 항목(Items requested in e-databases), (U5) 방문 건수(# virtual visits), (P1) 전체 참조서비스 대비 전자참조서비스 비율(% of e-ref of total ref) 등이다. Hahn & Faulkner (2002)는 현재 구성된 장서의 가치를 콘텐츠와 이용자의 수요에 기초하여 평가하는 방안으로 아래 세 가지 지수를 제안하였다.

- 접근당 평균 비용(average cost per access) = 구독비용(subscription price)/논문 접근 건수(number of articles accessed)
- 논문당 평균 비용(average cost per article) = 구독비용(subscription price)/온라인 논문 수(number of articles online)
- 수록 논문수를 감안한 이용율(content-adjusted usage) = 원문 접근 건수(number of full text accesses)/온라인 논문 수(number of articles online)

Luther(2001)는 무엇(콘텐츠)이 누구(이용자)에 의해 어떻게(방법) 언제(이용시간) 이용되는지를 측정하여 이용자와 이용자의 정보추구 행태에 대한 정보를 제공해야 한다고 하였다. 미국 멜런 재단의 후원을 받아 스탠포드 대학 도서관에서 데이터마이닝 기법을 이용해 HighWire의 트랜잭션 로그를 분석해 이용자 정보추구 행태와 추이에 관한 분석을 시도하였다.

McDonald(2006)는 인용과 이용 통계에 기

반해 인쇄 저널 이용, 온라인 저널 이용, 온라인 저널 검색 툴과 저널 인용 간의 상관관계를 분석하였고, Morrison(2005)은 전자정보의 이용통계가 경제적 요소로써 학술 커뮤니케이션에 미치는 영향을 살펴보기도 하였다.

개별 도서관에서는 보다 직접적이고 다양한 이유로 이용통계를 필요로 한다.

심원식(2005)은 전자정보를 제공하고 지원하는 비용이 도서관 전체 예산의 대부분을 차지함에 따라 도서관 운영의 분석, 보고 및 의사결정을 지원하기 위해 전자정보 이용통계의 역할이 부각되었음을 지적하였다.

컨소시엄 운영기관에서는 컨소시엄 기반 빅딜 모델에 대한 반발과 그 한계를 극복하는 방안으로 이용통계 데이터를 이용한다(김성진 외 2008). 반면 이전에 구독되지 않았던 학술지 중 오하이오-링크의 빅딜을 통해 구독된 학술지 중 약 15%에서 35%는 실제로 이용된 것으로 나타나, 이용통계를 이용해 빅딜의 한계로 지적되던 이용되지 않는 자원의 번들 구매에 대한 비판에 반론을 제기하기도 하였다(Gatten & Sanville 2004). 국내 연구에서도 전자저널 구독이 효율적이고 합리적으로 진행되기 위해 이용통계 데이터의 중요성을 인지하고 이용통계 공급자 및 활용자 간의 표준화 노력이 있어야 한다고 제안했다. 또한 이용통계 데이터가 전자정보 장서관리 정책에 반영되고 가격책정의 기반으로 활용될 수 있도록 이용통계 데이터 관련 연구를 지원하고 수행해야 한다고 지적한다(김혜선 2004).

### 2.2 이용 지수 프로젝트 (Usage Factor Project)

Shepherd(2012)는 ISI의 저널 영향력 지수(Impact Factor, IF)가 미국에서 출판되는 저널 중심이어서 비영어권 저널을 포함하지 못하고, 생명과학, 의과 분야에만 최적화되어 있으며 자기-인용으로 조작될 가능성이 있고, 저널에 대한 평균치만 제공할 뿐 개별 논문에 대한 분석이 불가능하고, 고품질의 군소 저널을 과소평가하는 등 많은 한계가 있다고 지적하였다. 그리고 IF의 한계를 보완할 수 있는 측정도구로 저널 이용 지수(Journal Usage Factor, JUF)를 제안하면서 JUF의 장점에 대해 다음과 같이 기술하였다.

- JUF는 대규모의 저널에 적용할 수 있다.
- 온라인 저널이 유통되는 학계 모든 분야에 적용할 수 있다.
- 전문직 지향 저널의 영향력이 이용에 더 잘 반영된다.
- 논문 출판 즉시 이용이 기록되고 보고된다.
- JUF의 활용은 기존의 IF로의 지나친 집중을 줄여줄 수 있다.
- 저자들은 이용에 기반한 저널 가치 측정을 환영할 것이다.

이에 UK Serials Group(UKSG)의 후원을 받아 2007년에서 2011년까지 두 단계로 프로젝트가 진행되었다. 1단계에서는 JUF의 타당성과 수용 가능성에 대한 시황 조사를 진행하

였고, 2단계에서는 JUF를 산출하기 위한 공식의 정립과 해당 공식을 실제 COUNTER 표준을 따른 데이터에 대입하여 검증하는 작업을 수행하였다. 2단계 말에 이용 지수 프로젝트 운영의 책임이 COUNTER로 이관되었고, 2011년 10월부터 COUNTER의 Peter Shepherd가 3단계 프로젝트의 총괄을 맡아 이를 진행하고 있다.

1단계 프로젝트에서 저자, 편집자, 사서, 출판사 담당자로부터 얻어진 의견들은 JUF가 전자저널의 가치와 품질의 측정을 위한 의미있는 도구인지 판단하는데 도움이 되었을 뿐 아니라, 실제 적용 방안에도 많은 식견을 제공했다. 또한, 향후 연구에서 다루어져야 할 주제들이 제안되었다. 주된 내용은 아래와 같다.[3)]

- 영향력 지수(IF): 저자, 사서, 대부분의 출판사들은 IF의 대안으로 저널의 가치를 평가하는 보다 포괄적이고, 계량적이며 비교 가능한 측정 도구의 개발을 강력히 열망하고 있다. 아직까지는 대안이 존재하지 않지만 미래의 그러한 측정 도구로 이용 데이터가 사용이 될 것이다. 설문에 답한 저자들 중 70%가 새로운 이용 기반 측정도구를 환영한다.
- JUF에 의한 저널 순위: 대부분의 출판사들은 JUF에 의해 자사 저널의 순위를 매겨볼 의사는 있으나, JUF의 산출이 제대로 나와야 하며 공정해야 한다고 주장한다. 사서들은 새로운 저널의 구독을 결정하는 두 번째로 중요한 요소로 JUF를 활용할 수 있을 것이라 답하였다.
- JUF 데이터 총괄 기구: 현재 도서관과 출판사 양측의 신임을 받고, JUF 데이터를 집계/검증할 수 있는 역량을 가진 기구는 없다. 따라서 도서관과 출판사 양측의 협력관계가 필요하며, 개별적으로도 역할을 담당해야 할 것이다.
- 대부분의 출판사들은 원칙적으로는 자사의 저널에 대해 JUF를 계산하고 이를 공시하는데 찬성한다고 밝혔고, 어떤 출판사들은 이를 완전히 지지하지는 않지만 UF가 정의되고 구현된다면 어떤 방법으로든 시장에 수용될만한 방법으로 이에 참여할 수 있다고 언급했다.

