한국기록관리학회지
21(2), 155-172, 2021.5
http://dx.doi.org/10.14404/JKSARM.2021.21.2.155
pISSN 1598-1487 eISSN 2671-7247

JKSARM
Journal of Korean Society of Archives and Records Management

FAIR 원칙 : 데이터 관점의 디지털 아카이브 구현을 위한 고려사항

FAIR Principles: Considerations for Implementing Digital Archives from a Data Perspective

김학래(Haklae Kim)

E-mail: haklaekim@cau.ac.kr

중앙대학교 사회과학대학 문헌정보학과 교수

OPEN ACCESS

논문접수 2021-04-20
최초심사 2021-04-22
게재확정 2021-05-10

ORCID
Haklae Kim
https://orcid.org/0000-0002-2616-421X

초 록

디지털 아카이브는 디지털 자원을 보존하고 지속적으로 활용하기 위한 전자화된 저장소이다. 디지털 아카이브에 대한 이론적 연구는 활발하게 진행되고 있고, 다양한 도메인의 디지털 자원을 기록하기 위한 아카이브가 구축되어 서비스되고 있다. 그러나 디지털 아카이브의 자원은 디지털화라는 본래의 목적은 만족할 수 있지만, 자원의 검색과 재사용에 있어 여전히 제한이 있는 것이 현실이다. 본 연구는 FAIR 데이터 원칙을 자세히 살펴보고, 디지털 아카이브에 적용하기 위한 성숙도 평가 프레임워크를 제안한다. FAIR 데이터 원칙은 디지털 자원을 기계가 읽고 처리할 수 있게 만드는 일련의 지침으로 웹에 존재하는 모든 자원을 대상으로 적용할 수 있다. FAIR 데이터 원칙의 평가 모델은 계획 수립과 적용 단계를 구분해서 정의하고 있다. 그러나, 개별 원칙의 적용 여부를 평가하기 위한 명확한 기준이 모호하고, 디지털 아카이브 분야를 위한 평가 기준에 대한 논의가 미흡하다. 본 연구는 디지털 아카이브에 FAIR 데이터 원칙을 적용하기 위한 프레임워크를 제안하고, 향후 적용을 위한 이슈를 논의한다.

ABSTRACT

Digital archives are electronic storages used to preserve and utilize digital resources sustainably. Theoretical research on digital archives is being conducted actively, and digital archives for recording various resources in heterogeneous domains are being built and serviced. However, although the original purpose of digitizing the resources of digital archives is achievable, the discovery and reuse are still limited. This study examines the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles in detail and proposes a maturity assessment framework for digital archives. The FAIR Data Principles is a set of guidelines that enable machines to read and understand digital resources that are applied to any online resource. The evaluation model of the FAIR data principle defines the planning and application stages separately. However, criteria for evaluating the application of individual principles are still ambiguous, and discussions on evaluation criteria for the field of digital archives are insufficient. This study proposes a framework for applying the FAIR data principle to digital archives and discusses issues for future application.

Keywords: FAIR, 탐색가능성, 접근가능성, 상호운용성, 재사용성, 기계판독
FAIR, findability, accessibility, interoperability, reusability, machine readability

▪이 논문 또는 저서는 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A6A3A01078538).

1. 서 론

디지털 아카이브는 지속적으로 보존할 가치를 가진 디지털 자원을 보존하고 지속적으로 활용하기 위한 저장소이다(김유승, 2010). ‘디지털’이라는 용어에 함의되었듯이, 디지털 아카이브는 시공간에 대한 제약 없이 자원이 축적되고 동시에 활용될 수 있다(이규철, 2020). 이런 맥락에서 디지털 아카이브는 자원의 생산과 전달, 공유를 극대화하고 접근성을 향상시킬 수 있다. 한편, 모든 디지털 자원은 장기적 보존과 지속적 가치를 유지하기 위해 네트워크로 연결되고, 디지털 아카이브의 물리적 장소에 제한되지 않고 가상의 공간 안에 존재한다(Niu, 2016). 그러나, 디지털 아카이브에서 제공하는 서비스는 종종 온라인 웹 페이지와 유사한 특징을 갖고 있다. 디지털 자원이 인터넷을 통해 제공되지만, 자원에 대한 메타데이터의 부족, 텍스트로 표현된 메타데이터, 서로 연결되지 않는 메타데이터는 디지털 아카이브의 본질적 특징을 제약하는 요소이다. 한편, 도메인, 자원 유형에 따라 구축되는 디지털 아카이브의 특징이 다르기 때문에(신재민, 곽승진, 2013), 디지털 아카이브의 공통 기능을 도출하고, 이를 평가하기 위한 합의된 프레임워크를 만드는 것이 어려운 것이 현실이다.

넓게 보면, 인터넷에 존재하는 모든 자원은 데이터이다. 이런 맥락에서 디지털 아카이브의 자원은 데이터로 해석할 수 있다(Collins et al., 2018). Wilkinson et al.(2016)은 디지털 자원의 재사용성을 지원하기 위해 과학자와 기관이 연합해 제안한 일련의 이행원칙으로 FAIR 데이터 원칙을 소개하고 있다. 데이터를 찾을 수 있으려면(findable), 데이터와 부가적인 정보에 DOI(Digital Object Identifier)와 같은 고유하고 영구적인 식별자 뿐만 아니라 충분히 풍부한 메타데이터가 있어야 한다. 접근하려면(accessible), 메타데이터와 데이터를 사람과 기계가 이해할 수 있어야 한다. 데이터는 신뢰할 수 있는 저장소에 저장되어야 한다. 상호 운용이 가능하려면(interoperable), 메타데이터는 지식 표현을 위해 공식적이고 접근 가능하며 공유되고 광범위하게 적용 가능한 언어를 사용해야 한다. 재사용이 가능하려면(reusable), 데이터에 명확한 라이선스가 있어야 하며 출처에 대한 정확한 정보를 제공해야 한다. 이 원칙은 연구 데이터의 생산과 재사용 측면에서 시작되었으나(김성욱, 김선태, 2020), 전통적인 학문의 경계를 넘어 학제간 오픈 사이언스(Open Science)의 맥락으로 확산되고 있다(Stall et al., 2019). FAIR 디지털 자원은 영구 식별자, 메타데이터 규격, 실행 가능한 정책과 데이터 관리 계획, 데이터의 탐색을 위한 저장소를 핵심 요소로 정의한다(Boeckhout, Zielhuis, & Bredenoord, 2018). 특히, 데이터 공유, 데이터 형식, 메타데이터 표준, 소프트웨어와 인프라 등 데이터 수준의 상호운용성을 위한 프레임워크를 정의하고 있다.

디지털 아카이브 관점에서 FAIR 데이터 원칙은 디지털 자원의 탐색, 상호 연결, 지속 가능성을 위한 프레임워크를 제공할 수 있다. FAIR의 세부 원칙은 디지털 자원의 생산, 축적, 공유에 포괄적으로 적용할 수 있다. 특히, 디지털 아카이브의 기능에 대한 합의된 프레임워크를 구성하고 평가할 수 있는 출발점이 될 수 있다. 본 연구는 FAIR 데이터 원칙에 대한 개념과 현황을 소개하고, 디지털 아카이브에 적용하기 위한 평가 프레임워크를 제안한다. 논문의 구성은 다음과 같다. 2장은 FAIR 데이터 원칙과 관련된 이론적 연구를 소개한다. 3장은 FAIR 원칙의 개념을 소개하고, FAIR 데이터를 구성하는 프로세스를 기술한다. 4장은 FAIR 데이터의 세부 원칙을 소개한다. 5장은 FAIR 데이터의 수준 평가를 위한 모델을 소개하고, 6장에서 디지털 아카이브 관점에서 FAIR 데이터 적용을 위한 고려사항을 논의하고 평가 프레임워크를 제안한다. 마지막으로 6장에서 연구 결과를 정리한다.

2. 관련 연구

FAIR 데이터는 과학 데이터의 저장과 공유를 위한 프레임워크로 제안되었다(Wilkinson et al., 2016). FAIR 원칙은 식별자(예: DOI), 데이터와 메타데이터 표준, 통제 어휘를 사용하여 기계가 읽을 수 있고 실행 가능한

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

정보의 중요성을 강조한다. 따라서, FAIR 원칙을 만족하는 데이터는 탐색, 접근, 상호운용과 재사용이 가능한 것을 말한다(FORCE11, 2016). FAIR 데이터 원칙은 데이터의 공유와 재사용을 장려하고 활성화하기 위한 일련의 지침으로, 연구와 학술 분야에서 논의가 시작되어 다양한 분야로 빠르게 확산되고 있다(Mons et al., 2017).

