GeoAI 데이터 학회 춘계 컨퍼런스

GEO DATA 저널의 지리 정보 관리를 위한
데이터 리포지터리 요구 기능 설계

2024.06.27

한국과학기술정보연구원
연구데이터공유센터 DataON 개발팀
엄정호

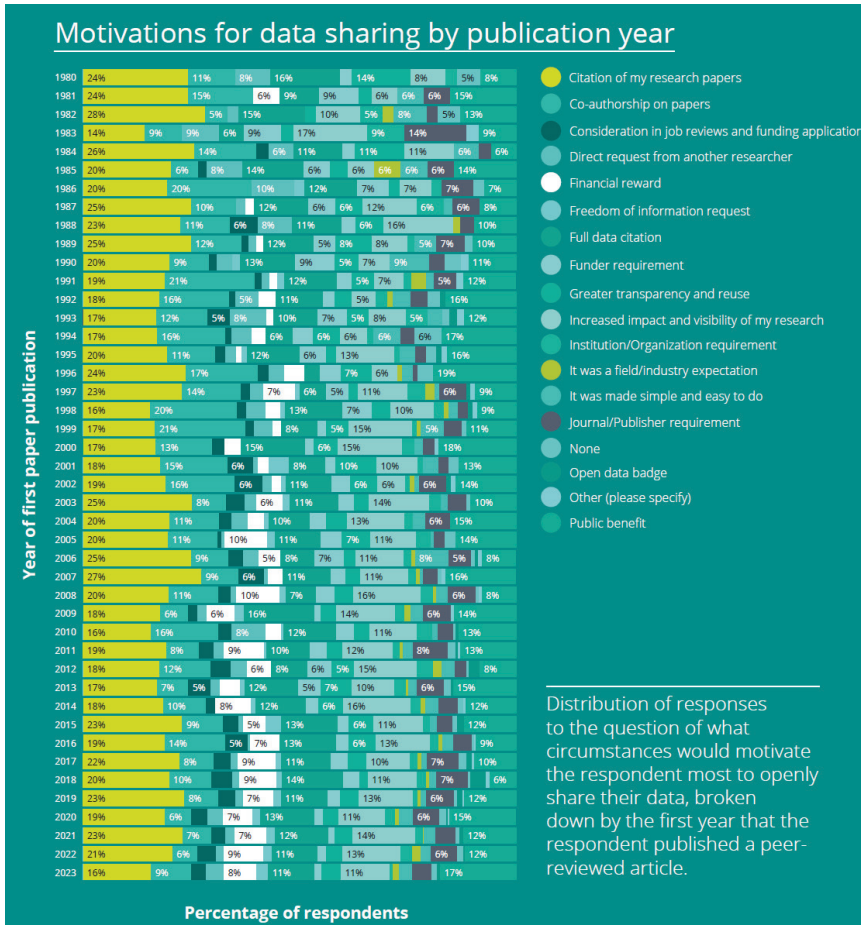
KiSTi www.kisti.re.kr 한국과학기술정보연구원 Korea Institute of Science and Technology Information

KISTI SINCE 1962

1. Introduction

데이터 격차 해소와 전지구적 문제 해결을 위한 오픈 데이터/오픈 사이언스에 대한 문화 확산

* 데이터 공유에 대한 인식은 기존 나의 평판을 올리기 위한 도구로 생각하였으나, 23년도 결과에서는 공공의 이득을 생각하는 방향으로의 인식 변화가 있었음. (Figshare, The State of Open Data 2023)



2022 Journal Metrics

On this page you will find a suite of citation-based metrics for *Scientific Data*. Brief definitions for each of the metrics used to measure the influence of our journals are included below the journal metrics. Data has been produced by Clarivate Analytics.

For recently launched journals, metrics are calculated from available citation data. If a metric uses multiple years of data, new journals may have partial metrics.

While the metrics presented here are not intended to be a definitive list, we hope that they will prove to be informative. The page is updated on an annual basis.