결론적으로, 1단계 프로젝트를 통해 JUF에 대한 지지를 확인했고, 과연 JUF가 신뢰도를 가질 수 있는가에 대한 조사가 필요하다는 점을 확인했다. 이에 2단계에서는 과연 JUF가 신뢰할만하고, 구현 가능하며 비용면에서 효과적인 도구인가를 검증하기 위해 설계되었으며, UKSG 등의 후원으로 Frontline GMS와 John Cox Associate에서 프로젝트를 수행하였다. 7개 출판사로부터 5개 주제 분야(공학, 인문학, 의학, 자연과학, 사회과학)의 2006년에서 2009년에 발간된 326종 저널의 이용통계를 테스트

3) 〈http://www.projectcounter.org/usage_factor.html〉.

데이터로 확보하였다. 2단계 프로젝트에서는 JUF의 4개 변수의 영향력이 분석되었는데 이를 요약하면 다음과 같다.

- 콘텐츠 유형: 모든 유형의 콘텐츠 대비 논문의 JUF를 비교분석한 결과, 모든 유형의 콘텐츠의 JUF 산출이 논문만의 JUF를 산출한 것보다 모든 주제 분야에 더 일관적인 측정 결과를 보여주었다.
- 논문 버전: 균형적이고 공정한 평가를 위해서는 논문의 모든 버전에 대해 JUF가 산출되어야 한다.
- 출판 기간: 출판 기간은 1년 단위보다 2년 단위로 JUF를 산출해야 보다 일관적이며 신뢰할만한 결과를 얻을 수 있다.
- 이용 기간: 온라인 출판월 이후 1-24개월 단위로 이용 기간을 잡아 JUF를 산출하는 것을 권고한다.

또 CIBER는 테스트 데이터에 대한 통계적 분석과 JUF의 중요한 속성에 대해 실험을 진행하여 이에 대한 분석 결과로 JUF의 적용과 관련된 12개 권고안을 제시하였다.

3단계 프로젝트는 아래 5가지 목표를 이루기 위해 설계되었고 2012년 3월에 완료될 예정이다(COUNTER 2012).

- Code of Practice for the Journal Usage Factor 초안 마련
- JUF 권고 산출 방안에 대한 추가 실험
- 저널 분류를 위한 적절하고 탄력적인 주제 분류 연구
- JUF의 지속가능한 구현을 지원하기 위한 기반 연구
- UF 개념의 타 분야 출판물에 적용 타당성 연구

본 절에서 살펴본 바와 같이 21세기 초반까지 전 세계 저널과 학자의 역량 평가에 지대한 영향력을 끼쳐 온 IF를 보완하는 새로운 측정 도구로 이용에 기반한 지수인 JUF에 대한 연구가 야심차게 진행되고 있음을 알 수 있다. 이용 지수 프로젝트에서 밝힌 바대로 결함이 있는 IF가 가진 막대한 영향력에 대응하기 위해 이용에 기반한 저널 평가는 연구자, 사서들이 환영하는 바이지만, 우선 이용통계 데이터의 품질과 신뢰도가 보장되어야 한다. 따라서 본 연구에서는 생성 주체별 이용통계 데이터의 장단점에 대해 논의하고, 현재 가장 보편적으로 활용되고 있는 출판사 생성 이용통계 데이터의 품질을 분석한다.

## 3. 생성 주체에 따른 이용통계 분류

### 3.1 도서관 웹사이트 로그 분석을 통한 이용통계

도서관에서는 도서관 웹사이트에서 전자저널을 포함한 데이터베이스로의 링크를 제공한다. 따라서 도서관 이용자의 이용에 관한 데이

터를 얻을 수 있는 가장 기본적인 방법은 웹 서버의 로그를 수집해 분석하는 것이다. 이용자가 도서관 웹사이트에 접근을 할 때마다 해당 트랜잭션이 서버의 로그 파일에 기록된다. 로그 파일에는 '파일을 요청하는 컴퓨터의 주소', '요청한 날짜와 시간', '요청된 파일의 URL', '요청에 사용된 프로토콜', '요청된 파일의 크기', '참조 URL', '요청된 컴퓨터의 브라우저와 운영 시스템' 등이 기록된다. 이러한 로그 파일을 특정 기간 단위로 수집해 웹 로그 분석 프로그램을 이용하여 분석할 수 있다. 현재 다양한 웹 로그 분석 프로그램이 있는데, 대표적인 웹 로그 분석 프로그램으로 'Deep Log Analyzer', 'Google Analytics', 'Webalizer' 등의 프리웨어도 있고, 'WebTrends'와 같은 상용 로그분석 프로그램도 있다. 이와 같은 방법을 통해 수집되고 분석된 도서관 웹사이트 서버 통계로부터 '조회된 페이지의 총 수', '서비스를 받은 개별 컴퓨터의 총 수', '조회/다운로드된 도서관 전자자원의 총 수'를 산출하여 도서관 웹사이트에서 제공하는 전자자원에 대한 이용통계를 얻을 수 있다. 이 외에도 도서관 이용자 수, 웹사이트를 통한 도서관 업무 내용, 도서관 전자자원 관리, 참조 서비스, 상호대차 서비스, 이용자-사서 간 질의응답 등 다양한 도서관 서비스에 대한 분석이 가능하여, 도서관에서 보유한 자원 및 제공하는 서비스의 이용에 대한 전반적인 분석이 가능하다. 이에 워싱턴대학교의 환경보건학부 도서관에서 1995년에 이용 데이터를 수집하기 시작하였고, 통계 분석 결과 다음과 같은 결론을 내렸다(Stavin & Owen(1997), Welch (2005)에서 재인용).

> 이 프로젝트에서 얻어진 이러한 통계는 우리에게 워싱턴대학교 내・외부의 이용자들이 찾고자 하는 정보의 구체적인 유형에 대한 직관을 제공해 주었다. 이는 가장 많이 이용된 분야로 우리의 관심을 전환시켰다. (These statistics that we gathered during this project have provided us with an insight into the specific types of information that users inside and outside the University of Washington are looking for. This will enable us to shift our concentration to areas that get the most usage.)

잭슨빌대학교, 노스텍사스대학교, 로드아일랜드 대학교 등의 대학도서관도 도서관 웹사이트 로그 분석을 통해 생성한 이용에 관한 통계를 도서관이나 대학 웹사이트에 게시하기도 하였다(Welch 2005).

University of North Carolina(UNC) at Charlotte에서는 2003년에 참조전문 사서, 기술서비스 사서, 특수 장서 사서, 도서관 웹 마스터 등으로 구성된 웹사이트 통계업무 그룹을 구성하여 도서관 전자 자원 및 서비스에 관련된 전반적인 통계를 분석하였다(Welch 2005). UNC-Chapel Hill의 학술도서관에서는 이용자가 원하는 전자자원 링크를 클릭할 때마다 이용자(IP 단위)와 해당 자원에 대한 로그 엔트

리[4]가 생성되고, 이러한 클릭의 합산을 도서관 생성 세션으로 산출함으로써 각 자원의 타이틀별로 세션의 통계를 생성하였다(Kemp 2008).

