FAIR 원칙 세부 평가 기준

본 내용은 FAIR Data Self-Assessment Tool[1]을 디지털 아카이브 평가에 적용하기 위해 개별 지표를 해석한 결과임

Findability

F1: Does the dataset have any identifiers assigned?

- 개별 디지털 객체가 고유한 식별자가 있는지에 대한 항목
- ARDC 기준: 디지털 객체는 검색되고 연결, 인용하기 위해 고유한 식별자가 필요하며, 개별 객체의 메타데이터 레코드와 관련된 파일이나 메타데이터에 명확하게 명시되어야 한다고 설명함. 이때 웹 주소(URL)는 객체의 온라인 위치를 지정할 수 있으나, 변경될 수 있다는 점에서 객체의 저장 위치에 관계 없이 고정된 참조 정보인 DOI(Digital Object Identifier), PID(Persistent Identifier)에 비해 영구성이 떨어진다고 설명함. 이러한 식별자는 해당 아카이브를 서비스하는 기관이나 레포지토리를 관리하는 측에서 제공함.
- 연구자 기준: 순서대로 평가 점수가 높으며, 중복으로 해당되는 경우 더 높은 평가 점수에 해당하는 선지를 선택함

  (1) Globally unique, and persistent (e.g. DOI, PURL, ARK or Handle)

  ARDC의 설명과 같이, 전역적으로 쓰이는 고유한 식별자이며 URL과 같이 변경될 가능성이 높은 식별자는 제외함. 이때 메타 태그(meta tag)에 URI가 있으나, 해당 사이트의 URL등 실제로 개별 객체를 식별할 수 있는 고유 값이 기재되어 있지 않은 경우는 제외함

  (2) Web Address (URL)

  메타데이터 항목에 '식별자'라는 항목이 없더라도 개별 디지털 객체별로 고유한 웹 페이지가 존재하며, 해당 페이지의 URL이 존재하는 경우가 해당함. 이때 객체의 개별 메타데이터가 반드시 하나의 페이지에 존재할 필요는 없으나(한국학 아카이브는 메타데이터가 클릭하면 팝업으로 뜸) 해당 객체와의 연결점이 반드시 필요함

  (3) Local identifier

  디지털 객체가 개별 웹페이지로 분리되지 않고, 목록과 같은 형식으로 제공되지만 각 객체를 구분할 수 있는 식별자가 메타데이터 상에 존재하는 경우가 해당함. 이때의 식별자는 해당 디지털 아카이브 내부에서만 사용되는 자체 식별자도 가능하며, 순번과 같은 형태도 포함됨

  (4) No identifier

  개별 디지털 객체를 구분할 수 있는 어떠한 식별자도 존재하지 않는 경우

[1] https://ardc.edu.au/resource/fair-data-self-assessment-tool/

F2: Is the dataset identifier included in all metadata records/files describing the data?

- 각 디지털 객체의 메타데이터에 식별자가 포함되는지에 대한 항목.
- ARDC 기준: 기준을 명시하지 않음
- 연구자 기준:

  (1) Yes

  메타데이터에 객체의 고유한 식별자가 포함되어 있어야 함. 이때 식별자의 형태는 '식별자', '순번', '고유번호', 'finding aid ID', '참조코드' 등 다양할 수 있으며, 식별자의 형태나 제공 방식보다는 모든 메타데이터에 일관되게 포함되어 있는지를 중점으로 확인함. 일부 디지털 객체의 메타데이터 항목에만 존재하는 경우는 제외함

  (2) No

  웹 주소가 존재하더라도 메타데이터 항목에 '식별자', '순번', '고유번호' 등과 같은 항목이 존재하지 않는다면 'No'에 해당

F3: How is the data described with metadata?

