データサイテーションマイニングによる科学データの利活用分析

Data Citation Analysis Framework for Open Science Data

是津 耕司,村山 泰啓, 渡邉 堯
Koji Zettsu, Yasuhiro Murayama, Takashi Watanabe

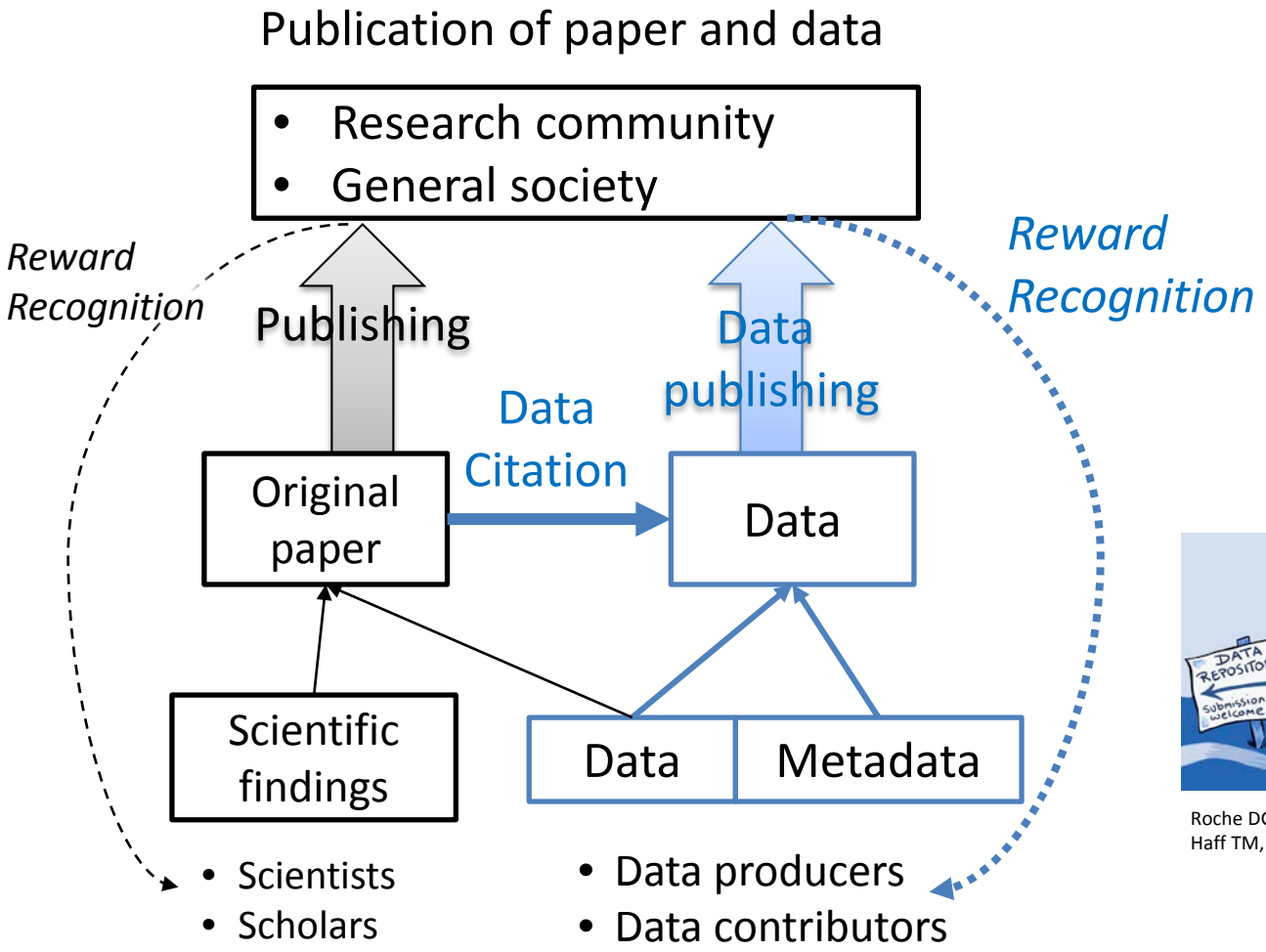
国立研究開発法人 情報通信研究機構
National Institute of Information and Communications Technology

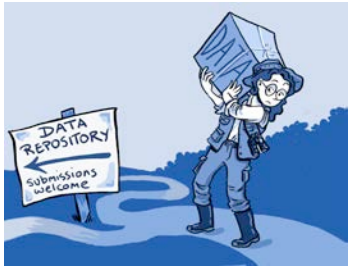
第2 回オープンサイエンスデータ推進ワークショップ
2015年12月7-8日

Out of Cite, Out of Mind

CODATA-ICSTI Task Group on Data Citation Standards and Practices (2013) Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal 12, CIDCR1–CIDCR75.

Data Citation as a Driver of Scientific Open Data





Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014)

Reference: Society of Geomagnetism, Earth, Planetary and Space Sciences, http:// sgepss.org/sgepss/shorai/SGEPSS_syorai_Jan2013.pdf [accessed on January 2013].

Data Citation Practices

Data Publisher	Description	Data Citation Example
PANGAEA: The Publishing Network for Geo-scientific and Environmental Data http://www.Pangaea.de/	Open access library and data publisher for earth and environmental science	*Gershanovich, DE; Zinkovskiy, AB (1987): Distribution of particulate matter and particulate organic carbon in waters of the Caspian Sea.doi:10.1594/PANGAEA.756520*
ICPSR: The Inter-university Consortium for Political and Social Research https://www.icpsr.umich.edu/	International consortium of about 700 academic institutions and research organizations that maintains and provides assess to social science data	*Escarce, Jose J., Nicole Lurie, and Adria Jewell. RAND Center for Population Health and Health Disparities (CPHHD) Data Core Series: Pollution, 1988-2004 [United States]. ICPSR27864-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2011-10-21. http://doi.org/10.3886/ICPSR27864.v1*
Dryad http://http://datadryad.org//	International data repository of peer reviewed scholarly literature specialized in bioscience data	*López-Rodríguez MJ, Tierno de Figueroa JM (2012) Data from: Life in the dark: on the biology of the cavernicolous stonefly Protonemura gevi (Insecta, Plecoptera). The American Naturalist http://dx.doi.org/10.5061/dryad.8m8r1*

… and more

Source: CODATA-ICSTI Task Group on Data Citation Standards and Practices: Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data, Data Science Journal, Vol. 12, pp. CIDCR1-CIDCR75 (2013)

Data Citation Practices (Cont'd)

- Data Citation Index (DCI) [Thomson Reuters]
  - Harvests citations to research data from papers indexed in the Web of Science

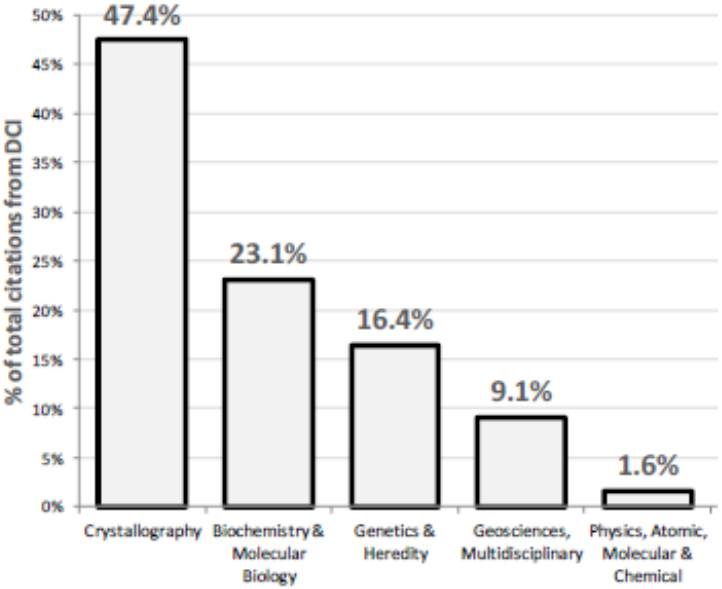
A. DCI overall statistics

	All Document Types	Data repositories	Data studies	Data sets
Total Citations	404,211	3,265	106,895	294,051
Total Records	2,623,528	90	154,674	2,468,736
Uncited Records	2,311,553	63	126,428	2,185,062
% Uncited	88.11	40.0	81.74	88.51
Citation Average	0.15	36.28	0.69	0.12
Standard Deviation	3.06	336.07	9.56	0.36