이와 같이 기관의 필요에 따라 서버 로그 분석 툴, 웹사이트 트래픽 분석 툴 등을 이용하여 만들어지는 도서관 생성 이용통계는 자관의 목적에 맞게 일관적인 데이터를 생성할 수도 있다는 장점이 있다. 그러나 개별 기관에서 통계 산출을 위한 시스템을 구축하는데 시간과 전문성이 요구되어(심원식 2005) 정보자원을 담당하는 사서가 감당하기가 어렵고 기관 예산을 들여 수행하기가 어렵다. 더욱이 도서관에서는 자관의 서버를 통해 수집되는 로그는 분석할 수 있지만 이용자가 도서관 웹사이트를 통하지 않고 Google Scholar나 NDSL과 같은 검색 서비스를 통해 출판사 웹사이트에 바로 접속하여 전자정보를 이용하는 경우에 대한 통계는 포함할 수 없다는 한계가 있다.

### 3.2 링크 리졸버를 통한 이용통계

링크 리졸버는 다수의 출판사에서 제공되는 전자학술정보의 원문을 하나의 플랫폼에서 손쉽게 이용할 수 있어 전자정보의 이용을 촉진시키는 효과가 있다. 링크 리졸버에서 제공하는 통계는 이용자들이 전자자원에 접근하는 경로에 대한 많은 정보를 제공한다. 특히, 이용자가 얼마나 특정 저널에 대한 원문을 찾고자 시도하는지, 해당 자원에 대해 얼마나 클릭하는지를 알 수 있다. 이를 통해 이용자가 어떤 자원을 검색하는지, 그러한 자원으로의 접근이 얼마나 성공적으로 이루어지는지를 도서관에서 파악할 수 있어, 검색한 자원 중 원문으로의 접근이 성공적으로 이루어지지 못한 자원을 모아 구독 후보 저널리스트를 생성할 수 있다. 또 어떠한 경로가 이용자로 하여금 도서관의 전자원문으로 가장 많이 연결하는지를 파악할 수 있다. 예를 들어 링크 리졸버에서 생성된 이용통계를 분석함으로써 한 도서관 내 이용자들이 원문에 접근하는 경로로 Google Scholar가 타 데이터베이스보다 압도적으로 많이 사용되었음을 밝힐 수 있었다(Orcutt 2009). 또한 링크 리졸버를 통해 요청되는 상호대차서비스를 통한 이용량도 집계된다. 그러나 도서관 생성 이용통계와 마찬가지로 링크 리졸버를 통하지 않은 전자정보의 이용량에 대해서는 파악할 수 없다는 한계가 있다. 이에 도서관에서 활용되고 있는 이용통계는 대부분 각 정보자원을 생산하는 출판사에서 생성하여 제공하고 있다.

### 3.3 출판사 생성 이용통계

도서관 웹사이트, 링크 리졸버, Google Scholar, NDSL 등 다양한 경로에서 논문에 대한 링

4) 도서관 내 존재하는 자원에 대한 로그 엔트리는 자체 생성할 수 있으나 출판사 서버에 존재하는 자원의 이용에 대한 산출은 도서관 웹사이트에서 출판사 사이트로 넘어가는 중간 과정에 웹 페이지를 삽입하여 이용자가 전자자원으로 연결되는 링크를 눌렀을 때 중개 페이지를 거쳐서 해당 출판사 사이트로 이동하게끔 하고, 중개 페이지를 계수기로 삼아 전자정보의 이용량을 추적한다(심원식 2005).

크를 통해 원문 페이지로 유입되는 이용은 모두 출판사 서버에 기록된다. 따라서 도서관이나 링크 리졸버에서 산출되는 이용통계보다 출판사 생성 이용통계가 포괄적이다. 거의 모든 출판사에서 이용통계를 생성하여 제공하고 있는데, 이는 고객인 도서관으로부터 출판사가 판매하고 있는 전자정보의 이용통계를 제공하라는 요구에 부응하고, 출판사 자사의 전자정보 이용 현황을 모니터링하기 위해서이다. 이에 현재 출판사, 도서관, 컨소시엄에서 활용하고 있는 전자저널 이용통계는 대부분 출판사에서 생성하여 제공하고 있다. 출판사 생성 이용통계의 산출 항목이 출판사별로 달라 일관성이 없다는 사실은 많은 선행연구에서 지적되었다. 출판사 데이터에서 '이용'이라 함은 세션인지, 검색인지 아니면 원문 접근을 지칭하는 것인지 모호하다(Orcutt 2009). UNC에서는 출판사 생성 통계와 도서관 생성 이용통계 데이터를 비교한 연구를 통해 실제로 각 출판사에서 생성한 이용통계 데이터가 도서관 생성 데이터보다 비일관적임을 보여주었다. 예를 들어, 어느 벤더는 단일 쿼리로 여러 DB를 검색하는 수행을 한 건으로 본 반면, 다른 벤더에서는 단일 쿼리나 복수의 DB를 검색하였으므로 복수 검색으로 간주했다(Duy & Vaughan 2003). 이에 도서관과 출판사 양측 모두 전자정보의 이용에 대해 일관적인 산출과 보고에 대한 중요성을 인식하였고, 이러한 인식 공유의 결과로 2002년 COUNTER 표준이 등장하였다(COUNTER 2008). COUNTER 표준의 등장 이후, 출판사에서 자체 포맷의 데이터 대신 표준 포맷을 따른 이용통계 데이터를 생성함으로써 데이터의 일관성 측면이 많이 향상되었다(Orcutt 2009). 앞에서 설명한 바와 같이 출판사 생성 이용통계가 포괄적이고, COUNTER 표준 포맷을 적용하여 데이터 일관성도 갖추었다는 전제 하에 출판사 이용통계를 기반으로 저널의 가치를 측정하고 평가하는 이용 지수 프로젝트도 진행되고 있다(2.2절 참조).

그러나 출판사에서 제공하는 이용통계 데이터는 여전히 다양한 오류를 포함하고 있으며, 여러 요인에 의해 왜곡된 수치를 보여주고 있다. 4절에서는 출판사에서 생성한 전자저널 이용통계에서 발견되는 다양한 오류 유형을 기술함으로써 출판사에서 제공하는 이용통계의 품질이 떨어지고, 데이터의 신뢰도에 문제가 있음을 밝힌다.