- 디지털 객체의 메타데이터의 설명이 자세한/풍부한 정도에 대한 항목
- ARDC 기준: 포괄적, 종합적(comprehensive)인 메타데이터에 포함되어야 하는 최소한의 기준을 제시하며, 'one size fits all'은 없다고 설명함. 개별 디지털 아카이브의 특성에 따라 메타데이터는 상이할 수 있다는 의미로 해석함.
- 포괄적인 메타데이터에 포함되어야 하는 예시는 다음과 같음: '전역적으로 고유한 영구 식별자', '제목', '데이터 생성자, 관리자', '데이터 혹은 파일 형식의 액세스 방법', '데이터 생성 방법', '데이터에 대한 설명', '데이터 인용 방법에 대한 정보', '기계가 읽을 수 있는 데이터 라이센스', '출처와 문맥 정보', '공간적, 시간적 범위(해당되는 경우만)'
  이때 말하는 문맥 정보(contextual information)은 해당 객체와 관련된 출판물, 프로젝트, 서비스, 소프트웨어에 대한 정보를 의미하며, 데이터가 생산된 방법론과 프로세스와 관련된 정보를 포함함
- 연구자 기준:
  개별 디지털 아카이브의 특성에 따라 메타데이터는 상이할 수 있음에 따라 메타데이터 항목명의 다양성을 고려함. 만약, 메타데이터를 여러 형식으로 제공하는 경우 메타데이터 요소가 많은 형식을 기준으로 평가함

  (1) Comprehensively using a formal machine-readable metadata schema

  메타데이터를 기계가 읽을 수 있는 형식으로 제공하고, ARDC의 기준에 따라 충분히 포괄적이어야함. 디지털 아카이브의 수준이 아닌, 디지털 객체(아이템) 수준으로 상세한 메타데이터가 제공되어 야함.

  (2) Comprehensively, but in a text-based, non-standard format

  웹 페이지에 구조화되지 않은 형태와 텍스트로만 메타데이터를 제공하는 경우임

  표준 메타데이터 어휘(Dublin Core, Europeana Data Model 등)을 사용하더라도 구조화된 형식(RDF, XML 등)으로 메타데이터를 제공하지 않는 경우도 해당 선지에 해당함

  (3) Brief title and description

제목과 설명만 제공하는 경우

(4) The data is not described

디지털 객체(아이템)에 대한 어떠한 설명글이나 부가 정보가 제공되지 않음

F4: What type of repository or registry is the metadata record in?

- 디지털 객체의 메타데이터가 저장되어 있는 형태(type)에 대한 항목
- ARDC 기준: 디지털 객체를 인터넷에서 검색하여 접근하기 위해서는 풍부한 메타데이터가 존재할 뿐만 아니라, 레포지토리에 등록, 색인화 되어 있어야 함. 이러한 repository와 registry는 구글이나 구글 스칼라와 같은 검색 엔진에서 색인화 되는 것이 바람직함
- 연구자 기준:

  개별 디지털 객체 자체가 아닌 객체의 '메타데이터'가 저장되어 있는 저장소를 기준으로 판단함. 이때 repository는 데이터가 저장되는 물리적인 공간이며, registry는 메타데이터 항목만 저장되는 일종의 목록적 형태로 정의함

  (1) Data is in one place but discoverable through several registries

  Google과 같은 검색엔진에 색인되어 개별 아이템 단위의 검색이 가능한 경우에 해당함

  (2) Generalist public repository

  GitHub, Zenode와 같이 일반적인 데이터 레포지토리에 메타데이터를 등록한 경우에 해당함. 한국의 경우, 공공데이터포털(data.go.kr)이 특정 도메인과 관계없이 모든 공공데이터를 제공함

  (3) Domain-specific repository

  문화공공데이터포털, 문화예술포털 등 특정 분야의 데이터를 관리하는 레포지토리/레지스트리에 등록한 경우에 해당함

  (4) Local institutional repository

  기관 단위의 레포지토리/레지스트리를 운영하는 경우에 해당함. 예를 들어, 모기관이 통합 레포지토리를 운영하고 산하기관 또는 유관기관이 운영하는 디지털 아카이브의 메타데이터를 통합 레포지토리에 등록하는 경우에 해당함

  (5) The data is not described in any repository

  별도의 레포지토리/레지스트리의 등록하지 않고 기관 내부 시스템에서만 관리하는 경우에 해당함

Accessible

A1: How accessible is the data?