2-year Impact Factor - 9.8

5-year Impact Factor - 10.8

Immediacy index - 0.9

Eigenfactor® score - 0.04464

Article Influence Score - 33

* 대표 데이터 저널 Scientific Data는
2-year impact factor: 9.8
=> 연구자들의 많은 관심 증대

1. Introduction

국내 데이터 저널 GEO DATA
- KCI 후보지
- 현재 분기별 출판, 총 120편
  (Original Paper 114편, Review paper, Short Communication(Opinion, note, erratum) 총 6편)

학술지 상세

KCI 후보

GEO DATA (GEO DATA)

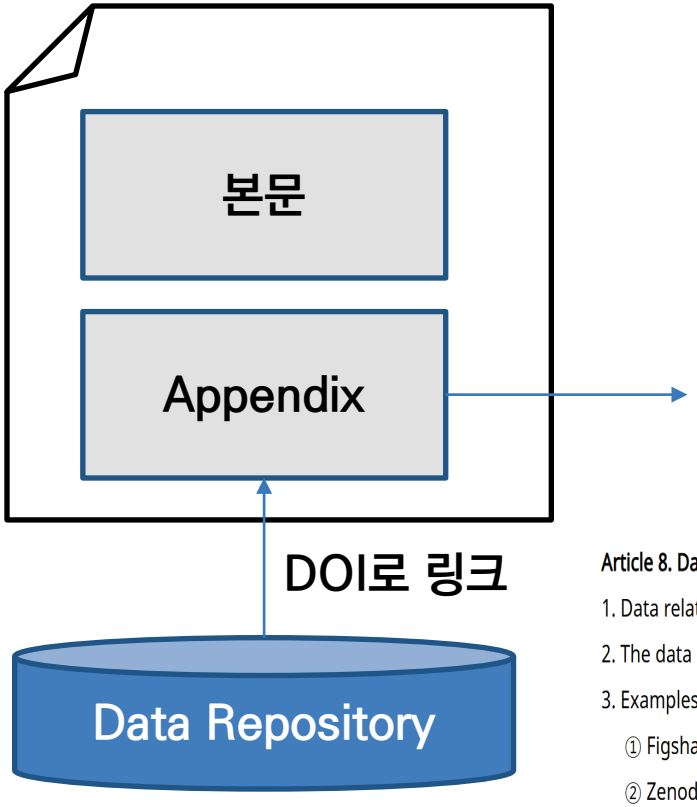
GEO DATA (GEO DATA)

연구분야	자연과학 > 지구과학	창간 연도	2019년 12월
ISSN		eISSN	2713-5004
최근 발행정보	2024년 06월, 6(2)	발행 간기	연4회 , 03월 31일, 06월 30일, 09월 30일, 12월 31일
언어	한국어 , 영어	학술지 홈페이지	https://geodata.kr/
발행 논문 수	연간 평균 23 건, 총 70 건 전체 논문 검색 인용보고서		
저작권 정책	* 저작권 : 학술지 발행기관 저작권 정책 더보기		

볼륨 (year-vol.-no)	Original	Opinion	review	note	erratum	합계
2019-v1-1	9	1				10
2020v2-1	7					7
2020v2-2	10					10
2021v3-1	4					4
2021v3-2	5					5
2021v3-3	3	1				4
2021v3-4	8					8
2022v4-1	4		1			5
2022v4-2	4			1		5
2022v4-3	3					3
2022v4-4	5					5
2023v5-1	8					8
2023v5-2	8					8
2023v5-3	13					13
2023v5-4	18		1			19
2024v6-1	5				1	6
	114	2	2	1	1	120

1. Introduction

GEO DATA 저널의 출판 절차(데이터 저널과 데이터 리포지터리*와의 관계)
* 데이터 리포지터리: 데이터 저장/보존 및 웹 상에 게시/검색 지원 시스템



[Metadata for Dataset]

1. The data investigated and utilized in this thesis should be described in the following technical items and written in English.
2. The data must be labeled with a Digital Objective Identification (DOI).

Sort	Field
Essential	Title
	DOI name
	Category
	Abstract
	Temporal coverage
	Spatial coverage
	Personnel
	CC License
Optional	Project
	Instrument

Original Paper에서만 관련 데이터 정보를 기술,
관련 데이터는 데이터 리포지터리의 DOI로 표기,
시간은 연월일로 표기하나 자유로운 서술
공간정보는 주로 위경도(WSG84)로 Point/Line/Polygon으로 표기하나 주소지(시/군/구)만 표기되기도 함

Article 8. Data Deposit Method

1. Data related to an article shall be deposited in the appropriate data repository.
2. The data repository must support the DOI issuance system to issue DOI addresses to the deposited data.
3. Examples of data repositories are as follows.
   ① Figshare
   ② Zenodo
   ③ Dryad
   ④ EcoBank
   ⑤ GeoBigdata OpenPlatform