## 4. 출판사 생성 이용통계 품질 평가

정영임 외(2011)는 KESLI 컨소시엄을 통해 유통되는 국내 STM 분야 전자저널의 이용 현황을 국가적 차원에서 파악하기 위해 전자저널 이용통계 자동수집 시스템을 구현하였다. 해당 시스템의 스크래핑 엔진이 각 출판사 이용통계 데이터를 제대로 스크래핑하여 구축하였는지 확인하기 위해, 구축된 이용통계 보고서를 임의추출하여 출판사 웹사이트에서 제공

되는 이용통계 데이터와 비교하였다. 총 11개 출판사 데이터를 검사하였는데, 해당 검사에서 출판사에서 제공하는 이용통계 데이터 자체의 오류를 발견하였다. 이에 본 절에서는 31개 출판사[5]의 이용통계 데이터를 분석하고 데이터에서 발견되는 다양한 오류를 분석함으로써 현재 출판사 생성 이용통계의 품질을 평가하였다.

## 4.1 단순 수치 오류

AR과 APS에서 제공하는 이용통계 데이터 중 〈그림 1〉과 같이 1월부터 12월까지 모두 이용량이 '0'건이나 연간 합산한 이용량이 '0'이 아닌 오류 데이터가 발견된다. 이러한 오류는 월별 이용량이 이용통계 보고서에 결여되어 나타나는 오류이다.

AR에서 공급되는 저널 48종 중 40종의 2010년도 이용통계 데이터에 이러한 오류를 보이고 있으며, APS의 41종 저널 중 'Physical Review A/B/E' 및 'Reviews of Modern Physics' 등 6종의 2007년부터 2010년 이용통계 데이터에 이러한 오류가 나타난다. 반대로 〈그림 2〉와 같이 ProQuest's의 1,354종 저널의 2010년, 2011

| 출판사명 | YEAR | 저널명 | 플랫폼 | PISSN | OISSN | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | YTD_HTML | YTD_PDF | YTD_FULL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Annual Rev | 2010 | Annual Review of Medicine | Annual R | 0066-4219 | 1545-326X | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 2 |
| Annual Rev | 2010 | Annual Review of Microbiology | Annual R | 0066-4227 | 1545-3251 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 1 | 11 |
| Annual Rev | 2010 | Annual Review of Pathology: Mechanisms o | Annual R | 1553-4006 | 1553-4014 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| APS | 2007 | Physical Review Online Archive (PROLA) | APS | n/a | n/a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 69 | 10 | 34 |
| APS | 2007 | Reviews of Modern Physics | APS | 0034-6861 | 1539-0756 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| APS | 2008 | Physical Review B | APS | 1098-0121 | 1550-235X | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 5 | 11 |
| APS | 2008 | Physical Review Letters | APS | 0031-9007 | 1079-7114 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 12 | 5 |
| APS | 2008 | Physical Review Online Archive (PROLA) | APS | n/a | n/a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 31 | 6 | 33 |
| APS | 2009 | Physical Review B | APS | 1098-0121 | 1550-235X | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 10 |
| APS | 2009 | Physical Review Letters | APS | 0031-9007 | 1079-7114 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 3 |
| APS | 2009 | Physical Review Online Archive (PROLA) | APS | n/a | n/a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 21 | 30 | 28 |
| APS | 2010 | Physical Review B | APS | 1098-0121 | 1550-235X | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 8 | 3 |

〈그림 1〉 월별 이용 수치 결여 오류(2011년 9월 기준)

| 출판사명 | YEAR | 저널명 | 플랫폼 | PISSN | OISSN | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | YTD_HTML | YTD_PDF | YTD_FULL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ProQuest | 2010 | Telecomworldwire | ProQuest | 1363-9000 | NA | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 6 | 0 | 0 |
| ProQuest | 2010 | Telephone IP News | ProQuest | NA | NA | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| ProQuest | 2010 | Telephony | ProQuest | 0040-2656 | NA | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ProQuest | 2011 | Annual Review of Sociology | ProQuest | 0360-0572 | NA | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ProQuest | 2011 | Applied Biochemistry and Biotechnology | ProQuest | 0273-2289 | NA | 2 | 6 | 1 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ProQuest | 2011 | Interfaces | ProQuest | 0092-2102 | NA | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

〈그림 2〉 월별 이용 대비 연간 이용 합산 수치 오류(2011년 9월 기준)

5) 31개 출판사는 다음과 같다. American Association for the Advancement of Science(AAAS), American Chemical Society(ACS), American Institute of Physics(AIP), American Physical Society(APS), American Society of Civil Engineers(ASCE), American Society of Mechanical Engineers(ASME), Annual Review(AR), Berkely, BioOne, Brill Academic Publishers(Brill), British Medical Journal(BMJ), Cambridge University Press(CUP), Elsevier, IEEE, Institute of Physics Publishing Ltd(IOP), International Society for Optics and Photonics(SPIE), John Wiley & Sons, Inc(Wiley), Jstor, Karger, Mary Ann Libert(MAL), National Academic of Science(NAS), Nature Publishing Group(NPG), Optical Society of America(OSA), Oxford University Press(OUP), Pion Ltd(Pion), Project Muse, Royal Society of Chemistry(RSC), Sage Publications(Sage), Springer, Thieme Publishing Group(Thieme), Walter de Gruyter이다.

년 이용통계 데이터에 월간 이용량이 '0' 이상이나 연간 합산량이 '0'으로 산출된 오류가 발견되기도 한다.

## 4.2 이용통계 중복 산출

ASME, CUP, Springer, Wiley에서 제공하는 이용통계 데이터에 한 개 기관의 한 개 저널의 이용에 대한 통계를 중복 산출하였고, 이러한 중복 산출량이 적지 않다. ASME는 'Journal of Mechanical Design'을 포함한 15종의 저널, CUP는 'Breast Cancer Online', 'The Cognitive Behaviour Therapist'을 포함한 8종의 저널, Springer는 'International Journal of Angiology', 'Protection of Metals and Physical Chemistry of Surfaces'를 포함한 10종의 저널, Wiley는 'Design Management Journal', 'Dialysis & Transplantation'을 포함한 9종 저널의 이용통계를 중복하여 산출하였다. 한 개 저널의 중복 산출된 이용통계는 개별 저널의 수치만 볼 때는 문제가 되지 않는다. 그렇지만 중복 산출된 이용통계가 있는지 모르고 해당 저널을 공급하는 출판사 전체 저널의 합산 이용량을 구해 타출판사 전체 저널의 합산 이용량과 비교할 때는 문제가 될 수 있다.

## 4.3 동일 저널에 대한 다른 이용 수치

4.2절에 기술한 것처럼 한 개 기관의 한 개 저널의 이용통계가 2개 이상 산출되는 것 중 서로 다른 수치를 가지는 오류도 종종 발견된다. 〈그림 3〉을 보면 ASME 출판사의 'Journal of Heat Transfer' 저널을 A대학교에서 2005년과 2008년에 이용한 수치가 두 가지로 다르게 산출되고 있다.