FAIR 원칙은 FORCE11(FORCE11, 2016), 유럽 연합 집행위원회(European Commission)를 비롯한 다양한 조직에서 데이터를 최대한 사용하고 재사용 할 수 있는 방식으로 공유하는 데 유용한 프레임워크로 전 세계적으로 인정받고 있다(DTL, 2021). FAIR 데이터 원칙은 세계 각국의 정부와 연구기관에서 채택하고 있다. 2016년 G20 항저우 정상 회담에서 G20 지도자들은 연구 분야에 FAIR 원칙을 적용 할 것을 승인하는 성명을 발표했다(EU, 2016). 2016년에 호주 조직 그룹은 호주 정부의 연구 산출물에 대한 공정한 접근에 관한 성명서를 개발했고, FAIR 원칙을 보다 일반적으로 연구 산출물에 확장하는 목표를 갖고 있다(AFAWG, 2017). 2017년 독일, 네덜란드, 프랑스는 FAIR 이니셔티브를 지원하기 위한 국제 사무소인 GO FAIR 국제 지원과 조정을 맡는 사무소를 설립하기로 합의했다(Government of the Netherlands, 2017). CODATA(Committee on Data), RDA(Research Data Alliance)와 같은 연구 데이터 생태계에서 활동하는 다른 국제 조직도 해당 커뮤니티의 FAIR 구현을 지원하고 있다.

FAIR 데이터의 평가 방법은 다양한 접근이 시도되고 있다. FAIR 원칙은 그 자체로 실행력이 있지 않고, 개별 원칙의 내용과 설명을 기준으로 기관에 따라 조금씩 다른 평가지표를 제공하고 있다. 예를 들어, RDA의 FAIR 데이터 성숙도 모델은 평가지표, 우선순위, 평가방법으로 세분화시킨 프레임워크를 정의하고 있다(Bahim et al., 2020). 특히, 모든 지표를 지속성과 전역 고유성을 특징으로 구분하고, 데이터와 메타데이터를 별도의 측정 지표로 정의하고 있다. 한편, FAIR 성숙도 지표와 도구(FAIRMetrics, 2021)는 FAIR 원칙을 자동으로 평가하기 위한 기준과 소프트웨어 스크립트를 제공하고 있다(Wilkinson et al., 2019). FAIRsFAIR 데이터 개체 평가 매트릭스는 RDA 모델을 기본으로 하고 있고, 세부 평가 기준을 상세화하고 있다(Devaraju et al., 2021).

FAIR 원칙의 주요 목표는 향후 연구에 재사용할 수 있도록 귀중한 디지털 자산의 장기적 관리를 위한 것이다. 이러한 원칙의 범위는 일반적인 의미의 데이터를 넘어 과학 연구의 일부인 알고리즘, 워크플로우를 포함한 연구 프로세스의 모든 구성요소를 포함한다. FAIR 데이터 원칙은 오픈 사이언스(Open Science)를 위한 연구 데이터의 접근과 재사용을 극대화하기 위한 배경으로 시작되었으나(Cousijn et al., 2018; Helliwell et al., 2019), 다양한 학문 분야를 넘어 보편적 데이터의 관리를 위한 도구로 인식되고 있다(Corpas et al., 2018; David et al., 2020; Haux & Knaup, 2019). Candela et al.(2020)는 GLAM 기관의 디지털 자원을 재사용하기 위해 링크드 데이터를 적용하고, 상호 연결된 자원을 의미적으로 탐색하고, 시각화하는 애플리케이션을 소개한다. 이 연구에서 FAIR 원칙이 디지털 자원에 직접 적용되지 않았지만, 저자는 디지털 자원의 재사용을 위해 중요성을 강조하고 있다. Koster와 Koster와 Woutersen-Windhouwer(2018)는 LAM(Library, Archive, Museum) 컬렉션의 검색과 재사용을 극대화하기 위해 FAIR 원칙의 적용 방안을 제안하고 있다. 문화유산 컬렉션은 학술 산출물에 대한 FAIR 원칙과 유사한 성격으로 갖고 있기 때문에, FAIR 원칙의 확장 없이 개별 원칙의 특성을 구체화하고 있다. Barbuti(2020)는 디지털 문화유산 분야의 기록 보존을 위해 FAIR 원칙의 적용을 주장하고 있다. 특히, 기록 자원의 장기 보존을 보존하기 위해 원칙 R(Reusable)을 확장하는 방안을 제안하고 있다. 그러나, 도서관과 디지털 아카이브 분야에서 FAIR 원칙에 대한 논의가 있지만(Calamai & Frontini, 2018; Hettne et al., 2020; Nitecki & Alter, 2021), 디지털 자원에 적용하거나, 평가한 사례는 미흡한 것이 현실이다. 국내에서 FAIR 데이터 원칙은 연구데이터 분야를 중심으로 진행되고 있고(최명석 외, 2017; 김성훈, 오삼균, 2019, 김성태, 김선태, 2020), 연구 데이터의 발행과 공유를 위해 FAIR 데이터 원칙을 준수하기 위한 가이드라인을 제공하고 있다(국가과학기술연구회, 2019). 그러나, FAIR 데이터 원칙을 적용하고 평가하기 위한 심도 있는 연구는 미흡한 상황이다.

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

3. FAIR 원칙

3.1 개요

FAIR 데이터 원칙(Findable, Accessible, Interoperable, Reusable)은 2015년 네덜란드 라이덴(Leiden)의 Lorentz Center 워크숍에서 초안이 작성되었다. 초기 제안된 FAIR 데이터 원칙은 디지털 자산의 재사용성을 지원하기 위해 과학자와 기관이 공동으로 추진하기 위한 일련의 이행 원칙을 포함하고 있다. 한편, FAIR 원칙은 데이터와 메타데이터가 '기계 판독할 수 있도록' 여러 데이터세트의 수집과 분석을 통해 새로운 발견을 지원하는 목적을 갖고 있다. 그러나, 기계 가독성(machine readability)과 함께 데이터 재사용을 위해 사람의 이해를 강조하고 있다.

3.2 개념적 특징

FAIR 데이터 원칙은 다양한 개념과 기술적 특징을 융합했기 때문에, 유사한 개념이 있다. 먼저, FAIR 데이터 원칙은 표준의 관점으로 접근하지 않는다. 실제 GO FAIR는 FAIR 데이터를 실현하기 위해 인터넷 표준을 활용해야 하고, 다양한 표준, API와 네트워크 프로토콜을 준용하여 FAIR 원칙을 적용하는 것을 권고한다(GO FAIR, 2021).

FAIR 데이터 원칙은 오픈 데이터와 동일한 개념이 아니다. 오픈 데이터에서 '개방(open)'은 저작권, 특허 또는 기타 통제 방식의 제한 없이 누구나 원하는 대로 일부 데이터를 자유롭게 사용하고 다시 배포하는 것을 의미한다. FAIR 원칙의 "A"는 "잘 정의된 조건에 따라 접근" 할 수 있음을 의미하지만, 데이터의 개방성과 관련된 도덕적·윤리적 문제를 명시적으로 다루지 않는다. 예를 들어, 정부의 세금으로 생성·관리되는 데이터는 오픈 데이터로 개방을 권장하고 있으나, 개인정보보호 또는 국가 안보의 목적으로 제한된 접근만 허용할 수 있다. FAIR 원칙은 데이터의 사용과 개방 여부를 데이터 소유자의 권한으로 정의한다. 따라서 데이터의 무조건적 개방 여부보다 재사용 원칙을 준수하기 위해 명시적이고, 기계가 처리할 수 있는 라이선스 제공을 권장한다.

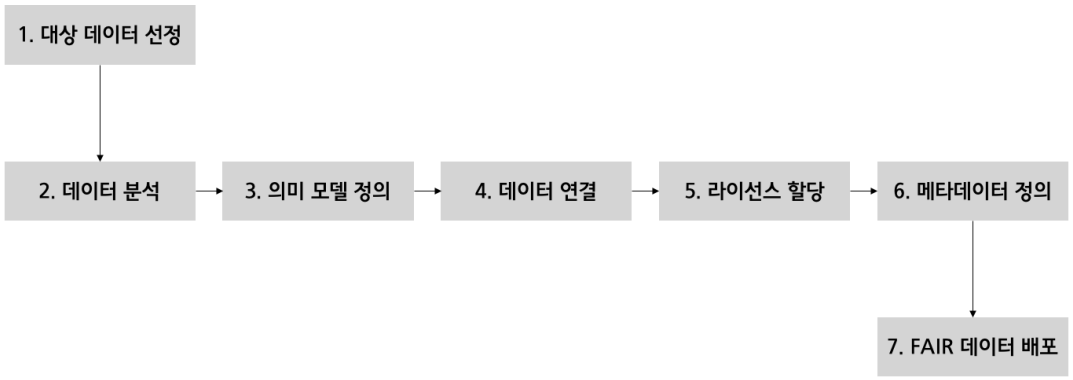
FAIR 원칙은 데이터와 메타데이터의 기계 판독을 중요한 요소로 정의하고, 지식 표현을 위해 보편적인 프레임워크의 활용을 권장하고 있다. RDF(Resource Description Framework)와 링크드 데이터(Linked Data)는 지식 표현과 교환, 표현된 지식의 연계를 지원하는 표준 프레임워크이다(김학래, 2017). 온톨로지는 메타데이터 수준에서 상호운용성과 지식 공유의 목적에 매우 효과적이다. FAIR의 도입 초기, RDF를 채택하여 지식 공유를 실현한 다양한 모범사례가 있다(Guizzardi, 2020). 그러나, 사물 인터넷, 빅데이터 분야와 같이 대규모 데이터의 고성능 분석을 위한 목적에서 지식 표현이 반드시 필요하지 않을 수 있고, 데이터의 특성에 따라 기계 판독의 방식도 다를 수 있다. 이런 이유에서, FAIR 원칙에서 링크드 데이터의 개념과 기술을 차용하고 있지만, FAIR 원칙은 RDF와 링크드 데이터 프레임워크를 필수요소로 규정하지 않는다.