DOI: Digital Object Identifier
WGS: World Geodetic System

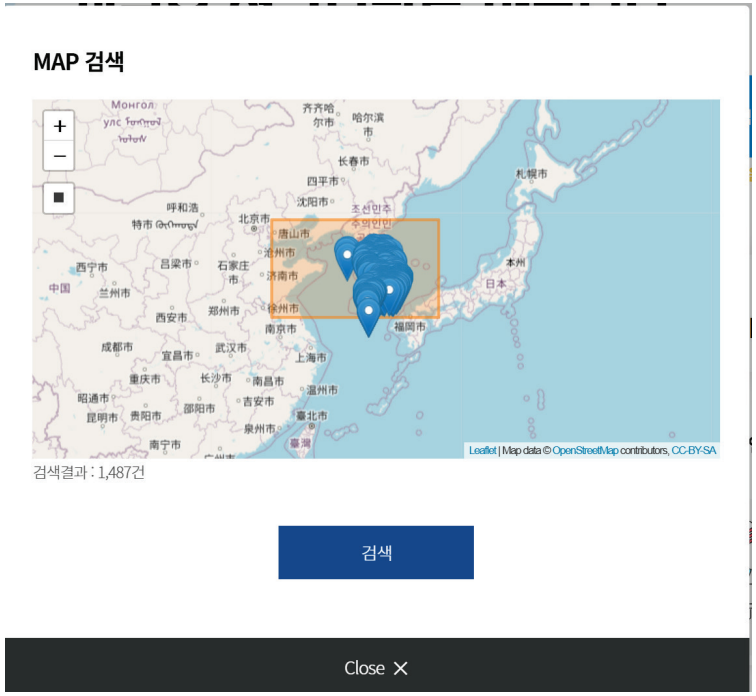
1. Introduction

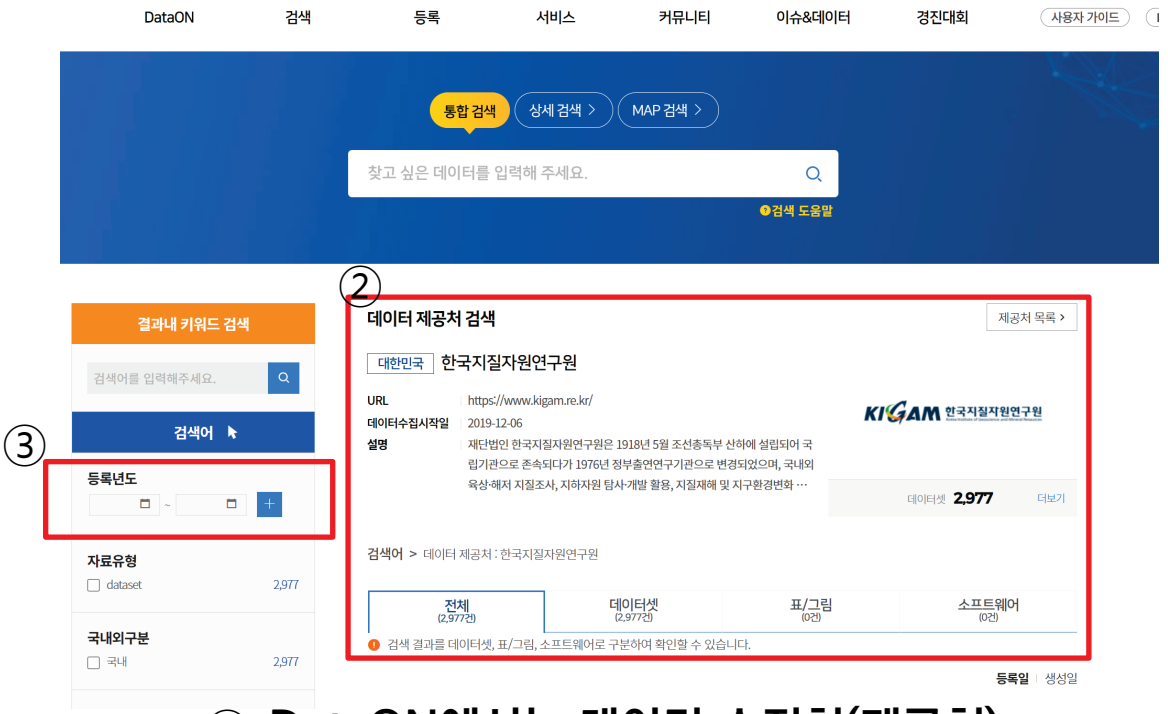
DataON: 2020년 연구데이터를 한 곳에서 수집부터 분석/활용을 지원하기 위해 개발된 플랫폼
(국제 리포지터리 신뢰 인증 CoreTrustSeal(CTS) 획득, 24.03)