B. DCI statistics by area of data studies

	Total Records	% Records	Total Citations	% Citations	Citation Average
Engineering & Technology	1,545	0.06	890	0.30	0.58
Humanities & Arts	44,588	1.81	1	0.00	0.00
Science	2,004,449	81.19	293,193	99.71	0.15
Social Sciences	424,952	17.21	7	0.00	0.00

C. Distribution of citations by subject



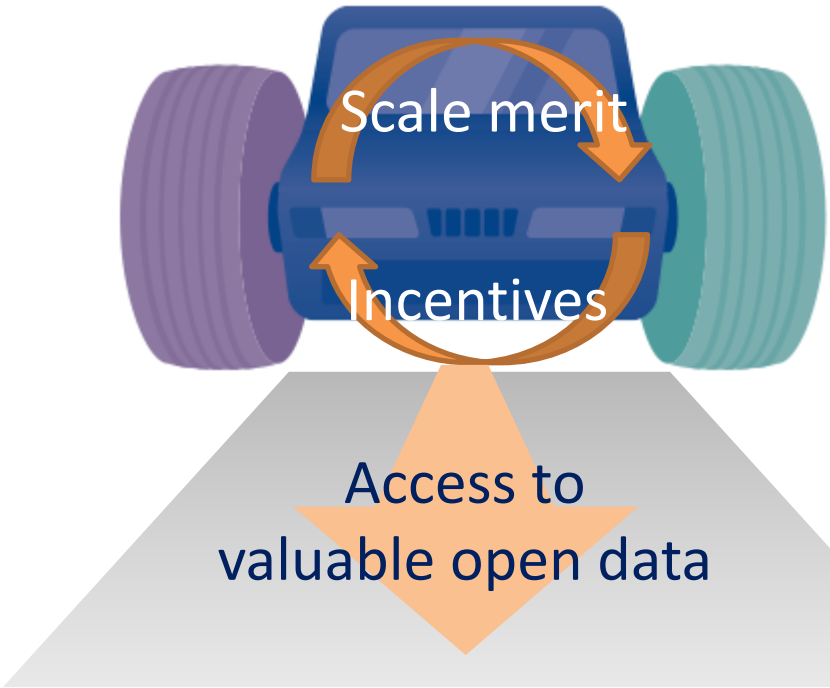
***Source**: Robinson-Garcia, N. et. al: Analyzing data citation practices using the Data Citation Index, Journal of the Association for Information Science and Technology, DOI: 10.1002/asi.23529 (June, 2015)*

