「研究データ共有の理想と現実、そして実践可能性」

～地球環境分野の研究基盤に関する意識調査から～

東京大学　地球観測データ統融合連携研究機構（EDITORIA）
小野 雅史

自己紹介：専門

- 専門は「空間情報科学」
- 様々な分野と接する学際性が特徴
- 最近の研究テーマ:
  - 空間最適化による気象＋交通・物流の融合

1) http://www.oki-gis.jp/
2) http://hiroshima.mapping.jp/

自己紹介：経歴

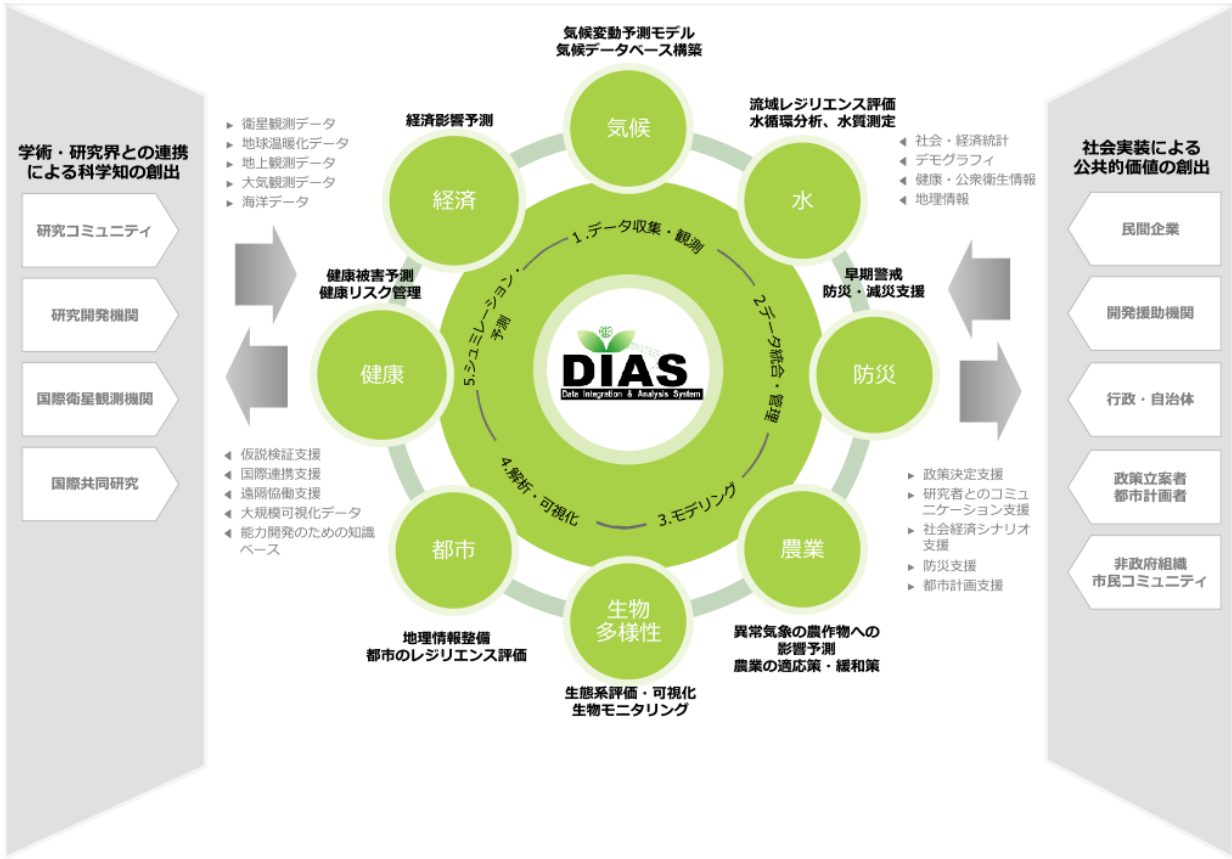
これまで関わった異分野融合プロジェクト

- **ISO/TC211**
  - 地理情報の国際標準
  - 地理情報標準メタデータ, 場所識別子(Place Identifier) などの検討
- **DIAS/GRENE-ei**
  - 地球環境分野を対象にした、DIASメタデータ(ISO準拠)の設計
  - 語彙管理システムの開発を担当
- **GEO (Group on Earth Observation)**
  - 各国の地球観測機関を中心とした国際グループ
  - オントロジー、ベストプラクティス・レジストリ等のタスクチームに参加
- **Belmont Forum e-Infrastructure and Data Management**