3.3 FAIR 데이터 프로세스

FAIR 데이터화(FAIRification)는 대상 데이터를 FAIR 데이터의 원칙을 적용하여 진단하고, FAIR 데이터로 구축하는 일련의 단계로 구성된다(Jacobsen et al., 2019; Sinaci, Núñez-Benjumea, & Gencturk, 2020). <그림 1>에서 보듯이, FAIR 데이터화는 7 단계로 구성되고, 각 단계는 데이터세트의 FAIR 상태(FAIRness)를 향상시키는 것을 목표로 한다. 각 단계의 순서는 엄격하게 정의되지 않으며, 필요에 따라 반복될 수 있다.

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

- 단계 1. 데이터 선정: 평가할 데이터를 선정하고 접근 권한을 확보한다.
- 단계 2. 데이터 분석: 데이터의 내용(데이터 요소 사이의 관계)과 물리적 정보(데이터 구조, 데이터 식별 정보)에 대해 분석한다. 예를 들어, 데이터세트가 관계형 데이터베이스라면 관계형 스키마는 데이터세트의 구조, 관련된 유형(필드명), 출현횟수 등에 대한 정보를 제공한다.
- 단계 3. 의미 모델 정의: 데이터세트의 개체와 관계를 의미 모델(semantic model)로 정의하고, 명시적으로 컴퓨터가 처리 가능할 수 있도록 기술한다. 좋은 의미 모델은 어떤 문제를 해결하기 위해 정의된 도메인의 합의된 관점을 포함하고 있어야 한다. 의미 모델은 기존에 개발되거나 커뮤니티에서 사용되고 있는 온톨로지 어휘를 적용할 수 있다. 이러한 개념 모델을 사용하면 제공된 용어, 개념, 그리고 개념 구조를 통해 데이터 모델과 항목을 정확하게 분류할 수 있다.
- 단계 4. 데이터 연결: 3 단계에서 정의한 의미 모델을 적용하여 대상 데이터를 연결 가능한 데이터로 변환할 수 있다. 이 단계는 링크드 데이터와 지식그래프 기술을 사용하여 수행된다. 이 단계는 상호운용성과 재사용을 촉진하여 서로 다른 유형의 데이터와 시스템을 쉽게 통합할 수 있다.
- 단계 5. 라이선스 할당: 라이선스 정보는 메타데이터의 일부이지만, FAIR 데이터화(FAIRification)에서 별도의 단계로 정의해 라이선스의 중요성을 강조한다. 명시적 라이선스가 없으면 데이터가 개방되어도 재사용하지 못하거나, 재사용에 있어 제약을 만들 수 있다.
- 단계 6. 데이터세트에 대한 메타데이터 정의: 적절하고 풍부한 메타데이터는 FAIR에서 정의한 모든 원칙에서 매우 중요하다.
- 단계 7. FAIR 데이터 자원 배포: 관련 메타데이터, 라이선스와 함께 FAIR 데이터 원칙으로 평가된 정보를 배포하거나 게시하여 메타데이터가 검색 엔진에 의해 색인화될 수 있고, 인증, 승인이 필요한 데이터에 적합한 방법을 통해 접근할 수 있다.



〈그림 1〉 FAIR 데이터 프로세스(GO FAIR, 2021)

4. FAIR 원칙의 요소

FAIR 데이터 원칙은 15개로 구성되어 있으며, 기관에 따라 원칙별 설명이 조금씩 차이가 있다. 본 연구는 GO FAIR(2021), FORCE11(2016), DTL(2021)에서 소개한 설명을 준용하여 소개한다. FAIR는 네 가지 특징의 앞 글자로 구성되어 있다. 각각의 원칙은 세부적인 특징으로 구체화되고, 원칙의 첫 글자와 숫자를 결합해서 정의

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

한다. 예를 들어, 탐색가능성(F)은 ‘F1’, ‘F2’로 정의하고, 접근가능성(A)은 세부 항목을 구체화시켜 ‘A1’과 ‘A1.1’, ‘A1.2’로 구분한다.

4.1 탐색성(Findability)

데이터를 사용하는 첫 단계는 탐색에서 시작한다. 기계가 읽을 수 있는 데이터는 자동 검색에 필수적이기 때문에 FAIR 데이터화에서 매우 중요하다. 탐색가능성을 지원하기 위해 데이터는 영구적인 식별자(예: DOI 또는 Handle)가 할당되어야 하고, 데이터에 대한 풍부한 메타데이터를 포함해야 한다. 더불어, 검색 포털(국내와 국제)을 통해 탐색할 수 있어야 한다.

4.1.1 F1: (메타)데이터는 전역적으로 고유하고 영구적인 식별자가 할당되어야 한다.

식별자(identifier)는 컴퓨터가 의미 있는 방식으로 데이터를 해석 할 수 있게 한다. F1은 식별자에 대해 두 가지 조건을 규정한다.

- 전역적(globally)으로 고유해야 한다. 특정 자원에 부여된 식별자는 웹에서 유일한 값을 갖고, 다른 자원과 구분되는 식별자로 표현해야 한다. 식별자의 고유성을 보장하기 위한 알고리즘을 활용할 수 있고, 특정한 자원을 저장하고 탐색하기 위한 레지스트리 서비스(registry service)는 전역적으로 고유한 식별자를 생성해야 한다.
- 지속적(persistent)이어야 한다. 웹에서 연결 정보를 유지하려면 시간과 비용이 필요하고 시간이 지남에 따라 해당 정보가 무효화될 수 있다. 레지스트리 서비스는 미래의 특정 시점까지 해당 링크에 접근할 수 있도록 보장해야 한다.

전역 고유 식별자(Globally Unique Identifier)가 항상 유일한 값이 만들어진다는 보장은 없지만, 적절한 알고리즘이 있다면 중복되는 숫자를 생성할 가능성이 매우 낮다. 전역적으로 고유하고(globally unique) 영구적인(persistent) 식별자는 메타데이터의 모든 요소와 데이터세트의 모든 개념에 고유 식별자를 할당하여 게시된 데이터의 의미를 명확하게 만든다. 이런 맥락에서 F1의 식별자는 URI(Uniform Resource Identifier)와 같은 인터넷 링크로 구성하는 것이 일반적이다.

4.1.2 F2: 데이터는 풍부한 메타데이터로 기술되어야 한다.

메타데이터는 데이터세트를 설명하는 정보이다. 디지털 자원을 생성할 때, 메타데이터는 데이터의 맥락, 품질, 상태 또는 특성에 대한 다양하고 일반적인 기술 정보를 포함해야 한다. 데이터에 대한 맥락은 데이터세트가 저장된 URL, 작성자, 출처, 목적, 시간, 지리적 위치, 접근 조건, 데이터 수집과 이용에 대한 약관을 포함할 수 있다. 데이터세트에 따라 제공되는 메타데이터의 범위는 다를 수 있다. 표준화된 통제 어휘(controlled vocabulary), 시소러스(thesaurus), 온톨로지는 메타데이터를 상호운용하고, 기계가 읽고 처리할 수 있는(machine actionability) 방식으로 만들 수 있다. 통제 어휘(controlled vocabularies)는 체계적이고 표준화된 용어 목록이며, 분야별로 다양하다. 통제 어휘는 기계와 사용자가 메타데이터를 훨씬 더 쉽게 이해하는데 사용되기 때문에, 데이터 탐색가능성을 향상시키는데 효과적이다. 텍소노미(taxonomy)는 정렬된 시스템의 개체 분류이다. 텍소노미에 정의된 용어는 콘텐츠・데이터 기술에 추가하여 데이터를 식별하는데 사용할 수 있다. 구조화된 방식으로 데이터를 식별하는 것은 하나의 검색 질의로 관련 있는 데이터를 쉽게 찾게 할 수 있다. 따라서 데이터세트 설명에 텍소노미 용어를 추가하

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

면 데이터세트의 탐색가능성이 향상된다. 온톨로지와 링크드 데이터는 데이터가 서로 다른 데이터를 의미적 수준에서 상호 연결하고, 접근할 수 있게 해준다(김학래, 2017). 데이터를 이종의 데이터에 연결하면 더 많은 지식과 연결할 수 있고, 궁극적으로 데이터의 탐색가능성을 높일 수 있다. 이런 특징 때문에, F2 원칙은 데이터세트의 접근성 (A), 상호운용성 (I), 재사용 (R) 원칙과 밀접한 관련이 있다.

4.1.3 F3: 메타데이터는 기술하는 데이터에 대한 명확하고 명시적인 식별자를 포함해야 한다.

일반적으로 데이터세트를 기술한 메타데이터는 원본 데이터세트와 별도로 저장되거나 개별적인 파일로 존재한다. F3은 메타데이터에 기술된 정보가 전역적으로 고유하고 영구적인 식별자로 기술되어야 한다는 원칙이다. 즉, 메타데이터 파일과 데이터세트 사이의 연결이 명시적으로 표현되어야 한다.