이러한 오류는 이용통계 중복 산출보다 더욱 심각한데, 한 개 저널에 대한 이용량 중 어느 수치가 맞는지 판단할 수가 없고, 또 출판사별 혹은 컨소시엄 패키지별 전체 저널 이용량을 합산하여 비교할 때 한 개 저널에 대한 두 이용 수치 중 임의로 하나를 선택해야 할지 아니면 두 수치를 모두 합해서 전체 이용량을 산출해야 할 지 결정하기 어렵다. 한 개 저널의 다른 이용량 산출 오류는 ASME, CUP, Springer, Wiley에서 제공하는 이용통계 데이터에서 발견된다.

| 출판사 | 기관명 | 연도 | 저널명 | 플랫폼 | PISSN | OISSN | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | YTD_HT | YTD_PD | YTD_FU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASME | A대학교 a캠퍼스 | 2007 | Journal of Heat Transfer | Scitation | 0022-1481 | 1528-8943 | 3 | 3 | 15 | 7 | 19 | 15 | 8 | 8 | 9 | 16 | 22 | 10 | 14 | 111 | 135 |
| ASME | A대학교 a캠퍼스 | 2007 | Journal of Heat Transfer | Scitation | 0022-1481 | 1528-8943 | 3 | 3 | 15 | 7 | 19 | 15 | 8 | 8 | 9 | 16 | 22 | 10 | 14 | 111 | 135 |
| ASME | A대학교 a캠퍼스 | 2005 | Journal of Heat Transfer | Scitation | 0022-1481 | 1528-8943 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ASME | A대학교 a캠퍼스 | 2005 | Journal of Heat Transfer | Scitation | 0022-1481 | 1528-8943 | 1 | 4 | 5 | 3 | 1 | 4 | 23 | 0 | 2 | 0 | 0 | 0 | 0 | 43 | 43 |
| ASME | A대학교 a캠퍼스 | 2008 | Journal of Heat Transfer | Scitation | 0022-1481 | 1528-8943 | 9 | 0 | 0 | 18 | 24 | 15 | 8 | 11 | 13 | 13 | 19 | 24 | 3 | 151 | 154 |
| ASME | A대학교 a캠퍼스 | 2008 | Journal of Heat Transfer | Scitation | 0022-1481 | 1528-8943 | 9 | 12 | 20 | 18 | 24 | 15 | 8 | 11 | 13 | 13 | 19 | 24 | 3 | 183 | 186 |
| ASME | A대학교 a캠퍼스 | 2006 | Journal of Heat Transfer | Scitation | 0022-1481 | 1528-8943 | 15 | 3 | 9 | 12 | 11 | 19 | 6 | 10 | 13 | 14 | 20 | 8 | 7 | 132 | 140 |
| ASME | A대학교 a캠퍼스 | 2006 | Journal of Heat Transfer | Scitation | 0022-1481 | 1528-8943 | 15 | 3 | 9 | 12 | 11 | 19 | 6 | 10 | 13 | 14 | 20 | 8 | 7 | 132 | 140 |

〈그림 3〉 동일 저널의 다른 이용량 산출 오류(2011년 9월 기준)

### 4.4 제공방식별 제공 데이터 차이

정영임 외(2011)가 밝힌 것처럼 이용통계를 제공하는 방식이 2가지 이상인 출판사가 적지 않다. Wiley에서는 출판사 웹사이트에서 이용통계 보고서를 html 포맷으로 조회할 수도 있고, csv 파일로 다운로드 받을 수도 있다. 그런데 Wiley 2005년 데이터 중 'Bruce R. Hopkins' Nonprofit Counsel', 'Children & Society', 'International Journal of Auditing'의 3종 저널의 이용통계가 웹사이트에서 조회할 수 있는 html 파일에는 없지만, 다운로드용 csv 파일에는 존재한다.

### 4.5 제공대상별 데이터 차이

전자정보 컨소시엄 주관기관에서는 해당 컨소시엄에 참여하는 전체 기관의 이용통계 데이터를 수집하여 컨소시엄을 통해 유통되는 전자정보의 이용 현황을 모니터링하고 있다. 이를 위해 몇몇 출판사는 컨소시엄 주관기관용 이용통계 조회 계정을 발급하여 컨소시엄 관리자가 개별 기관의 이용통계를 파악할 수 있도록 하였다. 그런데 〈그림 4〉와 같이 OUP의 2011년 이용통계 데이터를 보면 A 기관의 계정으로 확인한 이용통계 수치와 컨소시엄 관리자 계정으로 확인한 A 기관의 이용통계 수치가 다름을 알 수 있다.

4.1절에서 4.3절까지 기술한 오류는 SQL문 실행을 통해 자동으로 감지해낼 수 있는 오류이다. 1월부터 12월까지의 수치 합산과 연간 수치가 동일하지 않은 데이터를 뽑거나, 저널명, P-ISSN, O-ISSN을 키로 하여 하나의 저널이 두 번 이상 카운트되는 데이터를 뽑으면 되기 때문이다. 그런데 4.4절과 4.5절에서 밝힌 제공방식별/제공대상별 데이터 비일치 오류는 자동으로 찾아내기 힘들다. 어떤 출판사의 어떤 기관, 그리고 어느 기간의 통계 데이터가 제공방식별/제공대상별로 비일치 하는지 알 수 없기 때문에 결국 전체 데이터를 대상으로 비교해봐야 정확하게 파악할 수 있다. 그러나 다수 출판사에서 제공되는 다수 기관의 다년에 걸친 이용통계 데이터의 규모는 매우 크기 때문에[6] 이러한 오류를 찾아내기가 쉽지 않다.

4.6절 이하에서 기술하는 이용통계 데이터의 문제 유형은 오류는 아니나 출판사 간 이용통계 데이터를 비교 분석할 때 단순 수치 비교를 통해서 객관적인 결과를 얻기 힘든 유형이다.

| Institution A | YTD Total | YTD HTML | YTD PDF |
|---|---|---|---|
| | 1366 | 562 | 804 |
| Consortia admin data on Institution A | YTD Total | YTD HTML | YTD PDF |
| | 1212 | 515 | 697 |

〈그림 4〉 제공대상별 데이터 비일치(2011년 8월 기준)

6) 정영임 외(2012)에 따르면 31개 출판사로부터 자동수집하여 구축한 359개 도서관의 이용통계 데이터의 규모는 2012년 2월을 기준으로 총 7,539,083건이다.