Two Wheels for Driving Data Citation

Data Citation

Provision

- Principles
- Standards & practices

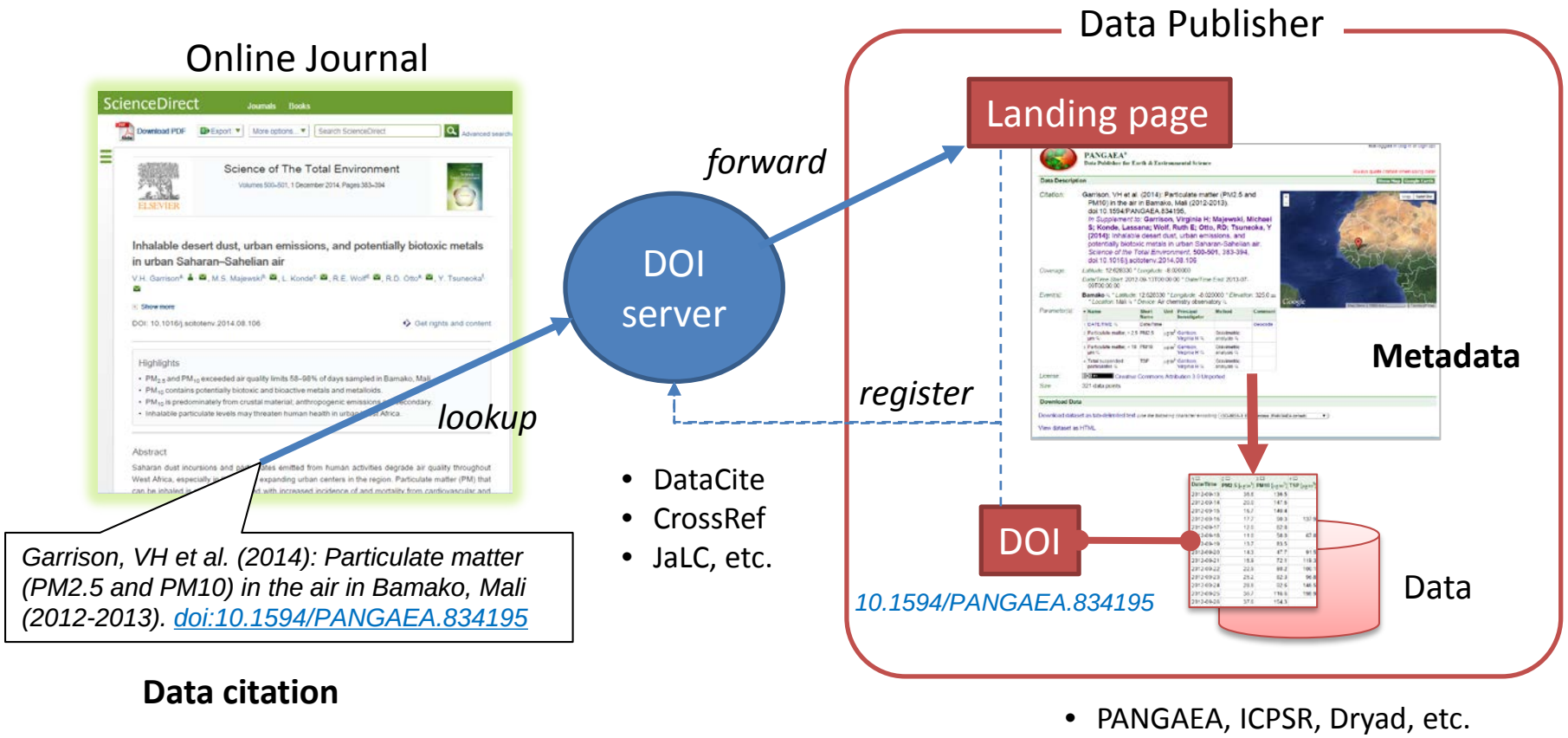


Usage

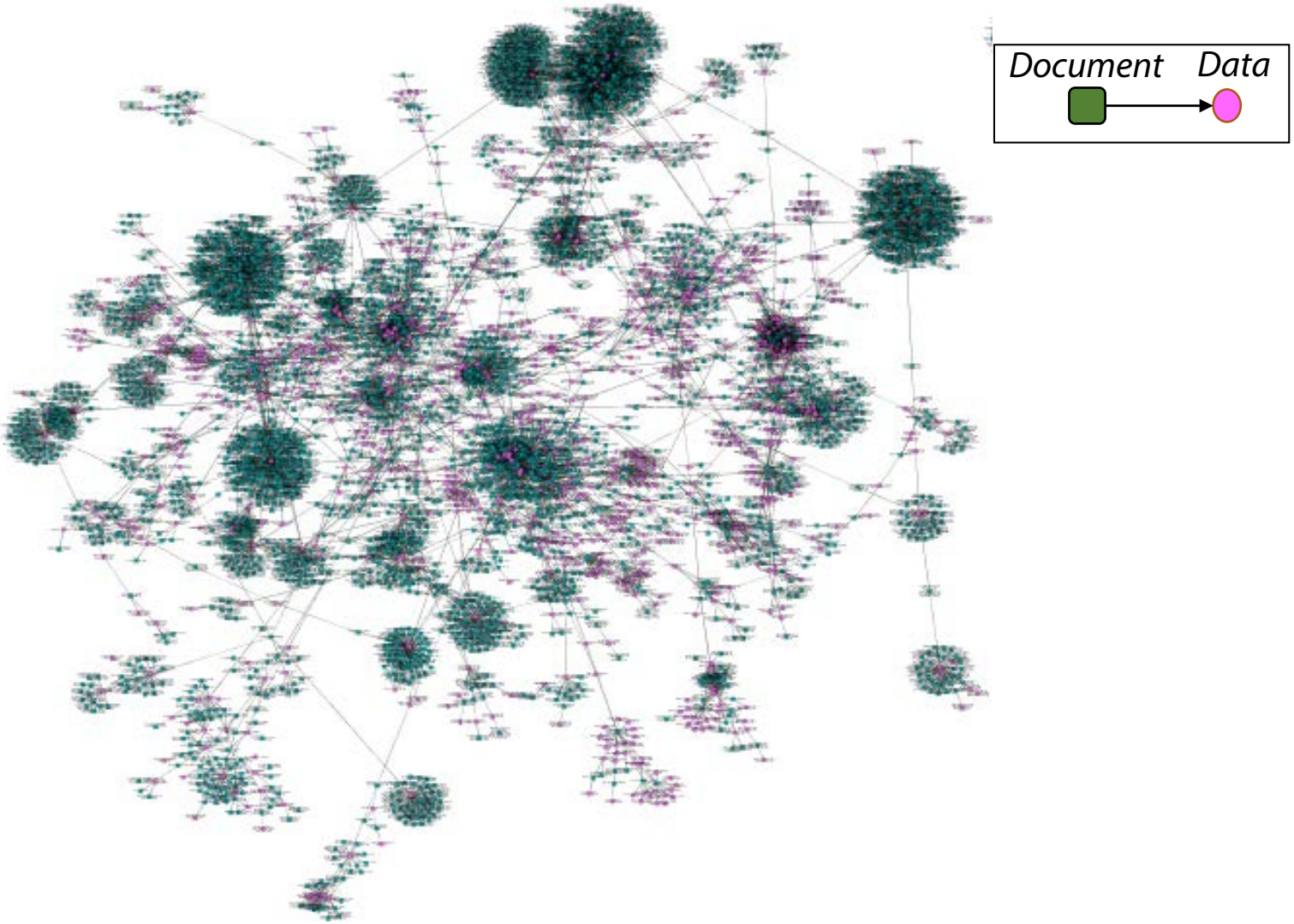
- Data evaluation
- Data discovery

Open Science

Data Citation Mechanism



Web of Data Citation



Inter-university Consortium for Political and Social Research (ICPSR)

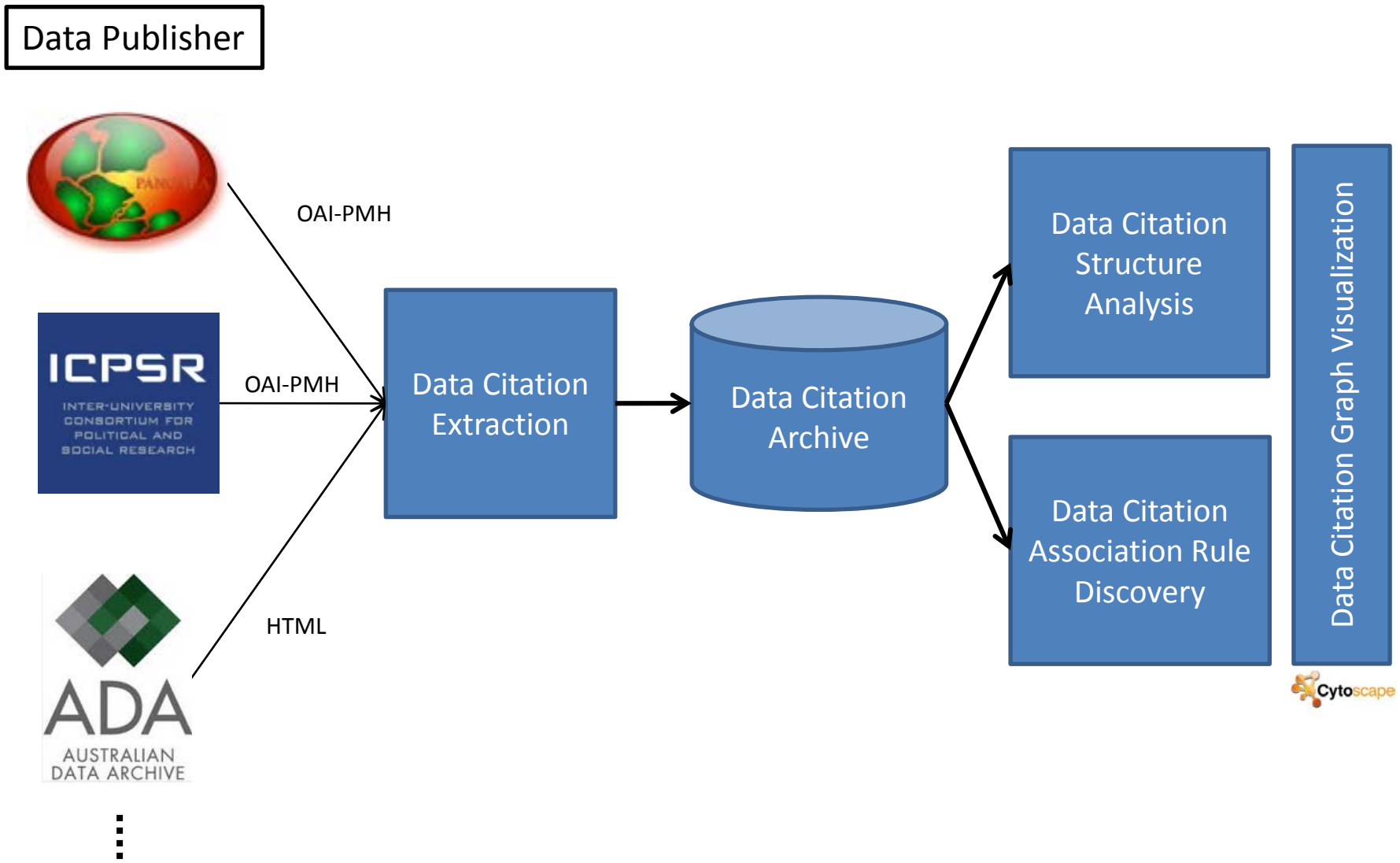
115,154 citations

Data Citation Analysis

- ***Macro analysis:***
  Analyze structure of data citation network
  - Discover communities of data citation, and characterize data by citations in a community

- ***Micro analysis:***
  Analyze associations between document and data
  - Discover typical associations (*i.e.* ***association rules***) between documented knowledge and evidential data

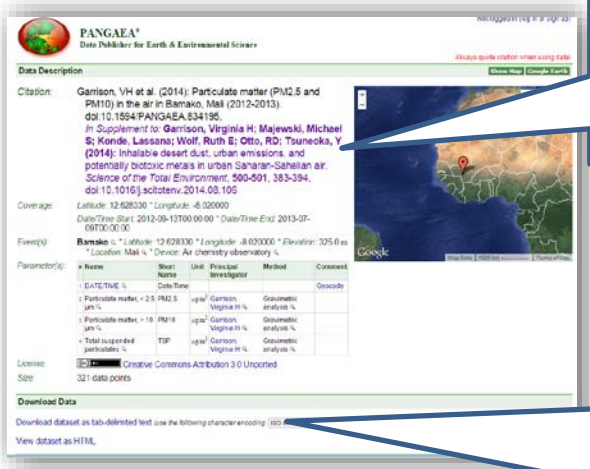
Data Citation Analysis Framework

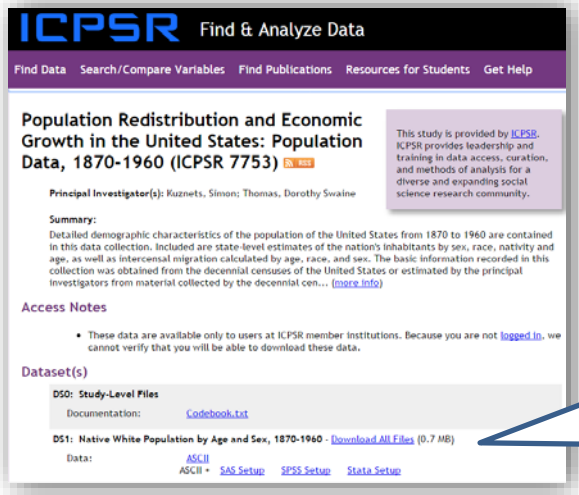


Data Citation Extraction

Data Citation Metadata

Landing pages





Referencing documents

Citation: Garrison, VH et al. (2014): Particulate matter (PM2.5 and PM10) in the air in Bamako, Mali (2012-2013). doi:10.1594/PANGAEA.834195,
In Supplement to: **Garrison, Virginia H; Majewski, Michael S; Konde, Lassana; Wolf, Ruth E; Otto, RD; Tsuneoka, Y (2014):** Inhalable desert dust, urban emissions, and potentially biotoxic metals in urban Saharan-Sahelian air. *Science of the Total Environment*, **500-501**, 383-394, doi:10.1016/j.scitotenv.2014.08.106