4.1.4 F4: (메타)데이터는 검색 가능한 자원에 등록 또는 색인화되어야 한다.

식별자와 풍부한 메타데이터 기술만으로 웹에서 탐색가능성을 보장하는데 한계가 있다. 완벽하고 좋은 데이터 자원이라고 해도 그 존재를 모르면 전혀 사용할 수 없다. 따라서 가능한 모든 디지털 자원은 검색이 가능할 수 있도록 특정한 서비스에 색인으로 저장되어야 한다. 예를 들어, 대부분의 검색 엔진은 웹에 공개된 웹 페이지를 수집하고, 해당 서비스에서 색인으로 저장함으로써 검색 서비스에서 활용할 수 있다. 원칙 F1 ~ F3은 데이터세트에 대한 색인과 검색 서비스를 위한 필수요소이다.

4.2 접근성(Accessibility)

원칙 A는 표준화된 프로토콜을 사용하여 데이터에 접근하는 것을 포함한다. 데이터세트와 해당 메타데이터의 접근성은 사용자가 데이터세트를 평가하고 잠재적으로 재사용하는데 필수적이다. 데이터 접근성에 대한 수준은 시간 경과에 따라 달라질 수 있다. 특히, 데이터 자체를 더 이상 사용할 수 없는 상황에서, 해당 저장소는 메타데이터에 대한 정보를 제공하고, 데이터세트에 대한 라이선스를 지속적으로 보장해야 한다. 라이선스의 보장은 데이터세트에 접근할 수 있는 범위 또는 상황을 결정하는 요소이다.

4.2.1 A1: (메타)데이터는 표준화된 통신 프로토콜을 사용하여 식별자로 검색된다.

대부분의 사용자는 웹에 있는 링크를 선택하여 데이터를 검색한다. HTTP 프로토콜은 웹에서 이루어지는 모든 데이터 교환의 기초이며, HTML 문서와 같은 자원을 가져올 수 있도록 해주는 프로토콜이다. 원칙 A1은 FAIR 데이터의 접근에 있어 전문적이거나 독점화되지 않은 통신 방법으로 디지털 자원에 접근하는 것을 정의하고 있다. 일반적으로, HTTP와 FTP는 TCP 프로토콜을 기반으로 다른 통신 프로토콜보다 쉽고 효과적으로 디지털 자원을 요청하고 제공할 수 있다. 그러나, 민감한 데이터(sensitive data)라면, 모든 상황에서 접근이 허용되지 않을 수 있고, 데이터 접근을 위한 제약사항이 존재할 수 있다. 원칙 A1은 민감 데이터에 대한 접근을 위한 정보를 명시적으로 기술하는 것을 권장하고, 접근 방법에 대해 논의할 수 있도록 담당자의 이메일 또는 연락처 정보의 제공을 권고하고 있다.

4.2.2 A1.1: 프로토콜은 개방형이며, 무료이고 보편적인 방법으로 구현 가능해야 한다.

데이터 재사용과 검색을 극대화하기 위해 프로토콜은 무료로, 오픈소스를 이용해 전역적으로 구현할 수 있어야 한다. 컴퓨터와 인터넷 연결만 있으면 누구나 메타데이터에 접근할 수 있어야 한다. 따라서 이 원칙은 데이터를 공유할 저장소를 선택하는데 중요한 요소이다. 대규모 데이터를 저장하고 탐색할 수 있는 오픈소스 기반의 데이터

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

포털(예: CKAN)은 이 원칙을 구현하기 위해 검토할 수 있다.

4.2.3 A1.2: 프로토콜은(필요한 경우) 인증과 권한 부여 절차를 허용해야 한다.

FAIR는 데이터에 접근하기 위한 정확한 조건을 제공하는 것을 권장한다. FAIR에서 원칙 'A'는 반드시 '개방' 또는 '무료'를 의미하지 않으며, 엄격하게 보호되는 개인 데이터도 포함할 수 있다. 접근성은 기계가 요구사항을 자동으로 이해하여 필요한 요구사항을 실행하거나, 사용자에게 요구사항을 공지하는 방식을 정의할 수 있다. 예를 들어, 사용자 계정을 통해 권한을 인증하고, 특정 데이터세트에 접근할 수 있다. 저장소 서비스에서 사용자 계정에 따라 접근할 수 있는 데이터세트가 차별화될 수 있다.

4.2.4 A2: (메타)데이터는 더 이상 사용할 수 없는 경우에도, 계속 액세스 할 수 있어야 한다.

데이터 자원에 대한 온라인 상태를 유지하는데 비용이 들기 때문에 데이터세트는 일정 시간이 지나면 사라질 수 있다. 링크가 무효화된 데이터세트는 사용자가 접근할 수 없다는 것을 의미한다. 일반적으로 데이터세트가 더 이상 존재하지 않을 때, 관련된 메타데이터도 함께 사라진다. 원칙 A2는 데이터가 더 이상 유지되지 않는 경우, 메타데이터가 유지되어야 한다는 것을 권고한다. A2는 F4에 설명된 데이터의 등록과 색인화 문제와 관련이 높다.

4.3 상호운용성(Interoperability)

메타데이터는 커뮤니티에서 합의한 표준, 어휘를 사용해야 하며, 식별자에 대한 정보를 기술한 링크를 포함해야 한다.

4.3.1 I1: (메타)데이터는 지식 표현을 위해 공식적이고 접근 가능하며 공유되고 광범위하게 적용 가능한 언어를 사용해야 한다.

일반적으로 상호운용성은 컴퓨터 시스템이 적어도 다른 시스템의 데이터 교환 형식을 인지 또는 합의된 것을 의미한다. 이를 위해 (1) 통제 어휘, 온톨로지, 시소러스의 사용, (2) 메타데이터를 기술하고 구조화하기 위해 정의된 프레임워크를 기반으로 적절한 데이터 모델을 사용하는 것이 중요하다.

4.3.2 I2: (메타)데이터는 FAIR 원칙을 따르는 어휘를 사용해야 한다.

데이터세트를 설명하는데 사용되는 통제 어휘는 문서로 명시적으로 기술하고, 전역적으로 고유하고 영구적인 식별자를 적용해야 한다. 통제 어휘, 온톨로지, 시소러스는 그 자체로 탐색에 활용할 수 있고, 접근 가능해야 한다. 동시에, 어휘 사이의 상호운용이 확보될 수 있도록 기계가 처리할 수 있는 방식으로 기술되어야 한다. 한편, 이러한 모든 명세서는 데이터세트를 사용하는 모든 사람이 쉽게 찾고 접근할 수 있어야 한다.

4.3.3 I3: (메타)데이터는 다른 메타데이터에 대한 정규화된 참조가 포함되어야 한다.

데이터세트는 다른 데이터세트의 연결을 통해 의미 있는 정보를 제공할 수 있다. 특정 데이터세트가 모든 정보를 포함하지 않고, 권위 있는 참조를 참조함으로써 데이터의 신뢰도를 높일 수 있다. 예를 들어, 위키피디아를 참조하는 방법은 디지털 자원에 대한 정규화된 참조를 제공하고, 해당 출처의 업데이트를 신속히 반영할 수 있는 장점을 확보할 수 있다.

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

4.4 재사용성(Reusability)

FAIR 데이터 원칙의 궁극적인 목표는 데이터의 재사용성을 촉진하는 것이다. 이를 위해 데이터를 재사용하는 기준을 명확하게 기술한 라이선스 정보가 필수적이다.

4.4.1 R1: (메타)데이터는 정확하고 관련성이 높은 속성이 여러 개 있어야 한다.

데이터에 많은 정보가 첨부되어 있으면 데이터를 찾고 재사용하는 것이 훨씬 쉽다. 원칙 R1은 F2와 관련이 있다. 다만, R1은 사용자(기계 또는 사람)가 특정 상황에서 데이터가 실제로 유용한지 여부를 결정하는 능력에 중점이 있다. 데이터 게시자는 검색을 허용하는 메타데이터뿐만 아니라 데이터가 생성된 맥락을 풍부하게 설명하는 메타데이터도 제공해야 한다.

4.4.2 R1.1: (메타)데이터는 명확하고 접근 가능한 데이터 사용 라이선스로 공개되어야 한다.

원칙 'I'는 기술적 상호운용성을 다루고 있지만, R1.1은 법적 상호운용성에 관한 것이다. 데이터에 어떤 사용권한을 부여했는지에 대한 명확한 설명이 포함되어야 한다. 라이선스에 대한 모호함은 데이터 재사용을 심각하게 제한할 수 있다. 데이터를 사용할 수 있는 조건은 기계와 사람에게 모두 명확하게 제공되어야 한다.

4.4.3 R1.2: (메타)데이터는 상세화된 출처와 관련이 있다.

데이터 재사용을 위해 데이터 출처 정보가 필요하다. 예를 들어, 데이터가 생성된 방식, 재사용 할 수 있는 맥락과 신뢰도, 데이터 출처, 인용에 대한 정보가 명확하게 기술되어야 한다. 특히, 출처는 데이터를 검증하는데 있어 과학 데이터베이스의 핵심 문제이다. 원칙 I3은 게시된 데이터세트를 재사용하기 위해 확인할 수 있는 중요한 정보를 포함한다.