- 데이터에 접근할 수 있는 방법에 대한 항목
- ARDC 기준: 데이터가 Findable하다고 해서 모두 자유롭게 접근할 수 있는 것은 아니며, 개인 정보 보호 및 상업적 이익과 같은 문제로 데이터에 대한 엠바고, 접근 제한 등의 상황이 발생할 수

있음. (연구자 해석으로 넘어가는 내용) 민감한 데이터의 경우에도 기관에 직접 문의, 데이터 신청 등과 같은 방법 제시를 통해 데이터에 접근할 수 있음

- 연구자 기준: 이때의 data는 디지털 객체의 '원문'에 해당하는데, 디지털 아카이브의 특성에 따라 원문의 개념이 다양할 수 있음. 예를 들어, 공연예술 분야는 해당 공연의 영상을 원문으로 인식할 수 있고, 건축 분야는 건축물의 사진을 원문이라고 할 수 있음. 따라서, 원문을 다운로드하거나 이용할 수 있는 것과 별개로 접근의 제한을 기준으로 평가함

  (1) Publicly accessible

  어떤 제약도 없이 공개적으로 접근이 가능한 경우에 해당함

  (2) Fully accessible to persons who meet explicitly states conditions, e.g. ethics approal for sensitive data

  윤리적으로 민감한 데이터 또는 별도의 사유로 인해 특정 조건을 만족하는 사용자만 접근이 가능한 경우임

  예를 들어, 접근하기 위한 조건과 절차를 명시적으로 안내하는 경우에 해당함. 로그인 과정 또는 별도의 폼을 통해 요청하는 경우도 포함됨

  접근 신청을 요청했더라도 조건이 모호하여 데이터에 접근이 가능한지 불명확한 경우는 해당하지 않음

  (3) A de-identified/modified subset of the data is publicly accessible

  원문 데이터를 비식별화 하거나 별도의 수정을 거친 데이터만 공개적으로 접근이 가능한 경우에 해당함

  (4) Embargoed access after a specified date

  엠바고(embargo)와 같이 제한된 기간 이후에 접근이 가능한 경우에 해당함

  (5) Unspecified conditional access e.g. contacts the data custodian for access

  명시적이지 않은 조건에 의해 접근이 제한된 경우에 해당함. (3) 선지와 차이점은 접근의 기준이 명시적이지 않다는 것과 사용자의 신청에도 접근이 불가능한 경우가 있음

  (6) Access to metadata only

  디지털 아카이브에서 제공하는 콘텐츠의 대부분이 메타데이터만 접근할 수 있는 경우에 해당함

  (7) No access to data or metadata

  디지털 아카이브의 데이터와 메타데이터에 접근할 수 없는 경우에 해당함

A2: Is the data available online without requiring specialised protocols or tools once access has been approved?

- 데이터에 접근하는 방법에 있어서 특정 규약이나 승인된 도구를 통해야 하는 지에 대한 항목
- ARDC 기준: 데이터에 접근하는 이상적인 방법은 인터넷에서 데이터를 찾은 뒤, HTTP, FTP와 같은 인터넷 프로토콜을 통해서 직접 파일이나 정보를 요청하고 사용하는 것임. API(Application

Programming Interface)는 정보가 기계가 읽을 수 있는 형식으로 제공되면, 웹을 통해 해당 정보를 요청하는 프로그램에서 해당 정보를 직접 사용할 수 있게 되며, 정보를 다른 기계에서 직접 사용할 수 있게 하는 방법임. 가장 이상적인 방법인 방법으로 예시를 든 OGC(Open Geospatial Consortium)의 API 서비스는 데이터 공유를 위한 개방형 표준에 해당함

- 연구자 기준: 메타데이터 뿐만 아니라 원본 데이터에 접근할 수 있는지를 기준으로 함. 이때, 접근은 원본 데이터에 접근하는 간접경로를 제공하는 경우도 포함됨

  (1) Standard web service API (e.g. OGC)

  데이터를 제공하는 프로토콜과 API를 표준 명세에 웹 표준 명세에 따라 설계한 경우에 해당함

  웹 표준의 예시는 UDDI(Universal Description, Discovery and Integration), WSDL(Web Services Description Language) 등이 해당함. API의 경우, SOAP 형식은 고도로 구조화되어 웹 표준 규정을 준수하는 형태임