  - 人的リソースと資金をもとに、国際連携するための枠組み
  - WP4「Data Sharing」グループのメンバー

DIAS http://www.diasjp.net/

地球環境分野の研究データ共有基盤

- DIAS1期: 2006-2010年 (※第3期科学技術基本計画)
  「データ統合・解析システム (**DIAS**)」
- DIAS2期: 2011-2015年 (※第4期科学技術基本計画)
  「地球環境情報統融合プログラム (**DIAS-P**)」
  「グリーンネットワーク・オブエクセレンス
  環境情報分野 (**GRENE-ei**)」
- DIAS3期：2016年-2020年 (※第5期科学技術基本計画)
  「地球環境情報プラットフォーム
  構築推進プログラム(**DIAS-PF**)」
  「基幹アプリケーション・フィジビリティスタディ(**DIAS-FS**)」



科学研究者
（データ提供者/利用者）

コミュニティ基盤

2期: GRENE-ei
3期: DIAS-FS

2期: DIAS-P
3期: DIAS-PF

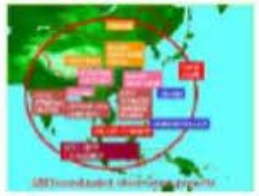
データ基盤

IT技術者

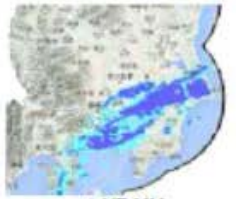
データ統合・解析システム(DIAS)
データインフラ上での科学知の創造と社会的利益の創出