4.4.4 R1.3: (메타)데이터는 도메인 관련 커뮤니티 표준을 충족해야 한다.

데이터세트의 재사용은 여러 가지 조건을 포함할 수 있다. 예를 들어, 동일한 유형의 데이터, 표준화된 방식으로 구성된 데이터, 잘 확립되고 지속 가능한 파일 형식, 공통 템플릿을 따르고 공통 어휘를 사용하는 문서(메타데이터)인 경우, 데이터의 재사용은 매우 수월할 수 있다. 한편, 데이터 저장과 공유에 대한 커뮤니티 표준 또는 모범 사례가 있다면 준수하는 것을 권장한다. FAIR 데이터는 최소한 이러한 기준을 충족해야 한다.

5. FAIR 데이터 성숙 모델

5.1 개요

FAIR 데이터 원칙은 기계가 읽고 처리할 수 있는 디지털 자원의 속성을 정의하고 있다. 그러나, FAIR 데이터의 개별 원칙의 평가 요소는 모호한 측면이 있고, 기관에 따라 조금씩 다르게 정의하고 있다. FAIR 데이터 원칙이 실용적으로 활용되려면, 개별 원칙과 이에 대한 평가 방법이 구체화될 필요가 있다. 예를 들어, ARDC(Australian Research Data Commons, 2021)와 DANS(Data Archiving and Networked Services, 2021)의 평가도구는 4개의 FAIR 원칙에 대해 12개의 평가 항목을 제공하고 있다. 그러나, ARDC의 탐색 가능성은 식별자, 메타데이터, 저장

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

소에 대한 항목이지만, DANS는 메타데이터 항목을 통제어휘, 텍소노미, 온톨로지로 구체화하고 있고, 부가적인 메타데이터 항목을 측정하고 있다. 즉, 현재 개발된 평가 프레임워크 사이에 차이점이 존재한다. 특히, 개별 평가 항목의 측정 지표에 대한 표준화된 방안이 미흡한 것이 현실이다.

본 연구는 평가 프레임워크의 상세한 특징을 기술하지 않고, 정교하게 설계한 FAIR 데이터 성숙도 모델을 전반적으로 소개한다. RDA 워킹그룹에서 개발한 FAIR 데이터 성숙도 모델은 FAIR 원칙의 수준을 평가하는 공통 기준을 정의하고 있다 (Bahim et al., 2020). FAIR 데이터 성숙도 모델의 프레임워크는 지표(indicator), 우선순위(priority), 평가 방법(assessment method)으로 구성된다. <표 1>에 보듯이, 이 모델은 측정하는 항목을 세분화시키고 있다. 현재 개발된 대부분의 모델은 FAIR 원칙을 기준으로 평가 방식을 차별화하고 있다. 예를 들어, FAIR Metrics(Wilkinson et al., 2019)는 FAIR 원칙을 프로그래밍 환경에서 측정할 수 있도록 자동화된 도구로 개발하고 있고, FAIRsFAIR(Devaraju et al., 2020)는 자체적으로 개발한 평가 지표와 세부 가이드라인을 제공하고 있다. RDA 모델과 비교하면, 두 가지 프레임워크에서 정의한 지표는 FAIR 원칙에서 크게 확장되지 않았다.

〈표 1〉 FAIR 원칙의 평가 프레임워크의 비교

FAIR	FAIR	RDA ID	설명	우선순위	FAIR Metrics	FAIRsFAIR
F	F1	RDA-F1-01M	메타데이터는 영구 식별자(persistent identifier)로 식별된다.	필수	FM-F1B	
	F1	RDA-F1-01D	데이터는 영구 식별자(persistent identifier)로 식별된다.	필수	FM-F1B	FsF-F1-02D
	F1	RDA-F1-02M	메타데이터는 전역 고유 식별자(globally unique identifier)로 식별된다.	필수	FM-F1A	
	F1	RDA-F1-02D	데이터는 전역 고유 식별자(globally unique identifier)로 식별된다.	필수	FM-F1A	FsF-F1-01D
	F2	RDA-F2-01M	풍부한 메타데이터가 탐색을 위해 제공된다.	필수	FM-F2	FsF-F2-01M
	F3	RDA-F3-01M	메타데이터는 데이터를 위한 식별자를 포함한다.	필수	FM-F3	FsF-F3-01M
	F4	RDA-F4-01M	메타데이터는 수집과 색인이 가능한 방식으로 제공된다.	필수	FM-F4	FsF-F4-01M
A	A1	RDA-A1-01M	메타데이터는 사용자가 데이터에 접근할 수 있는 정보를 포함하고 있다.	중요		FsF-A1-01M
	A1	RDA-A1-02M	메타데이터는 수동으로 접근할 수 있다(사람의 개입).	필수		
	A1	RDA-A1-02D	데이터는 수동으로 접근할 수 있다(사람의 개입).	필수		
	A1	RDA-A1-03M	메타데이터 식별자는 메타데이터 레코드로 확인된다.	필수		
	A1	RDA-A1-03D	데이터 식별자는 디지털 객체로 확인된다.	필수		
	A1	RDA-A1-04M	메타데이터는 표준 식별자로 접근할 수 있다.	필수		
	A1	RDA-A1-04D	데이터는 표준 식별자로 접근할 수 있다.	필수		
	A1	RDA-A1-05D	데이터는 자동으로 접근할 수 있다(컴퓨터 프로그램).	중요		
	A1.1	RDA-A1.1-01M	메타데이터는 무료 접근 프로토콜로 접근할 수 있다.	필수	FM-A1.1	FsF-A1-02M
	A1.1	RDA-A1.1-01D	데이터는 무료 접근 프로토콜로 접근할 수 있다.	중요	FM-A1.1	FsF-A1-03D
	A1.2	RDA-A1.2-01D	데이터는 인증과 권한 부여를 지원하는 접근 프로토콜로 접근할 수 있다.	유용	FM-A.1.2	
	A2	RDA-A2-01M	메타데이터는 데이터를 사용할 수 없는 상황에서도 사용할 수 있도록 보장한다.	필수	FM-A2	FsF-A2-01M

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

FAIR	FAIR	RDA ID	설명	우선순위	FAIR Metrics	FAIRsFAIR
I	I1	RDA-I1-01M	메타데이터는 표준 형식으로 표현된 지식 표현을 사용한다.	중요	FM-I1	FsF-I1-01M
	I1	RDA-I1-01D	데이터는 표준 형식으로 표현된 지식 표현을 사용한다.	중요	FM-I1	
	I1	RDA-I1-02M	메타데이터는 기계가 이해할 수 있는 지식 표현을 사용한다.	중요	FM-I1	FsF-I1-02M
	I1	RDA-I1-02D	데이터는 기계가 이해할 수 있는 지식 표현을 사용한다.	중요	FM-I1	
	I2	RDA-I2-01M	메타데이터는 FAIR 원칙과 호환되는 어휘를 사용한다.	중요	FM-I2	
	I2	RDA-I2-01D	데이터는 FAIR 원칙과 호환되는 어휘를 사용한다.	유용	FM-I2	
	I3	RDA-I3-01M	메타데이터는 다른 메타데이터에 대한 참조를 포함한다.	중요	FM-I3	
	I3	RDA-I3-01D	데이터는 다른 메타데이터에 대한 참조를 포함한다.	유용	FM-I3	
	I3	RDA-I3-02M	메타데이터는 다른 데이터에 대한 참조를 포함한다.	유용	FM-I3	FsF-I3-01M
	I3	RDA-I3-02D	데이터는 다른 데이터에 대한 정규화된 참조가 포함된다.	유용	FM-I3	
	I3	RDA-I3-03M	메타데이터는 다른 메타데이터에 대한 정규화된 참조가 포함된다.	중요	FM-I3	
	I3	RDA-I3-04M	메타데이터는 다른 데이터에 대한 정규화된 참조가 포함된다.	유용	FM-I3	
R	R1	RDA-R1-01M	재사용이 가능하도록 정확하고 관련이 높은 속성이 다수 제공된다.	필수		
	R1.1	RDA-R1.1-01M	메타데이터는 데이터를 재사용할 수 있는 라이선스에 대한 정보가 포함된다.	필수	FM-R.1.1	
	R1.1	RDA-R1.1-02M	메타데이터는 표준 재사용 라이선스를 나타낸다.	중요		
	R1.1	RDA-R1.1-03M	메타데이터는 기계가 이해할 수 있는 재사용 라이선스를 의미한다.	중요		
	R1.2	RDA-R1.2-01M	메타데이터는 커뮤니티에 특화된 표준에 대한 출처 정보를 포함한다.	중요		FsF-R1.2-02.M
	R1.2	RDA-R1.2-02M	메타데이터는 커뮤니티 사이의 언어에 대한 출처 정보를 포함한다.	유용	FM-R.1.2	
	R1.3	RDA-R1.3-01M	메타데이터는 커뮤니티 표준을 따른다.	필수	FM-R1.3	FsF-R1.3-01M
	R1.3	RDA-R1.3-01D	데이터는 커뮤니티 표준을 따른다.	필수		
	R1.3	RDA-R1.3-02M	메타데이터는 기계가 이해할 수 있는 커뮤니티 표준에 따라 표현된다.	필수		
	R1.3	RDA-R1.3-02D	데이터는 기계가 이해할 수 있는 커뮤니티 표준에 따라 표현된다.	중요		FsF-R1.3-02D

5.2 지표

FAIR 데이터 성숙도 모델에 정의한 지표는 FAIR 원칙에서 파생되었다. RDA 모델은 F, A, I, R의 세부 지표를 각각 7개, 11개, 11개, 9개 포함하고 있다. FAIR 원칙의 지표는 확장하거나 수정하지 않고, 개별 원칙과 설명에 있는 내용을 바탕으로 새로운 지표를 정의하고 있다. 예를 들어, RDA 모델은 메타데이터와 데이터에 대해 별도의 지표를 정의함으로써 F1에 '(메타)데이터'로 지칭한 모호함을 제거할 수 있다. F1 원칙은 영구적이고 전역적으로 고유한 식별자를 정의하고 있다. RDA 모델은 F1을 지속성과 전역 고유성으로 구분한다. 지속성은 각각 RDA-F1-01M과

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

RDA-F1-01D, 전역 고유성은 RDA-F1-02M과 RDA-F1-02D 지표로 정의한다. 이와 같은 방식으로 FAIR 원칙을 세부적으로 평가할 수 있는 지표를 정의하고 있다.