### 4.6 출판사 웹사이트 인터페이스에 따른 이용통계 왜곡

Davis & Price(2005)가 지적한 바와 같이 전자저널 웹사이트의 인터페이스가 COUNTER 기반 이용통계 산출에 영향을 미친다. COUNTER JR1 포맷을 따르면, 저널 원문을 html로 조회한 양과 pdf로 다운로드 받은 수치를 합산하여 원문의 전체 이용량을 산출하고 있다. 위 연구에서는 NPG와 HighWire Press의 두 개 출판사에서 제공하는 'The Embo Journal'이라는 동일한 콘텐츠에 대한 코넬대학교의 PDF 대 HTML 이용 평균이 두 출판사의 인터페이스에 의해 영향을 받아 유의미한 차이를 보였음을 밝혔다. 그리고 이러한 차이는 두 출판사의 인터페이스가 시각적으로는 유사해 보이나, 이용자가 논문의 PDF 원문을 탐색해 가는 방식의 차이에서 유래한다고 지적했다. HighWire의 목차 페이지에서 PDF 버전을 찾기 위해서는 HTML 버전의 논문을 먼저 다운로드 받아야 함에 비해 NPG의 인터페이스에서는 PDF 버전으로 바로 가는 링크가 목차 페이지에서 제공된다. Elsevier 인터페이스에서도 논문을 검색한 결과 페이지에서 원하는 논문 링크를 클릭하면 해당 논문의 초록 페이지가 아니라 html 원문 페이지로 바로 연결된다. 이로 인해 해당 논문의 서지 정보만 파악하려는 이용자의 경우에도 html 원문을 조회하게 되고, 해당 논문을 저장하거나 인쇄하여 보고 싶어하는 경우 다시 PDF 파일을 다운로드 받아야 한다.[7] 따라서 HighWire Press나 Elsevier 출판사의 인터페이스에서는 한 개 논문의 한 번 이용에 대해 수치가 과장될 수 있다.

### 4.7 비정상적 이용에 따른 왜곡

인쇄자원의 소장에서 전자정보의 접근으로 정보유통 패러다임이 전환되었음에도 불구하고, 도서관에서는 예산을 들여 구독한 전자자원의 저장에 대한 열망이 높다. 오히려 인쇄자원에 비해 전자자원의 휘발성이 크기 때문에 구독한 전자정보에 대한 접근만으로 자료의 지속적인 이용이 가능할까에 대한 의구심과 불안감이 크다. 이러한 담당자의 불안감을 반영한 것처럼 국내 도서관에서 전자정보의 구독 초기에 원문을 대량으로 다운로드 받는 일이 있었다. 또 대학의 이용자가 웹 로봇을 이용하여 전자저널을 대량으로 다운로드 받아 해당 기관에 서비스가 중단되거나 출판사의 경고가 내려진 사례도 다수 있었다.[8] 〈그림 5〉와 같이 B대학

7) 실제로 KESLI 컨소시엄 내 Elsevier 저널의 이용량이 타출판사 저널의 이용량을 압도한다. 이에는 다양한 원인이 있겠으나, Elsevier 인터페이스의 영향도 적지 않을 것으로 추측된다. 이러한 추측을 과학적으로 규명하기 위해 Davis & Price(2005)가 수행한 실험 방식대로 Elsevier와 타출판사 컨소시엄에 포함된 한 개 저널을 선정하여, 한 개 기관의 PDF 대 HTML 이용 평균을 비교하고 t-test를 통해 유의미한 차이를 보이는지 분석해야할 것이다.

8) 〈http://engr.korea.ac.kr/index.html?mode=read&msg=16&page=engrboard/notice01〉, 〈http://me.yonsei.ac.kr/info/news.asp?b_gbn=R&b_idx=410&b_type=L&cur_pack=1&num=101&page=20&s_field=&s_string=〉.

| 1월 | 2월 | 3월 | 4월 | 5월 | 6월 | 7월 | 8월 | 9월 | 10월 | 11월 | 12월 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 84,759 | 341,842 | 260,235 | 94,740 | 4,204 | 4,447 | 4,526 | 5,912 | 4,727 | 5,419 | 4,286 | 3,148 | 818,245 |

〈그림 5〉 B대학교 2007년 Elsevier 저널 이용량

교의 2007년 Elsevier 저널의 이용량을 보면, 1월부터 4월까지는 정상적인 이용에 의해 산출된 수치가 아님을 상식적으로 알 수 있다.

이 기관의 Elsevier 저널의 2008년도 이용 총량은 49,012건, 2009년은 44,830건, 그리고 2010년은 46,613건으로 2007년도의 818,245건과 크게 차이가 난다. 따라서 출판사 간 이용량을 비교할 때 비정상적인 이용에 의해 크게 부풀려진 이용 수치로 그 결과가 왜곡될 수 있기 때문에 주의해야 한다.

다음 4.8절과 4.9절에서 논하는 문제의 유형은 저널 단위로 이용량을 비교하거나 저널의 서지정보 및 IF정보를 결합하여 저널 이용에 대한 심층 분석을 하고자 할 때 문제가 될 수 있다.

## 4.8 저널명 표기의 비일관성

Orcutt(2009)는 출판사별로 동일 저널에 대한 저널명 표기가 일관적이지 않다고 기술했다. 어떤 출판사는 저널명에 'The'를 넣기도 하고 다른 출판사는 관사 없이 저널명이 시작되기도 한다. 출판사별로 동일한 저널에 대한 저널명 표기 방식이 매우 다른데 이를 정리하면 다음과 같다.

- 관사의 유무: 저널명 앞에 관사 'the' 삽입 여부가 다름(〈그림 6〉 참조).
- 다국어 제목 표기: 저널명을 영어로 표기, 영어 외 언어(다국어라 칭함)로 표기, 영어와 다국어를 병기 등 출판사별, 저널별로 다름(〈그림 7〉 참조).
- 부가 정보 표기 여부: 출판사에 따라 다양한 부가 정보를 표기하기도 하여 동일한

| 저널명 | P-ISSN | O-ISSN | 출판사 |
|---|---|---|---|
| Botanical Review | 0006-8101 | 1874-9372 | JSTOR |
| The Botanical Review | 0006-8101 | 1874-9372 | Springer |

〈그림 6〉 동일 저널명의 이표기-관사 유무

| 저널명 | P-ISSN | O-ISSN | 출판사 | 비고 |
|---|---|---|---|---|
| Journal of Ornithology | 0021-8375 | 1439-0361 | Springer | 영어 제목 |
| Journal f체r Ornithologie | 0021-8375 | 1439-0361 | Wiley | 다국어 제목 |
| International Statistical Review | 0306-7734 | 1751-5823 | Wiley | 영어 제목 |
| International Statistical Review / Revue Internationale de Statistique | 0306-7734 | NA | JSTOR | 영어, 다국어 병행 |

〈그림 7〉 동일 저널명의 이표기-언어

| 저널명 | P-ISSN | O-ISSN | 출판사 | 비고 |
|---|---|---|---|---|
| Erkenntnis | 0165-0106 | 1572-8420 | Springer | - |
| Erkenntnis (1975-) | 0165-0106 | NA | JSTOR | 연도 표기 추가 |
| Science | 0036-8075 | 1095-9203 | JSTOR | - |
| Science (including archive) | 0036-8075 | 1095-9203 | AAAS | archive 포함 여부 추가 |
| Wiener Medizinische Wochenschrift | 0043-5341 | 1563-258X | Wiley | - |
| WMW Wiener Medizinische Wochenschrift | 0043-5341 | 1563-258X | Springer | 약어 WMW 추가 |
| Somnologie | 1432-9123 | 1439-054X | Wiley | - |
| Somnologie - Schlafforschung und Schlafmedizin | 1432-9123 | 1439-054X | Springer | 하위 제목 추가 |
| Proceedings of the Aristotelian Society | 0066-7374 | NA | JSTOR | - |
| Proceedings of the Aristotelian Society (Hardback) | 0066-7374 | 1467-9264 | Wiley | 커버 형태 추가 |