Related to:

Bers, A Valeria; Momo, Fernando; Schloss, Irene R; Abele, Doris (2013): Analysis of trends and sudden changes in long-term environmental data from King George Island (Antarctica): relationships between global climatic oscillations and local system response. *Climatic Change*, **116**, 789-803

Klöser, Heinz; Ferreyra, Gustavo A; Schloss, Irene R; Mercuri, Guillermo; Laturnus, Frank; Curtosi, Antonio (1993): Seasonal variation of algal growth conditions in sheltered Antarctic bays: the example of Potter Cove (King George Island, South Shetlands). *Journal of Marine Systems*, **4**, 289-301, doi:10.1016/0924-7963(93)90025-H

Schloss, Irene R; Abele, Doris; Moreau, Sébastien; Demers, Serge; Bers, A Valeria; Gonzáles, Oscar; Ferreyra, Gustavo A (2012): Response of phytoplankton dynamics to 19-year (1991-2009) climate trends in Potter Cove (Antarctica). *Journal of Marine Systems*, **92**, 53-66

Schloss, Irene R; Ferreyra, Gustavo A (2002): Primary production, light and vertical mixing in Potter Cove, a shallow bay in the maritime Antarctic. *Polar Biology*, **25**, 41-48, doi:10.1007/s003000100309

Schloss, Irene R; Ferreyra, Gustavo A; Ruiz-Pino, Diana (2002): Phytoplankton biomass in Antarctic shelf zones: a conceptual model based on Potter Cove, King George Island. *Journal of Marine Systems*, **36**, 129-143, doi:10.1016/S0924-7963(02)00183-5

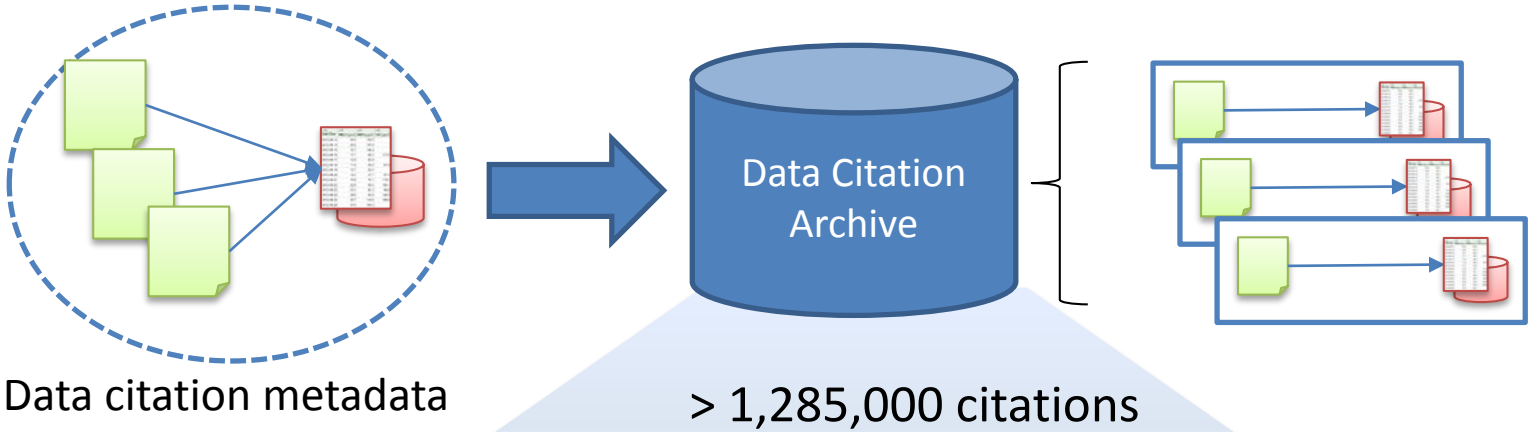
Related Publications

1988 Long, Larry E. Migration and Residential Mobility in the United States. New York: Russell Sage.
Export Options: RIS/EndNote

1964 Eldridge, Hope T., Thomas, Dorothy Swaine . Demographic Analyses and Interrelations. *Population Redistribution and Economic Growth, United States, 1870-1950 series, vol. 3.* Philadelphia, PA: American Philosophical Society.
Export Options: RIS/EndNote

1960 Kuznets, Simon Smith, Miller, Ann Ratner, Easterlin, Richard A. Analyses of Economic Change. *Population Redistribution and Economic Growth: United States, 1870-1950 series, vol. 2.* Philadelphia, PA: American Philosophical Society.
Export Options: RIS/EndNote

Data Citations Archive



Data citation metadata

> 1,285,000 citations

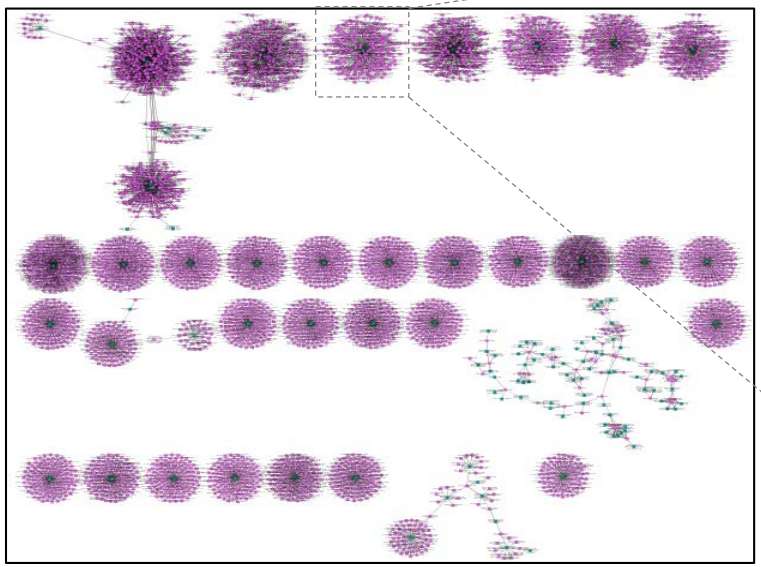
Data Site	Domain	# of DC
Pangaea	Earth & Environment	322,477
ICPSR	Social Science	114,815
DataCite	(any)	773,173
DRYAD	Bioscience	1,556
ADA	Social Science	16,062
ESDS	Economic & Social Science	59,471

Data Citation Structure Analysis (1)

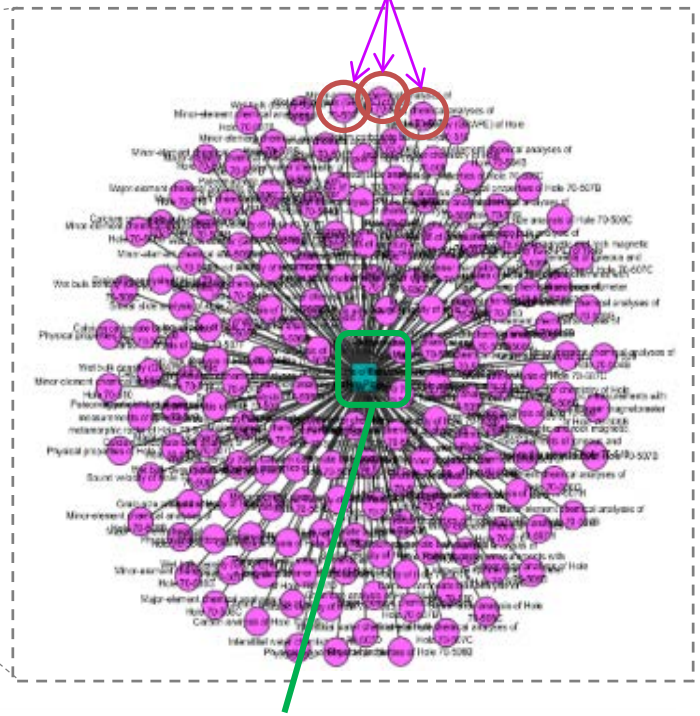
- Data Collection Community

Data collection

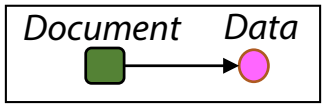
"Physical properties of Hole #"