5.3 우선순위

FAIR 데이터 성숙도 모델은 각각의 지표에 우선순위를 지정하고 있다. FAIRness를 평가하기 위한 지표는 데이터 제공자와 소비자의 관점에서 중요도가 서로 다를 수 있다. 지표의 우선순위는 세 가지 수준으로 구분하고 있다: 1) 필수(Essential): 대부분의 상황에서 FAIRness을 달성하는 데 가장 중요한 측면을 다루거나, 반대로 지표가 충족되지 않으면 FAIRness를 달성하는 것이 사실상 불가능하다, 2) 중요(Important): 특정 상황에서 가장 중요하지 않을 수 있는 측면을 다루지만 가능하다면 만족도는 실질적으로 FAIRness를 증가시킨다, 3) 유용성(Useful): 반드시 필수가 아닌 선택적 지표를 포함한다.

5.4 평가 방법

FAIR 데이터 성숙도 모델에 정의된 지표를 사용하여 데이터 자원과 해당 메타데이터를 평가할 수 있다. 워킹그룹은 해당 지표에 대해 점수를 평가하는 두 가지 방법을 제안한다. 지표별로 진행 상황을 5 단계 척도로 평가하거나, '예/아니오' 형식으로 구분하는 방법이 있다.

- 진행률 측정: 평가 대상 자원이 지표에 표현된 요구 사항을 충족하는 정도를 측정하는데 중점이 있다.
- 합격 또는 불합격 측정: 평가 중인 자원이 합격 또는 실패 척도의 지표 요구 사항을 충족하는지 여부를 결정하는 데 중점이 있다.

이 모델은 데이터 관리 계획을 수립하는 동안 데이터 자원이 생성되기 전에 자원이 달성할 것으로 예상되는 FAIRness 수준을 지정하는 데 사용될 수 있다. 또한 데이터 자원 생산 이후 자원의 FAIRness 달성 수준을 테스트하는 데 사용할 수 있다. 데이터 생산자는 이 모델을 사용하여 데이터의 FAIRness을 개선하는 방법을 결정할 수 있으며, 프로젝트 관리자, 자금 지원 기관은 이 모델을 사용하여 데이터 자원이 미리 정의된 예상 수준의 FAIRness을 달성하는지 여부를 결정할 수 있다.

6. 디지털 아카이브를 위한 고려사항

디지털 아카이빙에 대한 연구는 OAIS(Open Archival Information System) 참조모형 기반의 보전전략(이소연, 2002), 보존 메타데이터 표준과 모델(이소연, 2013; 박희진, 한상은, 오삼균, 2018), 디지털 아카이브의 시스템 현황과 구축(김성훈, 오삼균, 2018; 이규철, 2020) 등 다양한 주제의 연구가 활발하게 진행되고 있다. 기록 관리 분야에서 디지털 자원은 영구 보존의 대상이며 고유한 속성과 변화를 설명하기 위한 지적 정보와 다양한 속성을 포함한다. 전자기록은 "컴퓨터 등 전자적 처리 장치를 사용하여 생성・획득・이용・관리되는 기록이며, 디지털 형태로 존재한다는 의미에서 디지털 기록"으로 정의한다(한국기록관리학회, 2020). 기록 관리 분야에서 아카이빙은 기록의 보존을 처리하는 과정을 의미하며, 디지털 아카이빙은 "지속적인 가치를 갖고 있는 디지털 객체를

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

장기간 관리하는 활동"이다(이소연, 2002). 프로세스 관점에서 디지털 아카이빙은 정보 자원의 생산・수집・기술・보존・접근의 일련의 활동이 상호 연결되어 유기적으로 수행되는 종합적인 과정을 의미한다(박희진, 한상은, 오삼균, 2018).

그렇다면, 디지털화된 기록은 데이터라고 할 수 있을까? '디지털'의 의미는 손가락이나 발가락을 의미하는 라틴어인 디기투스(digitus)에서 유래되어, 컴퓨터의 등장이후 정보통신 기술로 처리할 수 있는 대상으로 그 의미를 확장하고 있다. 따라서 디지털 객체(digital object)는 컴퓨터 또는 소프트웨어로 읽고, 쓰고, 저장할 수 있는 물리적 형태를 의미한다. 이와 유사하게, 디지털 자원은 소프트웨어로 처리할 수 있는 정보 또는 자원으로 해석할 수 있다. 한편, 월드와이드웹은 가상의 정보 공간에서 거대한 데이터의 공간으로 개념을 확장하고 있다(Hall & Tiropanis, 2012). 팀 버너스리(Berners-Lee, 2006)는 웹에 존재하는 모든 자원은 공개 여부에 관계없이 데이터로 간주하고, 웹 표준을 기반으로 표현하고 서로 연계할 수 있다. 즉, 기록관리의 이론, 표준, 시스템을 통해 생산된 디지털화된 자원은 기록의 특수성을 포함하는 데이터의 새로운 유형으로 해석할 수 있다.

기록물에 FAIR 데이터 원칙을 수정 없이 적용하는 것은 심도 있는 논의가 필요한 주제이다. FAIR 데이터 원칙은 디지털 아카이브에 담겨 있는 기록관리의 이론적 체계와 특성을 평가하지 않는다. 현재 정의된 원칙은 기계가 디지털 자원을 읽고 이해하기(machine-actionable) 위한 목표를 갖고 있다. 따라서, 디지털화된 기록이 아닌 디지털 아카이브에 포함된 데이터를 대상으로 한정하는 것이 바람직하다. 그러나, 기록의 생애주기 관점에서 수집과 생성부터 디지털화된 특성을 고려한다면, FAIR 데이터 원칙의 포괄적 검토도 필요해 보인다.

<표 2>는 디지털 아카이브에 적용하기 위한 FAIR 데이터 평가 프레임워크를 요약하고 있다. 제안하는 프레임워크는 두 가지 측면에서 기존에 개발된 프레임워크와 차별성을 갖는다. 먼저, FAIR 원칙을 평가하기 위한 지표를 정의하고, 개별 지표를 정량적으로 평가하기 위한 세부 기준을 정의한다. 예를 들어, F 원칙의 대표적인 주제인 식별자는 식별자가 없는 상태(1점)에서 전역적이고 영구적인 식별자를 제공(5점)하는 5 단계로 구분한다. 메타데이터는 F와 A 원칙에서 공통적으로 적용된다. F 원칙의 메타데이터는 제공하는 메타데이터의 내용 중심이고, A 원칙은 데이터의 접근 여부를 판단할 수 있는 기준으로 정의한다. I와 R은 데이터 형식, 어휘, 라이선스에 대한 세부 평가 기준을 정의하고 있다. 특히, 제안하는 프레임워크는 개별 평가 지표를 평가함에 있어, 기록관리 분야에서 통용되는 표준, 어휘, 메타데이터, 프레임워크의 기준을 제공하고 있다. FAIR 원칙은 개별 원칙을 평가하는데 있어 지표를 강제하고 있지 않기 때문에, 도메인에 적합한 평가 지표를 선정하는 것이 중요하다. 그러나, 제안한 프레임워크는 디지털 아카이브에 적용하기 위해 전문가의 의견 수렴과 현장 적용성 평가를 수행하는 것이 필요하다.