〈그림 8〉 동일 저널명의 이표기–부가 정보 표기

| 저널명 | P-ISSN | O-ISSN | 출판사 | 비고 |
|---|---|---|---|---|
| Population & Environment | 0199-0039 | 1573-7810 | Springer | '&' 표기 |
| Population and Environment | 0199-0039 | NA | JSTOR | 'and'로 표기 |
| In Vitro Cellular & Developmental Biology - Plant | 1054-5476 | 1475-2689 | Springer | '-' 표기 |
| In Vitro Cellular & Developmental Biology. Plant | 1054-5476 | NA | JSTOR | '.' 표기 |
| Recherche Transports Sécurité | 0761-8980 | | Springer | '-' 없음 |
| Recherche - Transports - Sécurité | 0761-8980 | NA | Cell Press | '-' 있음 |
| Archivos de Bronconeumolog챨a (English Edition) | 1579-2129 | NA | Cell Press | ';' 없음 |
| Archivos de Bronconeumolog챨;a (English Edition) | 1579-2129 | NA | Elsevier | ';' 있음 |

〈그림 9〉 동일 저널명의 이표기–특수 문자 이형 표기

저널에 대한 저널명 표기가 달라짐(〈그림 8〉 참조).

- 특수 문자의 이형 표기: 저널명에 들어가는 'and'를 특수 문자인 '&'로 표기한 저널명도 다수 있다. 이외 '이음표(-)', '반쌍점(;)' 등 다양한 특수 문자 포함 여부에 따라 동일한 저널에 대한 표기가 달라짐(〈그림 9〉 참조).

## 4.9 동일 저널명의 ISSN 비일치 및 ISSN 결여

저널을 식별하는데 ISSN은 중요한 요소이다. 그런데 〈그림 10〉과 같이 저널명은 동일하나 P-ISSN 혹은 O-ISSN이 일치하지 않은 경우, 두 저널이 동일한 저널인지 판단하기 어렵다. Wiley 출판사에서 제공하는 이용통계 데이터에서 저널명은 동일하나 ISSN 정보가 다른 저널들이 다소 발견된다.

이 외에도 Cell Press, Jstor, Springer, IOP 등 다수의 출판사에서 P-ISSN 혹은 O-ISSN이 결여된 저널의 이용통계를 제공하기도 한다. 해당 정보가 아예 공란으로 나오기도 하고 'Not Available(NA)'로 표기되기도 한다. ISSN 정보가 결여되면 동일한 저널의 저널명 표기가 비일관적일 때나(4.8절 참조) 저널명이 일부 다른 경우 해당 저널들을 동일한 저널로 인식하는 데 문제가 된다.

| 저널명 | P-ISSN | O-ISSN | 출판사 | 비고 |
|---|---|---|---|---|
| Diabetic Medicine | 0742-3071 | 1096-9136 | Wiley | O-ISSN만 다름 |
| Diabetic Medicine | 0742-3071 | 1464-5491 | Wiley | O-ISSN만 다름 |
| Microbial Biotechnology | 1751-7907 | 1751-7915 | Wiley | P-ISSN만 다름 |
| Microbial Biotechnology | 1751-7915 | 1751-7915 | Wiley | P-ISSN만 다름 |

〈그림 10〉 ISSN 비일치

### 4.10 미구독 통계 제공/구독 통계 미제공 문제

국내 도서관들은 주요 전자저널의 경우 컨소시엄을 통해 패키지로 구독하고 일부만 개별 구독하고 있다. 그러나 출판사에서 제공하는 이용통계 데이터에는 컨소시엄을 통해 구독하는 저널과 개별 구독하는 저널 그리고 free trial로 제공되는 저널의 이용통계를 구별하지 않고 제공한다. 이용통계에 구독 정보가 결합되지 않으면 이용 대비 비용과 같은 저널의 비용 분석을 하기 어렵다.

Orcutt(2009)가 비판을 제기한 바와 같이 일부 출판사는 해당 기관이 구독하지 않는 저널에 대한 이용통계를 제공하기도 하는데, 미구독 저널에 대한 이용량은 모두 '0'으로 표시되어 있다. 그런데 미구독 저널의 이용통계에 표시된 '0'과 구독 저널의 실제 이용량이 '0'인 경우를 구별할 수 없어, 도서관에서 구독하지만 이용량이 저조한 저널을 찾고자 할 때 문제가 된다. 반대로 Talyor and Francis와 같은 일부 출판사에서는 기관이 구독하지만 이용량이 '0'인 저널의 이용통계를 제공하지 않는 경우도 있는데 이는 더 심각한 문제를 야기한다. 도서관에서 이용량이 저조한 저널의 구독을 취소하고자 할 때 이용량이 '0'이라 제외된 저널은 아예 빠지게 되기 때문이다. 대신 '0'보다는 이용량이 많아서 이용통계가 제공이 되는 다른 저널이 구독 취소의 대상이 된다.

이상과 같이 4절에서는 국내 STM 분야 컨소시엄에 참여하는 31개 출판사들이 생성하여 제공하는 이용통계 데이터를 분석하였다. 분석한 결과 단순한 수치 오류부터 이용통계를 활용하여 구독 및 취소 결정을 내릴 때 심각한 문제를 발생시킬 수 있는 문제까지 다양한 유형의 문제가 발견되었다. COUNTER 표준 보고서 포맷을 적용하여 출판사 간 데이터 일관성은 많이 향상되었으나, 데이터를 깊이 들여다보면 여전히 많은 오류를 포함하고 있어, 출판사 생성 이용통계 데이터가 완전하지 못하며, 완전히 신뢰할만한 수준은 아니다.

## 5. 결론 및 제언

본 논문에서는 최근 연구가 활성화되고 있는 전자저널 이용통계의 개념과 필요성에 대해 알아보고 COUNTER가 수행 중인 이용 지수

프로젝트 동향을 조사하였다. 그리고 생성 주체별 이용통계 데이터가 가지는 장점 및 한계점을 살펴보았고, 31개 출판사의 실제 이용통계 데이터를 분석하여 5가지 오류 유형과 5가지 문제 유형을 발견함으로써 출판사 생성 이용통계 데이터의 품질이 완전히 신뢰할만한 수준이 아님을 밝혔다.