  (2) Non-standard web service (e.g. OpenAPI/Swagger/informal API)

  오픈API, Swagger API와 같이 유연한 구조인 API를 사용하여 데이터를 제공하는 경우임.예를 들어, REST API 유형은 XML만 지원하는 SOAP API 유형과 달리 다양한 형식을 지원함

  (3) File download from online location

  웹상에서 디지털 객체를 다운로드 받을 수 있는 경우에 해당함. 단, 공식적으로 기능을 제공하는 경우에 한하며, 사용자가 별도의 방법(웹크롤링 등)으로 수집하는 것은 해당하지 않음

  (4) By individual arrangement

  사용자가 개별적으로 문의하여 데이터를 요청하거나 수집하는 경우에 해당함. 특정 뷰어나 웹상에서 열람만 가능한 경우도 해당 선지임

  (5) No access to data

  웹상에서 데이터를 열람할 수 없는 경우에 해당함

A3: Will the metadata record be available even if the data is no longer available?

- 인터넷 상에서 디지털 아카이브의 객체가 더이상 존재하지 않아도 해당 객체의 메타데이터에는 접근할 수 있는 지에 대한 항목
- ARDC 기준: 온라인 상에서 장기간 데이터를 유지하기 위해서는 많은 노력이 필요하며, 일부 경우 데이터와 메타데이터의 연결이 끊길 수 있음. 데이터에 대한 정보가 손실된 경우에도 메타데이터에 대한 최소한의 설명이 존재하는 경우, 데이터를 추적할 수 있는 가능성이 생기므로, 데이터 접근성 관점에서 데이터의 존재를 확인할 수 있는 정보가 남아야 있어야 함을 설명함
- 연구자 기준:

  (1) Yes

  디지털 아카이브는 데이터 손실 시에도 메타데이터 접근이 가능함을 명시적으로 안내해야 함.접근이 가능하다는 것은 웹 인터페이스, API, 별도의 메타데이터 저장소 등을 통해 메타데이터에 접근하는 것을 의미

- 예시1) 디지털 아카이브의 '데이터 관리 정책' 페이지에 "모든 메타데이터는 관련 데이터의 가용성과 관계없이 영구적으로 보존됩니다"라고 명시되어 있는 경우
- 예시2) 아카이브의 API 문서에 "데이터 삭제 시에도 메타데이터 엔드포인트는 계속 유지됩니다"라는 설명이 있는 경우

(2) No

디지털 아카이브가 데이터 이용 불가 시 메타데이터도 접근 불가능함을 명시적으로 밝히는 경우

(3) Unsure

데이터가 없어져도 메타데이터에 접근할 수 있는지에 대한 명시적인 정책이나 정보가 없는 경우

Interoperable

I1: What (file) format(s) is the data available in?

- 데이터의 포맷(형식)에 대한 항목
- ARDC 기준: 아카이브에 적합한 데이터 형식은 비 독점적이고, 암호화되지 않고, Uncompressed(?)하며, 연구 커뮤니티에서 일반적으로 사용되는 것이여야 하고, 다양한 플랫폼과 애플리케이션 간 상호 운용이 가능해야 함. 또한 로열티없이 사용할 수 있는 등 지적 재산권의 제한 없이 다양한 플랫폼에서 독립적으로 구현할 수 있어야 하고 개방형 표준 기관에서 개발 및 유지 관리를 지속적으로 해야 함. 각 선지에 해당하는 포맷의 예시는 다음과 같음

- Structured, open standard, machine-readable format e.g. (text) PDF/A, HTML, Plain text, (images) TIFF, JPEG 2000, GIF, (audio) MP3, AIFF, WAVE, (video) MOV, MPEG, AVI, (Tabular data) CSV
- Structured, open standard, non-machine-readable format, e.g. PDF, HTML, JPG
- Proprietary format, e.g. doc (Word), .xls (Excel), .ppt (PowerPoint), .sav

- 연구자 기준: ARDC에서 제시한 포맷 예시를 기준으로 하며, 예시에 포함되지 않은 형식(hwp 등)이 존재할 경우 특성을 파악하여 포함함

(1) In a structured, open standard, machine-readable format

ARDC의 설명에 해당하는 형식인지 파악하여 분류함. 단, HWPX와 XLSX는 개방형 표준이 아니므로 해당하지 않음

(2) In a structured, open standard, non-machine-readable format

ARDC의 설명에 해당하는 형식인지 파악하여 분류함. PDF/A가 아닌 PDF는 구조화되지 않은 형식으로 (2)에 해당함.