〈표 2〉 디지털 아카아브를 위한 FAIR 데이터 평가 프레임워크

원칙	평가대상	점수	평가지표	설명	관련 표준
F	식별자	1	No PID	. 웹 표준(W3C), 디지털 아카이브 관련 표준 기반의 식별자 제공 . 링크드 데이터의 URI 식별 체계 적용(전역적인 식별자 생성)	doi, Handle, Archival Resource Key(ARK), ISO 9834-8 UUIDs, PURLs
		2	내부 시스템 ID(예: 데이터베이스 ID)		
		3	URL		
		4	전역 URI		
		5	전역, 영구적 PID		
	메타데이터	1	메타데이터 없음	. 디지털 기록의 포괄적 메타데이터(더블린코어, schema.org) . 디지털 기록의 수집, 생성, 기술, 출처에 대한 메타데이터 . Readme 파일, 버전관리 파일 . 검색엔진이 디지털 아카이브의 메인페이지를 통해 데이터 형식을 처리할 수 있는 메타데이터(schema: Collection 등)	RDA, AMREMM, BIBFRAME, MARC 21, MODS, VRA Core, DCMI, Schema.org
		2	메타데이터가 충분하지 않음		
		3	제한된 메타데이터 제공		
		4	충분한 메타데이터 제공		
		5	확장된 메타데이터 제공과 부가적인 문서 제공		

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

원칙	평가대상	점수	평가지표	설명	관련 표준
A	메타데이터	1	데이터 또는 메타데이터에 접근 불가	. 개방형 프로토콜을 통해 디지털 아카이브에 접근(예: HTTP, HTTPS, FTP, TFTP, SFTP) . 명시적으로 데이터 접근을 기술한 문서 제공	OAI-PMH, Search/Retrieve via URL(SRU)
		2	메타데이터에 한정해 접근		
		3	제한된 방식으로 접근(상용목적의 접근, 엠바고 정책 적용)		
		4	일반적인 접근(public access) - 등록 필요		
		5	제한없는 전면 접근		
I	데이터형식	1	독점적인 형식	. 특정 소프트웨어에 제한되지 않는 개방형 파일 형식 . 기계가 읽고 이해할 수 있는(machine-actionable) 메타데이터 형식으로 제공 . 디지털 자원의 내용협상(content negotiation) 제공 . SPARQL 엔드포인트를 통해 디지털 자원에 접근, 검색 기능 제공	X3D, FLAC, MP3, DPX, MPEG 4, TIFF, Jpeg2000(JP2), ASCII Text, CSV
		2	높은 비율의 독점적인 형식 제공		
		3	개방형 포맷		
		4	개방형, 구조적인 형식		
		5	개방형, 구조적, 기계가 처리할 수 있는 형식		
	어휘	1	표준어휘를 사용하지 않음	. 디지털 아카이브와 관련된 통제 어휘, 텍소노미, 온톨로지의 적용 . 기록관리 표준(국내) . 기록관리 메타데이터 표준: EAD, ISAD(G), OAIS, PREMIS, AGRkMS . 온톨로지 어휘: RiC-O, Schema.org	. 아카이브 메타데이터 구조: ISAD(G), RAD . 메타데이터 구조: ISO 15836, ISO 23081 . XML DTD/스키마: EAD, EAG, MODS,RDF . 디지털 보존: ISO/TR 18492, PREMIS . 레퍼런스모델/프레임워크: ISO 15489, ISO 14721, PAIMAS(ISO 20652), BSI PAS 197
		2	자체적으로 정의한 어휘 사용		
		3	제한적으로 표준 어휘 사용		
		4	표준어휘 사용		
		5	의미 기반 표준 어휘 사용		
R	라이선스	1	라이선스 없음	. 데이터 라이선스에 대한 설명, 링크 정보 제공 . 기계가 처리할 수 있는 형식의 라이선스 정보 제공	Creative Commons(CC0, CC BY, CC BY NC, CC BY NC ND), BSD, MIT, GNU GPL, Open Database license
		2	이용 약관(Terms of Use)		
		3	라이선스가 있으나 구체적인 정보를 제공하지 않음		
		4	라이선스 있음(기계가 읽을 수 있는 방식 X)		
		5	기계가 읽을 수 있는 라이선스		

7. 결 론

본 연구는 FAIR 데이터의 개념과 원칙을 소개하고, FAIR 데이터 원칙을 디지털 아카이브에 적용하기 위한 고려사항과 프레임워크를 제안했다. FAIR 데이터 원칙은 데이터를 찾을 수 있고(Findable), 접근 가능하며(Accessible), 상호 운용성이 확보되고(Interoperable), 재사용하기(Reusable) 위한 일련의 지침이다. FAIR 원칙은 기계가 읽을 수 있는 개방형 환경을 염두해 두고 설계되었으나, 궁극적으로 기계와 사람이 데이터를 효과적으로 탐색하고, 재사용하는 것을 목표로 갖고 있다. 디지털 아카이브의 관점에서 FAIR 데이터 원칙은 디지털 자원의 공유와 재사용을 위한 핵심 요소인 메타데이터 표준, 데이터 형식, 데이터의 상호운용성을 위한 프레임워크로 활용할 수 있다. FAIR 원칙을 만족하는 디지털 자원은 검색, 인용, 재사용이 가능하고, 커뮤니티에서 합의된 메타데이터 적용과 관리 계획을 통해 디지털 아카이브의 지속적 가치를 유지할 수 있다.

http://dx.doi.org/10.14404/JKSARM.2021.21.2.155

국내에서 FAIR 데이터 원칙은 심도 있게 논의되지 않은 것이 현실이다. 연구 데이터 분야에서 FAIR 데이터 원칙을 권장하고 있지만, 개별 원칙을 적용하기 위한 세부 지표가 모호하고 지표를 평가하기 위한 프레임워크도 본격적인 연구가 미흡하다. 본 연구는 FAIR 데이터 원칙의 전문을 소개하고, 평가 프레임워크의 특징을 기술하고 있다. 더불어 디지털 아카이브 관점에서 FAIR 데이터 원칙을 적용하기 위한 프레임워크를 제안했다. 그러나, 기록관리의 이론 체계, 디지털 자원의 고유한 특성으로 고려하기 위한 논의가 필요하고, 프레임워크의 현장 적용 가능성을 검토할 수 있는 후속 연구도 필요하다. 더불어, FAIR 데이터 원칙이 디지털 아카이브에 적용하는 것이 적합한지, 기록관리와 보존의 프레임워크와 표준 관점에서의 관계와 같은 원론적인 논의도 본격적으로 시작할 필요가 있다. 향후 연구는 기록관리학과 FAIR 데이터 원칙의 이론과 구현 가능성을 검토하고, 국내에서 핵심적인 역할을 하는 주요 아카이브를 대상으로 제안한 프레임워크를 적용할 계획이다.

참고문헌

국가과학기술연구회 (2019). 연구데이터 관리 가이드라인. 국가과학기술연구회

김성욱, 김선태 (2020). 응집물질물리분야 연구데이터 관리 방안 연구. 정보관리학회지, 37(3), 77-106. http://dx.doi.org/10.3743/KOSIM.2020.37.3.077

김성훈, 오삼균 (2018). 연구데이터 관리서비스의 구현 시 고려사항에 관한 연구. 정보관리학회지, 35(2), 141-165. http://dx.doi.org/10.3743/KOSIM.2018.35.2.141

김유승 (2010). 아카이브 2.0 구축을 위한 이론적 고찰. 한국기록관리학회지, 10(2), 31-52. http://dx.doi.org/10.14404/JKSARM.2010.10.2.031

김학래 (2017). 지식그래프. 서울: 커뮤니케이션북스. https://doi.org/10.979.11288/05141

박옥남 (2012). PREMIS 기반 보존 메타데이터 요소 개발에 관한 연구: 국립중앙도서관 디지털 자료를 중심으로. 한국문헌정보학회지, 46(2), 83-113. http://dx.doi.org/10.4275/KSLIS.2012.46.2.083

박희진, 한상은, 오삼균 (2018). E-ARK 기반 디지털 아카이빙 모델에 관한 연구. 정보관리학회지, 35(1), 83-101. http://dx.doi.org/10.3743/KOSIM.2018.35.1.083

신재민, 곽승진 (2013). 디지털 콘텐츠 아카이빙 정책수립을 위한 문헌 및 사례 고찰. 사회과학연구, 24(1), 305-330.

이규철 (2020). 디지털 아카이브의 발전과 구축 동향. 건축, 64(5), 35-38.

이소연 (2002). 디지털 아카이빙의 표준화와 OAIS 참조모형. 정보관리연구, 33(3), 45-68. http://dx.doi.org/10.1633/JIM.2002.33.3.045

이소연 (2013). 국내 디지털 보존 연구의 동향 분석. 한국기록관리학회지, 13(2), 247-283. http://dx.doi.org/10.14404/JKSARM.2013.13.2.247

최명석, 이승복, 이상환 (2017). 국내 과학기술분야 연구기관의 과학데이터 관리 현황. 한국콘텐츠학회논문지, 17(12), 117-126. http://dx.doi.org/10.5392/JKCA.2017.17.12.117

한국기록관리학회 (2020). 기록관리의 이론과 실제. 조은글터

AFAWG (2017). Policy statement on F.A.I.R. access to Austrailia's research outputs. Retrieved April 19, 2021, Available: https://www.fair-access.net.au/fair-statement

ARDC (2021). SATIFYD: Self-Assessment Tool to Improve the FAIRness of Your Dataset. Retrieved May 12, 2021, Available: https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/

Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., Russell, K., & Stall, S. (2020). The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. Data Science Journal, 19(1), 41. http://doi.org/10.5334/dsj-2020-041

Barbuti, N. (2020). Thinking digital libraries for preservation as digital cultural heritage: by R to R4 facet of FAIR principles.