서론 및 2장에서도 밝힌 바와 같이 이용통계 데이터는 도서관뿐만 아니라 정책 입안자, 기관의 예산부서, 컨소시엄 운영자, 학자들, 이용 지수 프로젝트 그룹에게도 중요한 데이터이다. 따라서 이용통계 데이터의 정확도와 신뢰도를 보장할 수 있어야 하는데, 이를 위해서는 출판사, 이용통계 데이터 사용자, 관련 표준기구, 컨소시엄 운영자 등이 다음과 같이 협력해야 한다.

출판사에서 생성하는 이용통계 데이터 품질에 대한 관리를 촉구해야 한다. 단순한 수치 오류 데이터 수정에서부터 제공방식 및 대상별 데이터의 일치 등 데이터 품질 관리를 해야 한다. 그리고 이용통계를 해당 저널의 메타데이터와 연계하여 보다 심층적인 분석을 하기 위해서는 이용통계 데이터에 결여된 저널의 식별 정보를 보충하고, 구독 계약시 제공되는 메타데이터와 이용통계에 나오는 저널 식별 데이터를 일치시키는 등 이용통계 데이터를 지속적으로 관리해야 한다.

이용 지수 프로젝트를 진행하는 그룹 및 출판사 생성 이용통계를 사용하는 그룹은 출판사 플랫폼의 인터페이스에 따라 이용통계에 영향을 미치기 때문에 보다 객관적인 비교 분석을 하기 위해서는 각 플랫폼의 인터페이스에 대한 분석이 선행되어야 한다. 반대로 이용자 측에서 비정상정인 이용행태를 보이는 경우도 있으므로, 이러한 이용행태는 다년간의 이용 추이를 분석하여 특정 기간의 이용량이 급속한 증가를 보이는 경우 이를 감지해 낼 수 있어야 한다.

마지막으로 이용통계 데이터는 개별 도서관에서도 생성할 수 있으나 예산 및 인력 부족의 현실적인 한계로 인해 출판사에서 생성하는 이용통계에 의존할 수밖에 없다. 그런데 출판사에서 제공하는 통계가 본고에서 밝힌 바와 같이 다양한 오류와 문제를 포함하고 있기 때문에 완전히 신뢰할 수 없다. 더욱이 이용통계가 저널 평가에 사용이 된다면, 각 출판사마다 자사에서 출판하는 저널의 이용량을 높이기 위해 인위적인 방법을 사용할 수도 있다. 따라서 출판사 생성 이용통계 데이터의 정확도와 신뢰도를 검증할 수 있는 방안으로 국가별 혹은 지역별 몇 개 도서관을 선정하여 해당 도서관의 이용량을 도서관 서버의 로그 분석을 통해 산출하여야 한다. 그리고 이를 출판사 생성 데이터와 비교하고 검증하는 일이 지속적으로 이루어져야 한다. 이는 개별 도서관에서 독립적으로 추진하기 어렵고 몇 개 도서관의 이용통계를 수집하여 분석하는 것이 더 신뢰도를 높일 수 있기 때문에 컨소시엄 운영기관에서 주도할 수 있다. 컨소시엄 운영기관에서는 전자학술정보 이용 로그를 분석할 수 있는 툴을 선정된 도서관에 제공해

주고, 해당 도서관에서 수집된 데이터를 확보하여 출판사에서 생성한 데이터를 비교한다. 이러한 데이터 비교 분석을 통해 출판사 생성 이용통계 데이터를 검증하고, 출판사 생성 이용통계 검증 결과를 출판사, 도서관, 타 지역 컨소시엄, 관련 표준 그룹 등 다양한 주체들과 공유할 수 있을 것이다.

## 참고문헌

김성진, 정은경, 한민혜. 2008. 전자저널 컨소시엄을 둘러싼 학술커뮤니케이션의 쟁점과 대응동향. 『정보관리연구』, 39(1): 27-52.

김혜선. 2004. 전자저널 이용통계서비스의 품질평가에 대한 연구. 『정보관리연구』, 35(4): 35-56.

심원식. 2005. 전자정보 이용통계 활용 전략. 『정보관리학회지』, 22(2): 5-21.

정영임, 김정환. 2012. 컨소시엄 기반 전자저널 이용통계 수집 및 분석 개선 방안. 『정보관리학회지』, 29(2): 7-25.

정영임, 김정환, 류범종. 2011. 『스크린 스크래핑 기반 전자저널 이용 통계 자동 수집 시스템 개발』, 2011. 6. [경주: 한국정보과학회 컴퓨터종합학술대회에서 발표].

Association of Research Libraries. Recommended Statistics and Measures for Library Networked Services, ARL New Measures Initiative The E-Metrics Project ARL E-Metrics Phase II Report. 〈http://www.arl.org/bm~doc/emetrics_module_3_statistics.ppt〉.

COUNTER. 2008. Release 3 of the COUNTER Code of Practice for Journals and Databases. 〈http://www.projectcounter.org/r3/r3_intro.pdf〉.

_____. 2012. Release 1 of the COUNTER Code of Practice for Usage Factors. 〈http://www.projectcounter.org/usage_factor.html〉.

Davis, P. M. & Price, J. S. 2006. "eJournal interface can influence usage statistics: implications for libraries, publishers, and Project COUNTER." *JASIST*, 57(9): 1243-1248.

Duy, J. & Vaughan, L. 2003. "Usage Data for Electronic Resources: A Comparison Between Locally Collected and Vendor-Provided Statistics." *Journal of Academic Librarianship*, 29(1): 16-22.

Gatten, J. & Sanville, T. 2004. "An Orderly Retreat from the Big Deal: Is It Possible for Consortia?" *D-Lib Magazine*, 10: 10. 〈http://www.dlib.org/dlib/october04/gatten/10gatten.html〉.

Hahn, K. & Faulkner L. A. 2002. "Evaluative Usage-based Metrics for the Selection

of E-journals." *College & Research Libraries*, 63(3): 215-227.

Kemp, R. 2008. E-resource Evaluation & Usage Statistics, VDM Verlag Dr.Muller

Luther, J. 2001. *White Paper on Electronic Journal Usage Statistics*. Council on Library and Information Resources, Washington, D.C.

MacIntyre, R. 2011. The Journal Usage Statistics Portal(JUSP). Paper presented at ICOLC 13th Europe Meeting, Istanbul, Turkey.

McDonald, J. D. 2006. "Understanding Online Journal Usage: A Statistical Analysis of Citation and Use." *Journal of the American Society for Information Science and Technology*, 57: 13.

Morrison, H. 2005. *The Implications of Usage Statistics as an Economic Factor in Scholarly Communications*. Usage Statistics of E-Serials Haworth Press. 〈http://hdl.handle.net/10760/6816〉.

Orcutt, D. edt. 2009. *Library Data: Empowering Practice and Persuasion*. Libraries Unltd Inc.

Shepherd, P. 2012. Journal Usage Factor: results, recommendations and next steps. 〈http://www.projectcounter.org/documents/Journal_Usage_Factor_extended_report_July.pdf〉.

Welch, J. M. 2005. "Who says we're not busy? Library web page usage as a measure of public service activity." *Reference Services Review*, 33(4): 371-379.