서비스 기능	서비스 개요
데이터 수집	각 연구기관에서 운영하는 데이터 서비스(리포지터리)로부터 메타데이터를 수집, 통합 검색 제공
데이터 등록	연구자가 DataON을 통해 연구데이터를 직접 등록 가능
데이터 검색	DataON에서는 수집 및 등록된 연구데이터, KISTI NRMS 에서 운영하는 보고서로부터 추출한 표/그림, DataON Canvas를 통해 등록한 소프트웨어 정보를 통합하여 키워드 및 맵 검색 가능
데이터 분석/활용	**App**: 데이터 분석 라이브러리를 재활용할 수 있도록 공유 **Workflow**: 가시화된 데이터 워크플로우를 제공하는 DataON Canvas **JupyterHub**: Python/R 프로그래밍을 Jupyter에서 수행할 수 있도록 제공 **MyDrive**: 나만의 데이터를 저장 및 분석에 활용할 수 있는 공간
데이터 공유/커뮤니티	동일한 관심분야의 연구자들이 토론하고 데이터를 공유할 수 있는 커뮤니티 기능 제공
IDR 호스팅	전산자원이 부족한 연구기관에서 데이터 리포지터리를 운영할 수 있도록 클라우드 자원을 통한 호스팅 서비스 제공

2. Motivation

GEO DATA의 요구사항: 저널 리포지터리에서는 시공간정보에 대한 검색 기능 부재
=〉 DataON의 맵검색을 활용 및 확장하여 GEO DATA 논문에 대한 시공간 검색 요청





① DataON에서는 OpenStreetMap을 활용하여 지도에 데이터를 표기,
Elastic Search로 데이터 정보 색인

② DataON에서는 데이터 수집처(제공처) 별로 데이터를 검색 가능,
③ 등록년도(시간정보) 패싯 검색 지원

DataON에서 GEO DATA 논문에 대한 시공간 검색을 위한 기능 요구사항
i) 저장: GEO DATA 논문에서 시공간 정보를 추출 및 정제 필요
ii) 검색: ①+②+③ 의 통합 필요

3. GEO DATA 저널의 시공간 정보 추출 및 이슈 도출

1. 저널 Paper의 XML 파일 정보 다운로드

2. XML 정보를 파싱 및 Appendix 추출

3. 시공간 정보 관리 정제 이슈 도출

시간정보(Temporal Coverage), 공간정보(Spatial Coverage)의 데이터 정형화(큐레이션) 필요

4. 시공간 정보 이슈 분석(1/2)

시간 정보(Temporal Coverage) 이슈

1) 시간 범위를 표현하는 기술이 자유로움.

```
DJ2021-3-4-002.xml,From 1 Jan 2003 UTC through 31 Dec 2020,
DJ2021-3-2-001.xml,July to October 2016
DJ2021-3-2-003.xml,2011 ~ 2017 (yearly)
```

2) 종료 시간이 현재인데, 현재 시점을 논문/데이터 등에서 찾아야 함.

```
<td valign="middle" align="left">*Temporal Coverage</td>
<td valign="middle" align="left">1986-Now</td>
<td valign="middle" align="left">*Temporal Coverage</td>
<td valign="middle" align="left">June 2016~Current</td>
```

3) 시작 및 종료가 없이 기간만 표기

```
<td valign="top" align="left">*Temporal Coverage</td>
<td valign="top" align="left">4 months</td>
```

4) 기간 및 측정 주기(hourly)가 포함되어 있는지가 모호함,

```
DJ2021-3-2-003.xml,2011 ~ 2017 (yearly)
DJ2021-3-4-006.xml,April 2011 ~ December 2012 (hourly)
```

- 주기가 있다면, 주기 단위로 데이터 분할 여부 결정 필요

4. 시공간 정보 이슈 분석(2/2)

공간 정보(Spatial Coverage) 이슈

1) 위, 경도에서 '도 ' 를 나타내는 문자 상이

```
Latitude 35oN-39oN,Longitude 128oE-132oE,WGS84
coastal waters of California,36°N ~ 39°N, -121°E ~ -123°E,
```

2) 위, 경도의 '도/분/초 ' 까지 관리 필요?

```
Upper right: N 38° 02′18.65″ E 128° 28′04.46″,- Lower left: N 38° 02′16.25″
```