International Journal on Digital Libraries, 1-10. https://doi.org/10.1007/s00799-020-00291-7

Berners-Lee, T. (2006). Design Issues: Linked Data. Retrieved April 19, 2021, Available: https://www.w3.org/DesignIssues/LinkedData.html

Boeckhout, M., Zielhuis, G. A., & Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: fair enough?. European journal of human genetics: EJHG, 26(7), 931-936. https://doi.org/10.1038/s41431-018-0160-0

Calamai, S. & Frontini, F. (2018). FAIR data principles and their application to speech and oral archives. Journal of New Music Research, 47, 339-354.

Candela, G., Sáez, M. D., Escobar Esteban, Mp., & Marco-Such, M. (2020). Reusing digital collections from GLAM institutions. Journal of Information Science. https://doi.org/10.1177%2F0165551520950246

Collins, S., Genova, F., Harrower, N., Hodson, S., Jones, S., Loaksonen, L., Mietchen, D., Petrauskaite, R., & Wittenburg, P. (2018). Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data. http://dx.doi.org/10.2777/1524

Corpas M, Kovalevskaya NV, McMurray A, & Nielsen FGG (2018). A FAIR guide for data providers to maximise sharing of human genomic data. PLoS Comput Biol, 14(3), e1005873. https://doi.org/10.1371/journal.pcbi.1005873

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., Murphy, F., Polischuk, P., Taylor, S., Martone, M., & Clark, T. (2018). A data citation roadmap for scientific publishers. Scientific Data, 5. https://doi.org/10.1038/sdata.2018.259

DANS (2021). FAIR self assessment tool. Retrieved May 12, 2021, Available: https://satifyd.dans.knaw.nl/

David, R., Mabile, L., Specht, A., Stryeck, S., Thomsen, M., Yahia, M., Jonquet, C., Dollé, L., Jacob, D., Bailo, D., Bravo, E., Gachet, S., Gunderman, H., Hollebecq, J. E., Ioannidis, V., Bras, Y. L, Lerigoleur, E., & Cambon-Thomsen, A. (2020). FAIRness Literacy: The Achilles' Heel of Applying FAIR Principles. Data Science Journal, 19(1), 32. https://doi.org/10.5334/dsj-2020-032

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., Vries, J., L'Hours, H., Davidson, J., & Whyte, A. (2020). FAIRsFAIR Data Object Assessment Metrics (Version 0.4). https://doi.org/10.5281/zenodo.4081213

Devaraju, A., Mokrane, M., Cepinskas, L., Huber, R., Herterich, P., de Vries, J., Akerman, V., L'Hours, H., Davidson, J., & Diepenbroek, M. (2021). From Conceptualization to Implementation: FAIR Assessment of Research Data Objects. Data Science Journal, 20(1), 4. http://doi.org/10.5334/dsj-2021-004

DTL (2021). European Commission embraces the FAIR principles. Retrieved April 19, 2021, Available: https://www.dtls.nl/2016/04/20/european-commission-allocates-e2-billion-to-make-research-data-fair/

EU (2016). G20 Leaders' Communique Hangzhou Summit. Retrieved April 19, 2021, Available: https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967

FAIRMetrics (2021). FAIR Maturity Indicators and Tools. Retrieved May 12, 2021, Available: https://github.com/FAIRMetrics/Metrics

FORCE11 (2016). The FAIR Data Principles. Retrieved April 19, 2021, Available: https://www.force11.org/group/fairgroup/fairprinciples

GO FAIR (2021). FAIRification process. Retrieved April 19, 2021, Available: https://www.go-fair.org/fair-principles/fairification-process/

Government of the Netherlands (2017). Progress towards the European Open Science Cloud. Retrieved April 19, 2021, Available: https://www.government.nl/latest/news/2017/12/01/progress-towards-the-european-open-science-cloud

Guizzardi, G. (2020). Ontology, Ontologies and the "I" of FAIR. Data Intelligence, 2, 181-191. https://doi.org/10.1162/dint_a_00040

Hall, W. & Tiropanis, T. (2012). Web evolution and Web Science. Computer Networks, 56, 3859-3865.

Haux, C. & Knaup, P. (2019). Using FAIR Metadata for Secondary Use of Administrative Claims Data. Studies in health

technology and informatics, 264, 1472-1473. https://doi.org/10.3233/shti190490

Helliwell, J. R., Minor, W., Weiss, M. S., Garman, E. F., Read, R. J., Newman, J., Raaij, M. J., Hajdu, J., & Baker, E. N. (2019). Findable Accessible Interoperable Re-usable(FAIR) diffraction data are coming to protein crystallography. Journal of applied crystallography, 52(Pt 3), 495-497. https://doi.org/10.1107/S2052252519005918

Hettne, K. M., Verhaar, P., Schultes, E., & Sesink, L. (2020). From FAIR Leading Practices to FAIR Implementation and Back: An Inclusive Approach to FAIR at Leiden University Libraries. Data Science Journal, 19(1), 40. http://doi.org/10.5334/dsj-2020-040

Jacobsen, A., Kaliyaperumal, R., Bonino, S., Mons, B., Schultes, E., Roos, M., & Thompson, M. (2019). A Generic Workflow for the Data FAIRification Process. Data Intelligence. 2(1-2). 56-65. https://doi.org/10.1162/dint_a_00028

Koster, L. & Woutersen-Windhouwer, S. (2018). FAIR Principles for Library, Archive and Museum Collections: A proposal for standards for reusable collections. Code4Lib Journal, 40. http://journal.code4lib.org/articles/13427

Mons, B., Neylon, C., Velterop, J., Dumontier, M., Da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; Revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services and Use, 37(1), 49-56. https://doi.org/10.3233/ISU-170824

Nitecki, D. A. & Alter, A. (2021). Leading FAIR Adoption Across the Institution: A Collaboration Between an Academic Library and a Technology Provider. Data Science Journal, 20(1), 6. http://doi.org/10.5334/dsj-2021-006

Niu, J. (2016). Linked Data for Archives. Archivaria, 82, 83-110. https://www.muse.jhu.edu/article/687083.

Sinaci, A. A., Núñez-Benjumea, F. J., & Gencturk, M., et al. (2020). From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods of information in medicine, 59(S 01), e21-e32. doi:10.1055/s-0040-1713684

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E., & Wyborn, L. (2019). Make scientific data FAIR. Nature, 570(7759), 27-29. https://doi.org/10.1038/d41586-019-01720-7

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Dumontier, M., Sansone, S. A., Prieto, M., Batisata, D., McQuilton, P., Kuhn, T., Rocca-Serra, P., Crosas, M., & Schultes, E. (2019). Evaluating FAIR maturity through a scalable, automated, community-governed framework. Scientific Data, 6(174), 1-12. https://doi.org/10.1038/s41597-019-0184-5

• **국문 참고자료의 영어 표기**
(English translation / romanization of references originally written in Korean)

Choi, M., Lee, S., & Lee, S. (2017). Research Data Management of Science and Technology Research Institutes in Korea. The Journal of the Korea Contents Association, 17(12), 117-126. https://doi.org/10.5392/JKCA.2017.17.12.117

Kim, Haklae (2017). Knowledge Graph. Seoul: CommunicationBooks, Inc. https://doi.org/10.979.11288/05141

Kim, S. & Kim, S. (2020). A Study on the Research Data Management Methods for the Condensed Matter Physics. Journal of the Korean society for information management, 37(3), 77-106. https://doi.org/10.3743/KOSIM.2020.37.3.077

Kim, S. & Oh, S. G. (2018). Key Factors in the Implementation of Research Data Management Services. Journal of the Korean society for information management, 35(2), 141-165. https://doi.org/10.3743/KOSIM.2018.35.2.141

Kim, You-Seung (2010). A Theoretical Study on Establishing Archive 2.0. Journal of Korean Society of Archives and Records Management, 10(2), 31-52. https://doi.org/10.14404/JKSARM.2010.10.2.031

Korean Society of Archives and Records Management (2020). Records and Archives Management: Theory and Practice. Goodwriting Publishing.

Lee, Geauchul (2020). Development and Current Trend of Digital Archive. Review of Architecture and Building Science, 64(5), 35-38.

Lee, S. (2002). Standardization of Digital Archiving and OAIS Reference Model. Journal of Information Science Theory and Practice, 33(3), 45-68. http://dx.doi.org/10.1633/JIM.2002.33.3.045

Lee, S. (2013). Trends Analysis of Digital Preservation Research in Korea. Journal of Korean Society of Archives and Records Management, 13(2), 247-283. https://doi.org/10.14404/JKSARM.2013.13.2.247

National Research Council of Science & Technology (2019). Guidelines of Research Data Management. National Research Council of Science & Technology.

Park, H., Han, S., & Oh, S.-G. (2018). A Study of a Digital Archiving Model based on E-ARK. Journal of the Korean Society for Information Management, 35(1), 83-101. https://doi.org/10.3743/KOSIM.2018.35.1.083

Park, O. (2012). A Study on Developing Preservation Metadata Based on PREMIS Focusing on Digital Data in National Library of Korea, 46(2), 83-113. https://doi.org/10.4275/KSLIS.2012.46.2.083

Shin, J. & Kwak, S. (2013). A Review of Literature and Cases for Developing Digital Content Archives. Journal of Social Science, 24(1), 305-330.