Pangaea



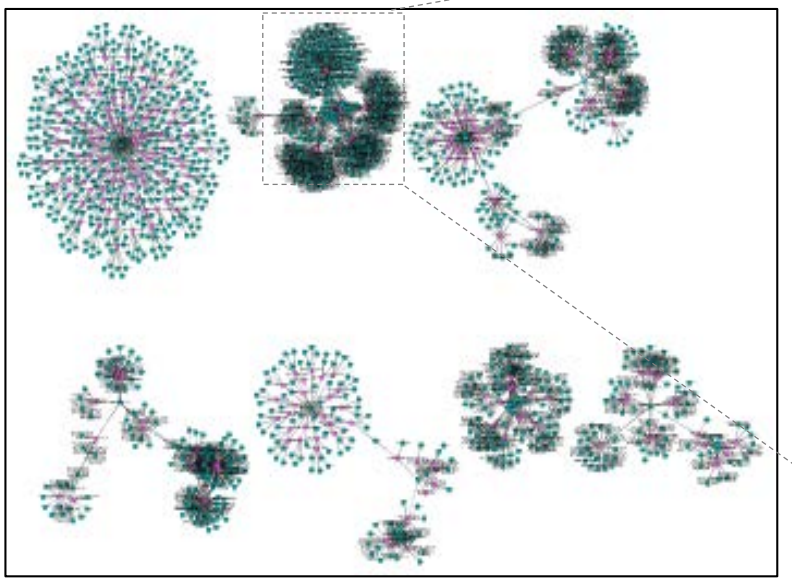
"Reports of the Deep Sea Drilling Project"



Catalogue document

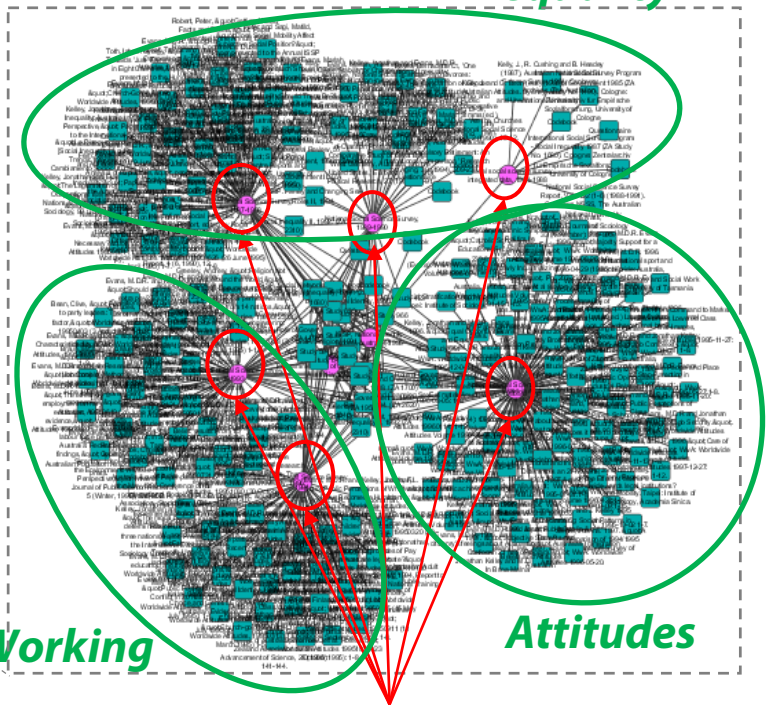
Data Citation Structure Analysis (2)

- Data Sharing Community

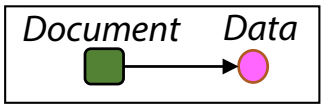


Australian Data Archive (ADA)

Data-sharing document clusters



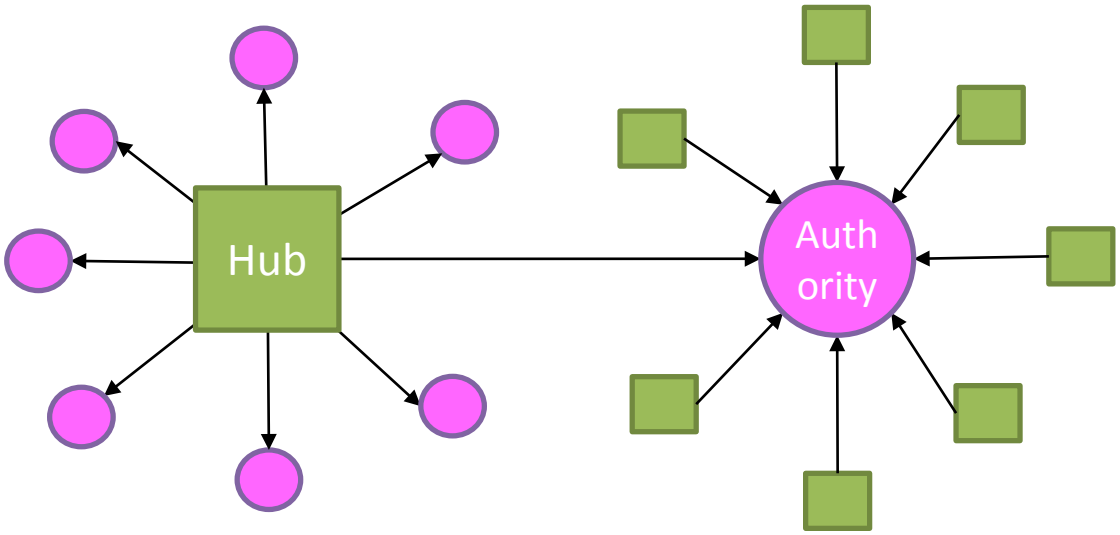
National Social Science Survey



Shared data

Topic-Specific Community Discovery

- HITS algorithm [Kleinberg 99]
  - A good **hub** links to many good authorities
    - Hub score: $H(x) = \sum_{y \leftarrow x} A(y)$
  - A good **authority** is referenced by many good hubs
    - Authority score: $A(x) = \sum_{y \rightarrow x} H(y)$
  - Discovery from result set



Hubs and Authorities

Hub document

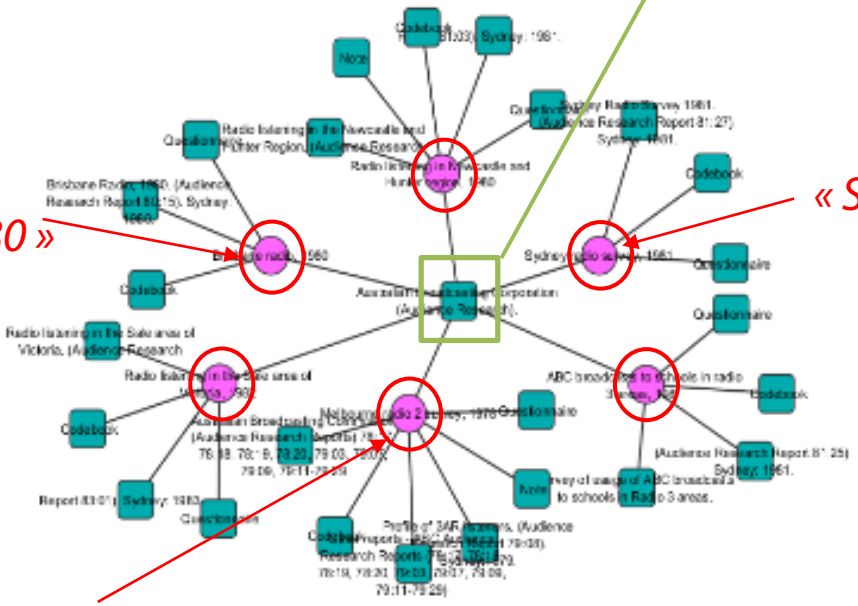
"Australian Broadcasting Cooperation (Audience Research)"