(3) Mostly in a proprietary format

ARDC의 설명에 해당하는 형식인지 파악하여 분류함. 특정 소프트웨어를 통해서만 열람할 수 있는 형식에 해당한다. 예를 들어, 디지털 아카이브에서 스캔한 이미지를 제공할 때 특정 뷰어를 설치가 필수적인 경우도 해당함

I2: What best describes the types of vocabularies/ontologies/tagging schemas used to define the data elements?

- 디지털 객체의 메타데이터를 표현할 때 어휘나, 온톨로지, 태깅 스키마 등을 사용했다면, 어느 수준으로 사용했는지에 대한 항목
- ARDC 기준: 표준 스키마는 ISO, DCMI와 같은 표준 기관에서 공식적으로 검증을 마친 것을 의미하며, 일반적으로 사용되는 메타데이터 스키마는 잘 문서화되어 유지 관리 되고 있음. 표준화된 개방 범용 스키마의 예시로는 DataCite Metadata Schema , PROV , Dublin Core 등을 제시하고, 특정 도메인 표준 스키마의 예시로는 HPO(Human Phenotype Ontology), MeSH(Medical Subject Headings), Marine Community Profile, DDI(Data Documentation Initiative) 등을 제시함. 또한 선지의 global identifiers는 F1의 식별자와 같은 맥락으로 이해할 수 있음
- 연구자 기준:

  (1) Standardised open and universal resolvable global identifiers linking to explanations

  웹 표준 어휘 또는 기록관리 관련 표준 어휘를 사용하고, 개별 기록물의 메타데이터를 RDF, XML 등 기계가 읽을 수 있는 형식으로 제공하는 경우에 해당함

  (2) Stanarsied vocabularies/ontologies/schema without global identifiers

  표준 어휘를 사용하여 관리하지만, 기계가 읽을 수 없는 형식으로 제공하는 경우에 해당함.

  (3) No standards have been applied in the description of data elements

  자체적인 기술 지침이 있더라도, 디지털 아카이브가 명시적으로 표준 어휘를 사용했음을 기재하지 않았다면, 표준을 적용하지 않은 것으로 평가함

  (4) Data elements not described

  기록물의 메타데이터를 제공하지 않는 경우에 해당함

I3: How is the metadata linked to other data and metadata (to enhance context and clearly indicate relationships)?

- 메타데이터가 다른 메타데이터, 데이터와 연계되는 방법에 대한 항목
- ARDC 기준: 메타데이터 간의 가능한 많은 연결이 있을 때, 데이터에 대한 맥락적 지식(contextual knowledge)이 풍부해질 수 있다고 설명함.
- 연구자 기준:

  (1) Metadata is represented in a machine-readable format, e.g. in a linked data format such as Resource Description Framework (RDF).

  메타데이터가 RDF 형식과 같이 기계가 읽을 수 있는 형식으로 제공되고, 다른 자원에 대한 URI 링크가 포함된 경우에 해당함

  (2) The metadata record includes URI links to related metadata, data and definitions

  기계가 읽을 수 있는 형식으로 메타데이터를 제공하지 않으나, 다른 자원을 탐색할 수 있는 URI를 제공하는 경우에 해당함. 예를 들어, 웹 페이지의 메타데이터 요소 중 분류체계, 키워드,

인물정보와 같이 다른 개체(Entity)로 재검색이 가능하도록 하이퍼링크를 제공하거나, 해당 기록물과 관련된 연관 데이터를 추천 또는 제공하는 경우에 해당함

(3) There are no links to other metadata

메타데이터 요소가 모두 텍스트로만 제공되고, 연관 정보를 제공하지 않는 경우에 해당함

Reusable

R1: Which of the following best describes the license/usage rights attached to the data?