3) 'Spatial Coverage', 'Spatial Coverage (WGS 84) 등 공간정보의 컬럼명 표기가 상이

```
<td valign="middle" align="left">*Spatial Coverage (WGS84)</td>
<td valign="middle" align="left">WGS84</td>
<td valign="middle" align="left">126.25, 33.75 / 126.5, 33.5</td>
```

4) 공간범위를 알 수 없음

```
GD-2023-0048.xml,2017-2020,Rural areas of the world,
```

5) 나라, 지역 등에 대한 boundary 정보 필요

```
<td valign="middle" align="left">*Spatial Coverage</td>
<td valign="middle" align="left">South Korea, China, France, Canada, Fiji</td>

<td valign="top" align="left" rowspan="2">*Spatial Coverage</td>
<td valign="top" align="left">Chungcheong-do (South Korea)</td>
```

5. 기타 이슈

시공간정보를 나타내는 테이블 명칭이 상이

```
if child.text == 'Metadata for Dataset' or child.text == '데이터셋에 대한 메타데이터'
or child.text == '메타데이터' or child.text == 'MetaData for Dataset' or child.text ==
'Metadata' or child.text == '데이터셋 에 대한 메타데이터':
```

XML 파일에서 시공간정보 테이블을 지칭하는 태그가 상이

```
if child.tag == 'title' or child.tag == 'label':
```

6. 결론 및 제언

DataON과 GEO DATA 저널의 시공간 정보 연결 및 검색을 위해서는 데이터에 대한 정형화가 필요

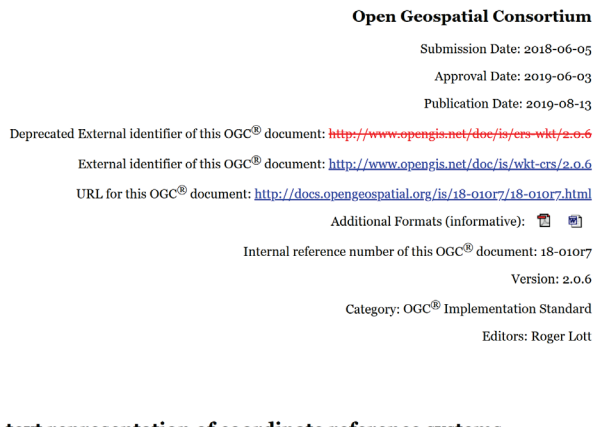
=〉 연관 데이터로 등록한 데이터 DOI를 통해, 실제 데이터 정보에 대한 분석

=〉 GEO DATA 저널의 Temporal, Spatial Coverage에 대한 정형화/표준화 가이드

시공간 정보에 대한 XML 스키마는 Geographical Markup Language(GML) 등의 표준 고려 필요.

The OpenGIS® Geography Markup Language Encoding Standard (GML) The Geography Markup Language (GML) is an XML grammar for expressing geographical features. GML serves as a modeling language for geographic systems as well as an open interchange format for geographic transactions on the Internet. As with most XML based grammars, there are two parts to the grammar – the schema that describes the document and the instance document that contains the actual data. A GML document is described using a GML Schema. This allows users and developers to describe generic geographic data sets that contain points, lines and polygons. However, the developers of GML envision communities working to define community-specific application schemas that are specialized extensions of GML. Using application schemas, users can refer to roads, highways, and bridges instead of points, lines and polygons. If everyone in a community agrees to use the same schemas they can exchange data easily and be sure that a road is still a road when they view it. Clients and servers with interfaces that implement the OpenGIS® Web Feature Service Interface Standard read and write GML data. GML is also an ISO standard (ISO 19136:2007).

OpenGIS에서 GML 정의 정보



Open Geospatial Consortium

Submission Date: 2018-06-05

Approval Date: 2019-06-03

Publication Date: 2019-08-13

Deprecated External identifier of this OGC® document: ~~http://www.opengis.net/doc/is/crs-wkt/2.0.6~~

External identifier of this OGC® document: http://www.opengis.net/doc/is/wkt-crs/2.0.6

URL for this OGC® document: http://docs.opengeospatial.org/is/18-010r7/18-010r7.html

Additional Formats (informative):

Internal reference number of this OGC® document: 18-010r7

Version: 2.0.6

Category: OGC® Implementation Standard

Editors: Roger Lott

Geographic information — Well-known text representation of coordinate reference systems

OpenGIS에서 Well Known Text 정의

질의 응답

KISTI SINCE 1962