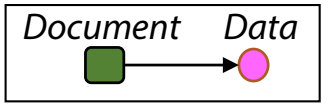
« Brisbane radio, 1980 »

« Sydney radio survey, 1981 »

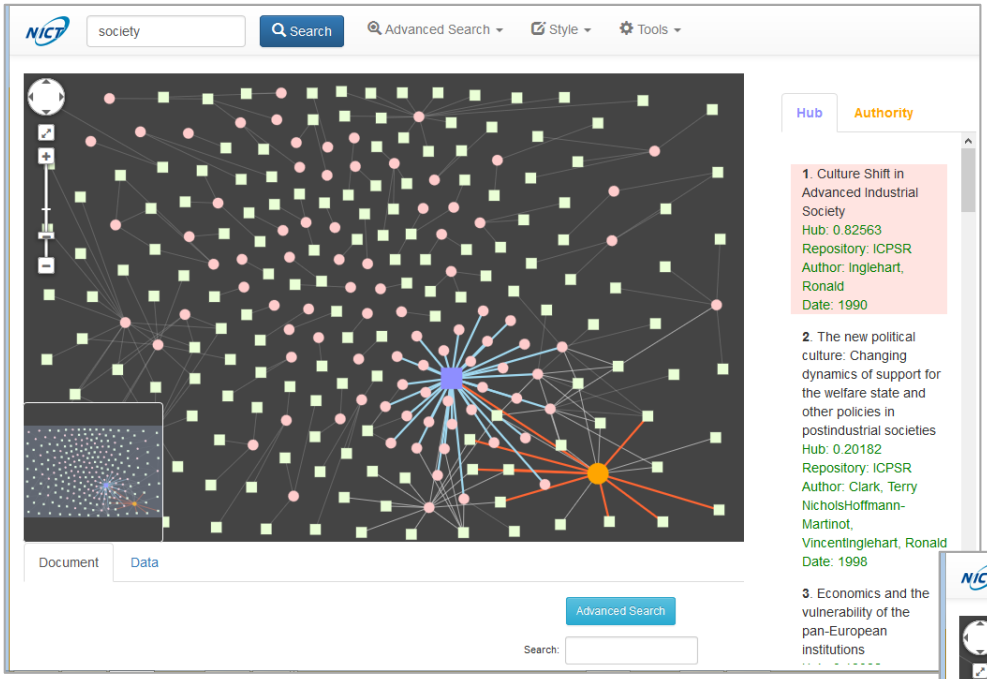
Authority data

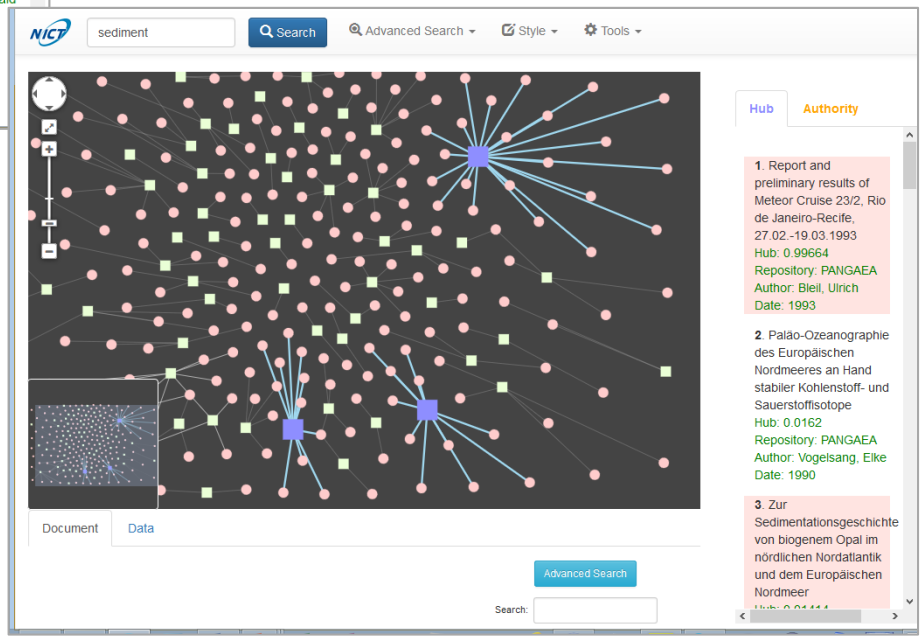
« Melbourne radio survey, 1978 »





Community Discovery Demo



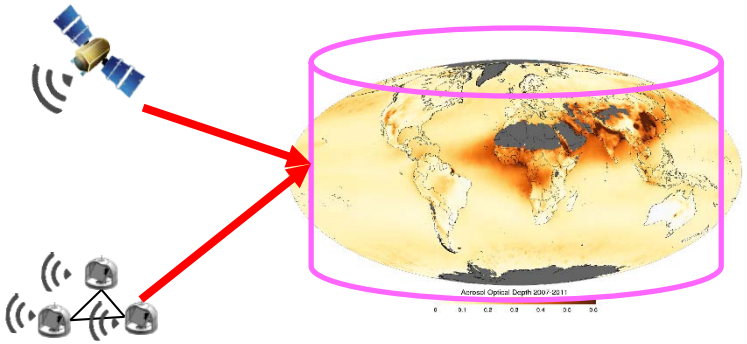


Referential Context: Basic Concept

Provider's view

User's view

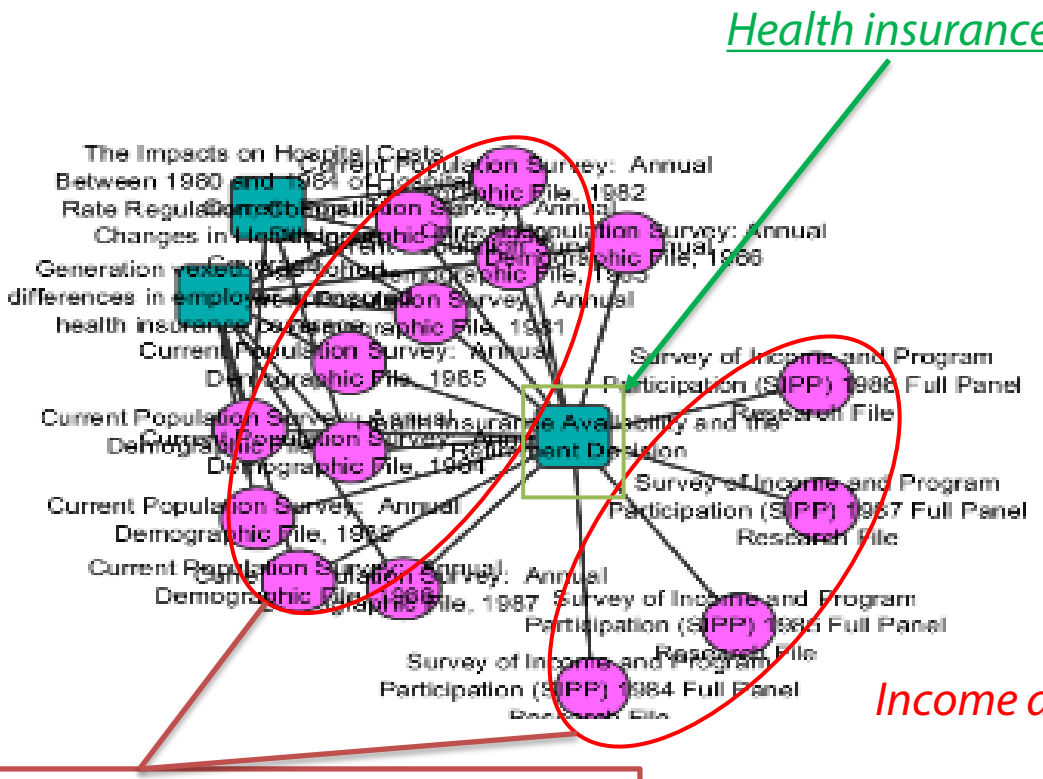
"aerosol optical depth"



"air pollution data"