- 디지털 객체의 라이선스가 기재되어 있는 방법에 대한 항목
- ARDC 기준: 호주의 경우 라이선스가 없는 것은 'all right reserved'와 동일하게 간주되어 재사용이 매우 제한됨. 이와 같이 데이터의 라이선스는 재사용을 위해 중요하며, 반드시 기재되어야 함. Creative Commons 라이선스는 데이터의 재사용을 명시할 수 있는 가장 간단한 방법임. 웹에서는 특정 라이선스가 적용된 작업물을 쉽게 인식할 수 있도록 "기계가 읽을 수 있는" 버전의 라이선스가 제공되는데, 이는 소프트웨어 시스템, 검색 엔진 및 기타 기술이 이해할 수 있는 형식으로 주요 자유와 의무를 요약한 것임. 기계 가독성(machine-readable)에 대해서는 CSV , JSON , XML 등과 같이 컴퓨터가 자동으로 읽고 처리할 수 있는 형식으로 정의함. (opendata handbook의 machine readable 정의: https://opendatahandbook.org/glossary/en/terms/machine-readable/)
- 연구자 기준:

  (1) Standard machine-readable license (e.g. Creative Commons)

  개별 기록물마다 Creative Commons, 공공누리와 같은 표준 라이선스를 기계가 읽을 수 있는 형식으로 제공하는 경우에 해당함. 라이선스 표시를 클릭할 경우, 해당 라이선스의 내용이 안내된 페이지로 이동하는 경우도 포함됨

  (2) Standard text-based license

  개별 기록물마다 표준 라이선스를 텍스트로만 제공하는 경우에 해당함

  (3) Non-standard machine-readable license (clearly indicating the data may be reused)

  개별 기록물마다 비표준 라이선스(기관에서 정의한 저작권 정책 등)을 기계가 읽을 수 있는 형식으로 제공하는 경우에 해당함

  (4) Non-standard text-based license

  개별 기록물마다 비표준 라이선스를 텍스트 또는 이미지(하이퍼링크를 제공하지 않는 형식)로 제공하는 경우에 해당함

  (5) No license

  개별 기록물에 라이선스를 명시하지 않은 경우에 해당함. 디지털 아카이브 수준의 이용약관 또는 저작권 정책만 제공하는 경우도 포함됨

R2: How much provenance information has been captured to facilitate data reuse?

- 디지털 객체의 출처 정보가 기재되어 있는 방법에 대한 항목
- ARDC 기준: 데이터의 출처는 데이터가 어디서 부터 생성되었으며, 어떤 과정에 의해서 생성되었는지에 대한 방법론에 대한 정보로, 데이터의 신뢰성과 재현성을 가능하게 하는 데이터의 진위성 확인에 중요한 정보임. 특히 연구가 데이터 데이터 집약적(intensive)이고 복잡한 데이터 변환과 절차를 포함하는 eScience 커뮤니티에서 더욱 중요해지고 있음.
- 연구자 기준:

  (1) Fully recorded in a machine-readable format

  기록물에 대한 출처 정보가 충분한(Fully) 설명이 존재하고, 기계가 읽을 수 있는 형식으로 제공된 경우에 해당함. GOFAIR의 "R1.2: (Meta)data are associated with detailed provenance" 원칙을 기반으로 아래 4가지 질문을 모두 충족한 경우에 'Fully'로 평가

  - Who generated or collected it? (기록물의 생산자 또는 수집자에 대한 정보를 명시하였는가?)
  - How has it been processed? (기록물의 획득방법 또는 처리방법에 대한 정보를 명시하였는가?)
  - Has it been published before? (기록물의 발행처 또는 기존에 출판된 적이 있었는가?)
  - Does it contain data from someone else that you may have transformed or completed? (기록물의 수정이나 변경된 사항에 대한 기술, 기록물 입수 시 원본이나 가공에 대한 정보를 명시하였는가?)

  (2) Fully recorded in a text format

  기록물에 대한 충분한(Fully) 설명이 존재하되, 텍스트로 기술된 경우에 해당함

  (3) Partially recorded

  기록물에 대한 출처 정보가 부분적(Partially)으로 기술된 경우에 해당함

  (4) No provenance information is recorded

  기록물에 대한 출처 정보가 없는 경우에 해당함