Referential Context in Data Citation Web

- Data usage is often different from the data content

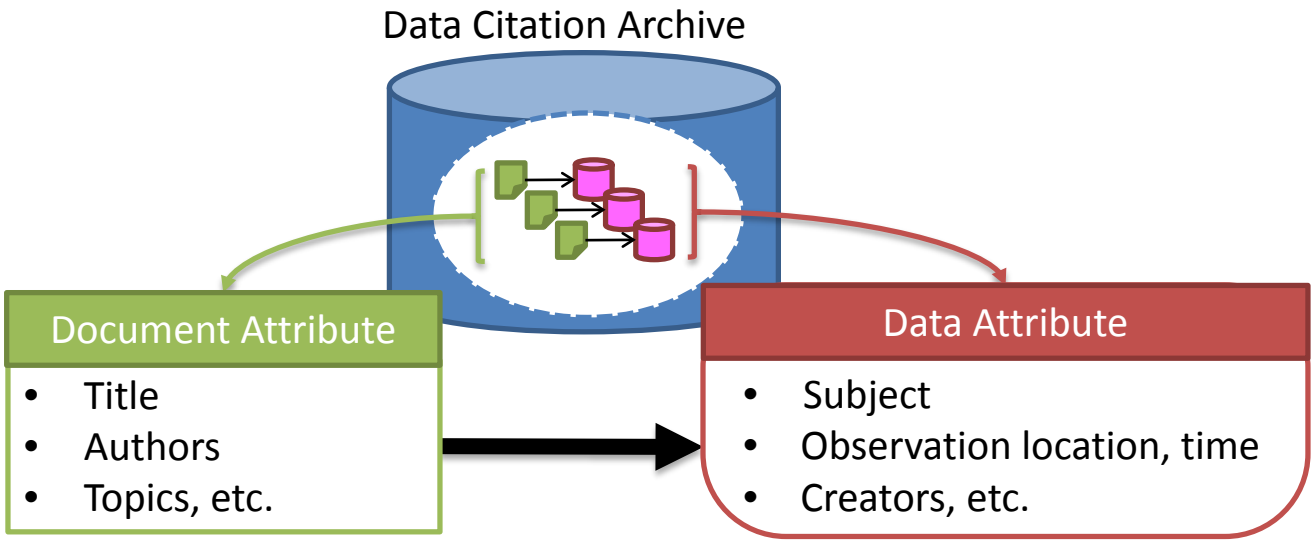


Both population data and income data are referenced by health insurance document

Data Citation Association Rule Discovery

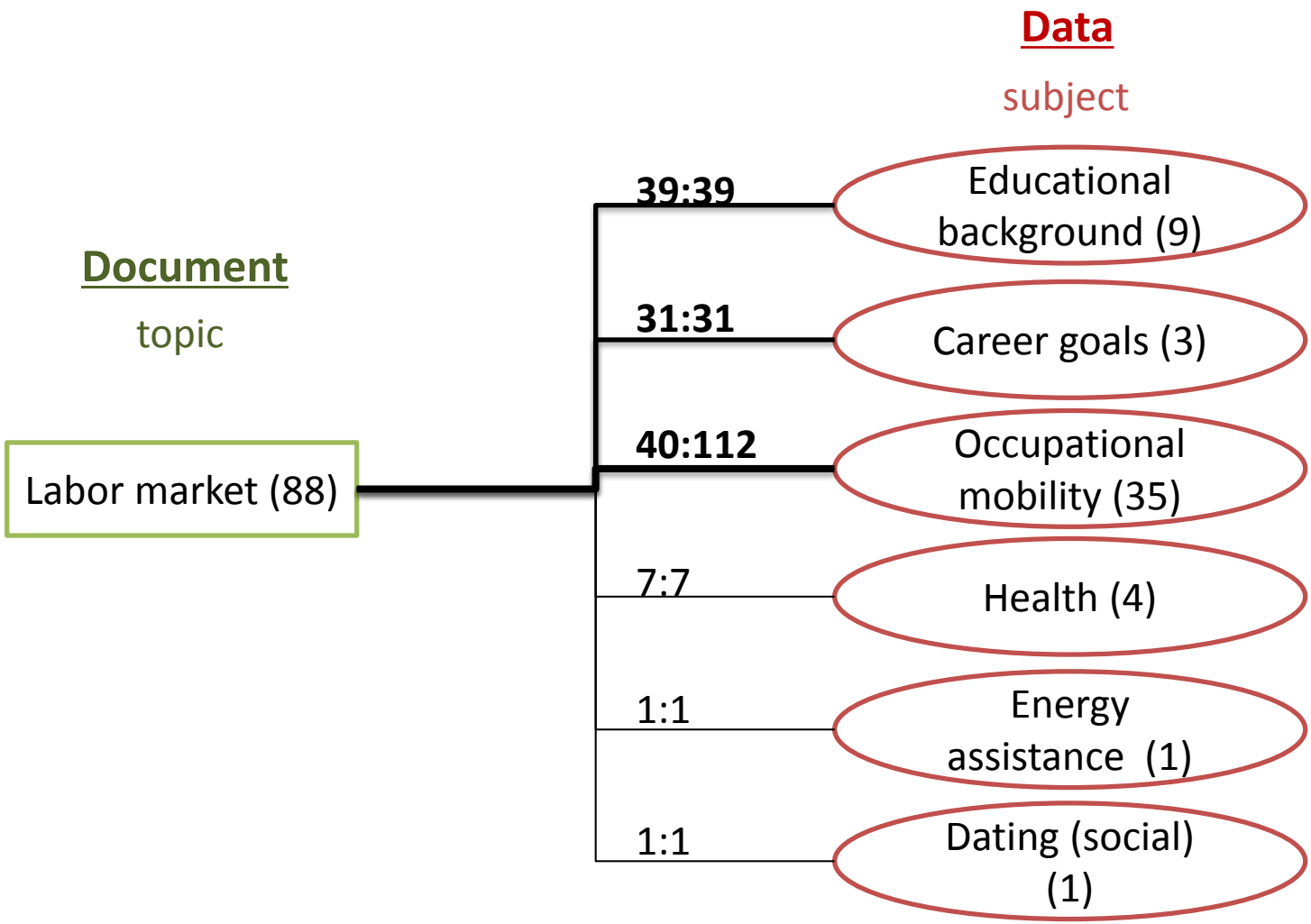
- Discover typical combinations of document and data attributes frequently co-occurring in data citations (referential contexts)

Referential Context:



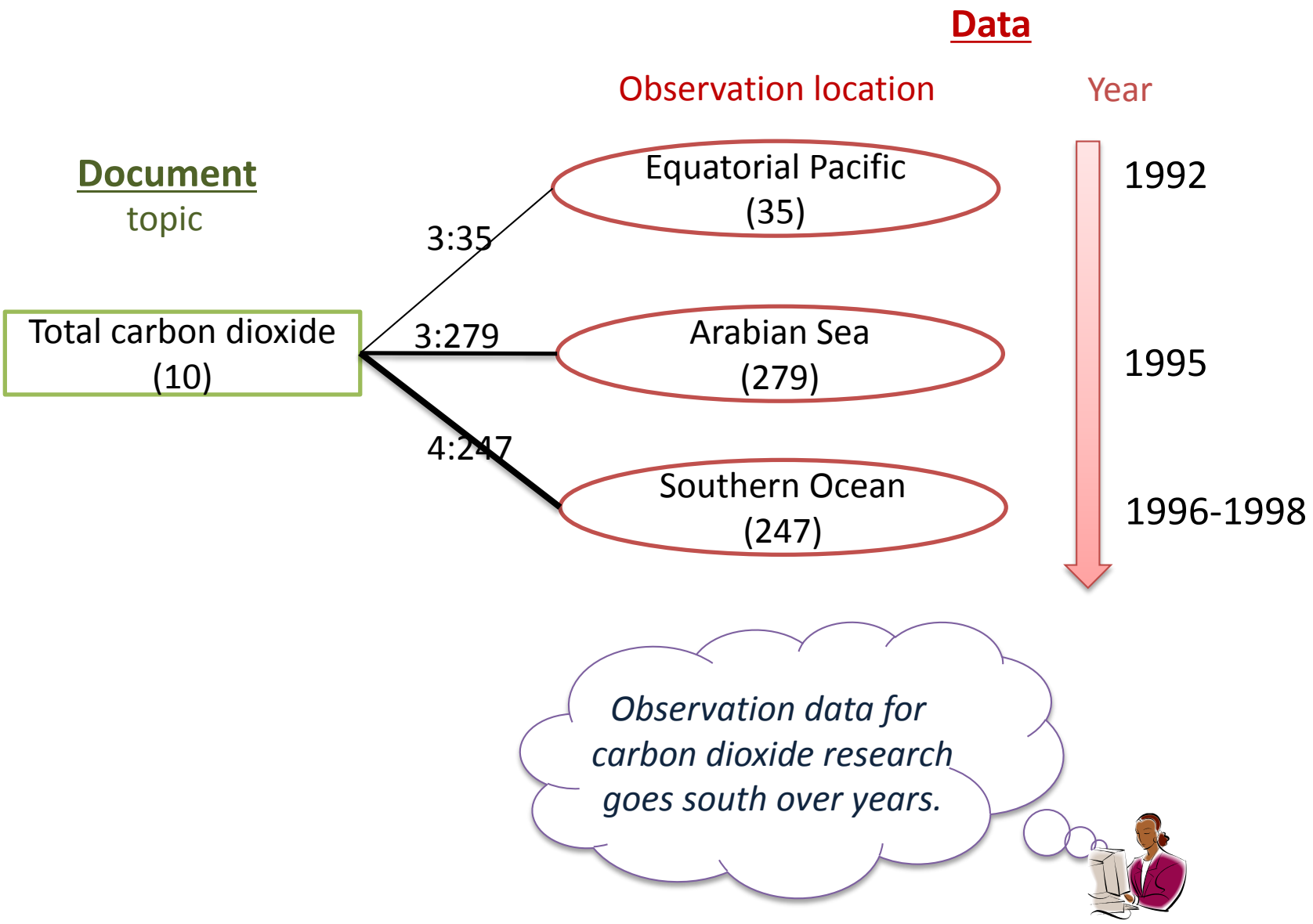
Association rule discovery tool

Typical Data Supporting Documented Knowledge



Data citations (*#source: #target*)

Trend of Knowledge-Supporting Data



Data Supporting Wide Knowledge

Document

Diverse topics

- Deep-Sea Research Part A. Oceanographic Research Papers (1) — 1 : 1
- Quaternary Research (2) — 2 : 2
- Quaternary Science Reviews (1) — 1 : 1
- Geophysical Research Letters (1) — 1 : 1
- Radiocarbon (1) — 1 : 1
- American Journal of Science (1) — 1 : 1
- Paleoceanography (2) — 2 : 2
- Marine Geology (1) — 1 : 1
- Earth and Environmental Science Transactions of the Royal Society of Edinburgh (1) — 1 : 1
- Earth and Planetary Science Letters (2) — 2 : 2
- Nature (4) — 4 : 4

Data

title

Sedimentation rates calculated on surface sediment samples from different site of the Atlantic and Pacific Oceans, 1991.

Data collected from 1963 ~ 1981 all over the world

Created by **Wallace Smith Broecker** (1931 -)

Use Cases

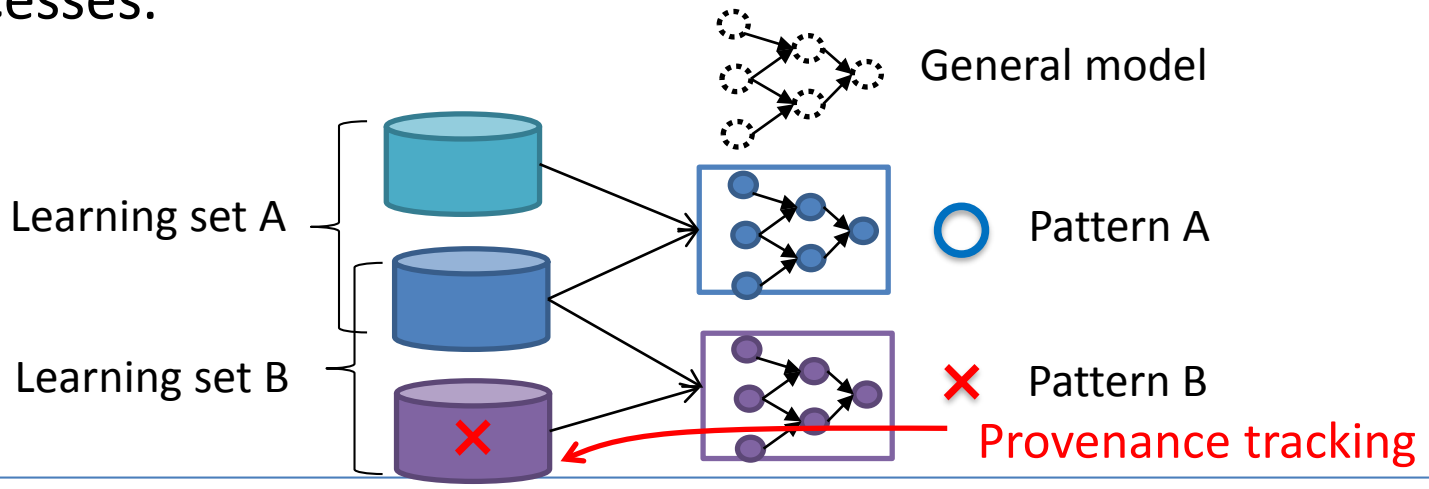
- Discover data-intensive community
  - Collaborate with research communities having same or similar data [*researcher*]
  - Survey the data common to a research community [*data repository*]
- Evaluate reputation of data
  - Reward data (creators) based on its popularity and/or authority [*funding organization*]
  - Manage quality of data [*data creator*]
- Provide superior discoverability for better reuse of data
  - Search data by both content and context keywords [*data repository*]
  - Discover related data for interdisciplinary research [*data curator*]
  - Find research publications actionable for reuse of data [*researcher*, *publisher*]

Open Issues

- Unstable metadata of data citation
  - Semantic compatibility among heterogeneous data citation metadata
    - E.g.) *relate to, supplement to (PANGAEA), related publications (ICPSR), is referenced by (Dryad), related materials (ADA), ….*
  - Up-to-date?
- More citations for better analysis
  - Unified and/or centralized access to distributed information of data citation
  - Citation-creating applications with harnessing citation analysis
    - E.g.) Data search with citation-based ranking (*more citations, more exposure*) → data citation optimization

New Challenges

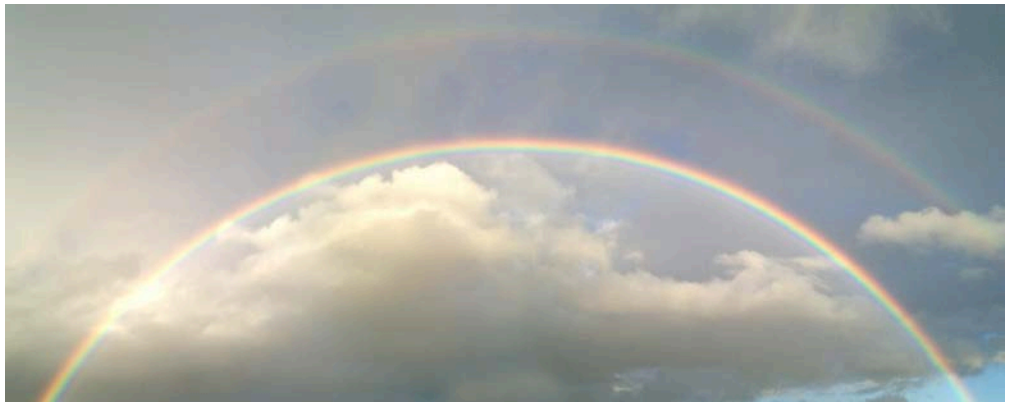
- Analyzing *dynamic* citations
  - **Behavior analysis** based on user-to-data citations obtained from data (DOI) access log
    - E.g.) "*Users Who Took This Item Also Took*" (social filtering)
  - **Intention analysis** based on keyword-to-data citations obtained from search query log
  - **Quality analysis of data-intensive AI models** based on process-to-data citations obtained from machine learning processes.



Summary

- Data citation = link from documented knowledge to evidential data
  - Instead of knowledge-to-knowledge link by document citation
- Analysis of ‘Web of data citation’
  - Data citation structure analysis (macro analysis)
  - Data citation association rule discovery (micro analysis)
- For better reuse & reward of data
  - Discover data-intensive community
  - Evaluate reputation of data
  - Provide superior discoverability

THANK YOU



Data Citation Mining
are going to
open source software project

Contact: zettsu@nict.go.jp