DIAS
Data Integration & Analysis System

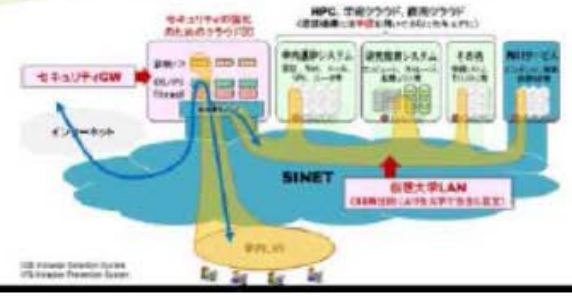
携帯電話によるタッカにおける300万人の移動



アジアモンスーン年プロジェクト

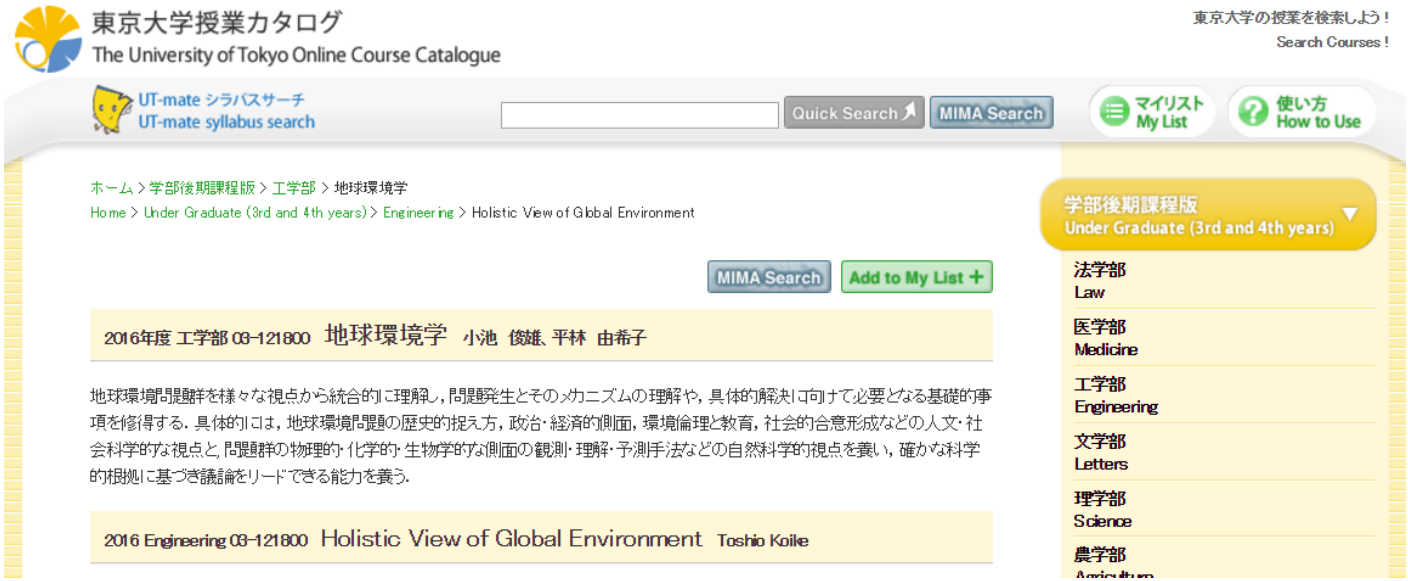


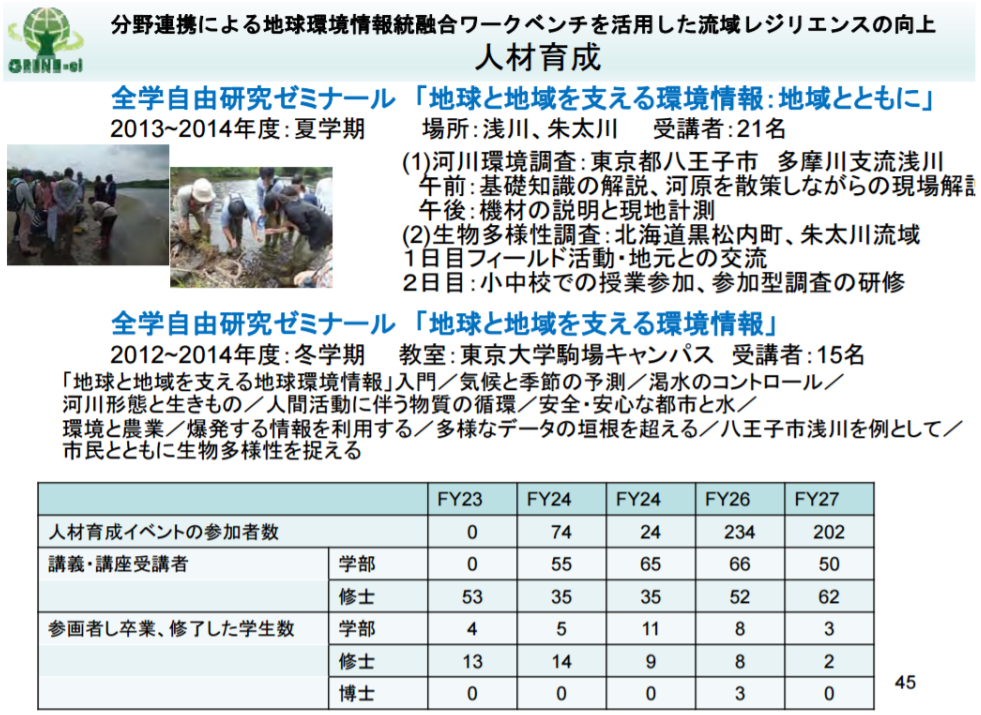
↑XRAIN
リアルタイムデータ利用
ひまわり8号↓



教育

- 大学内での講義
- サマースクールなど

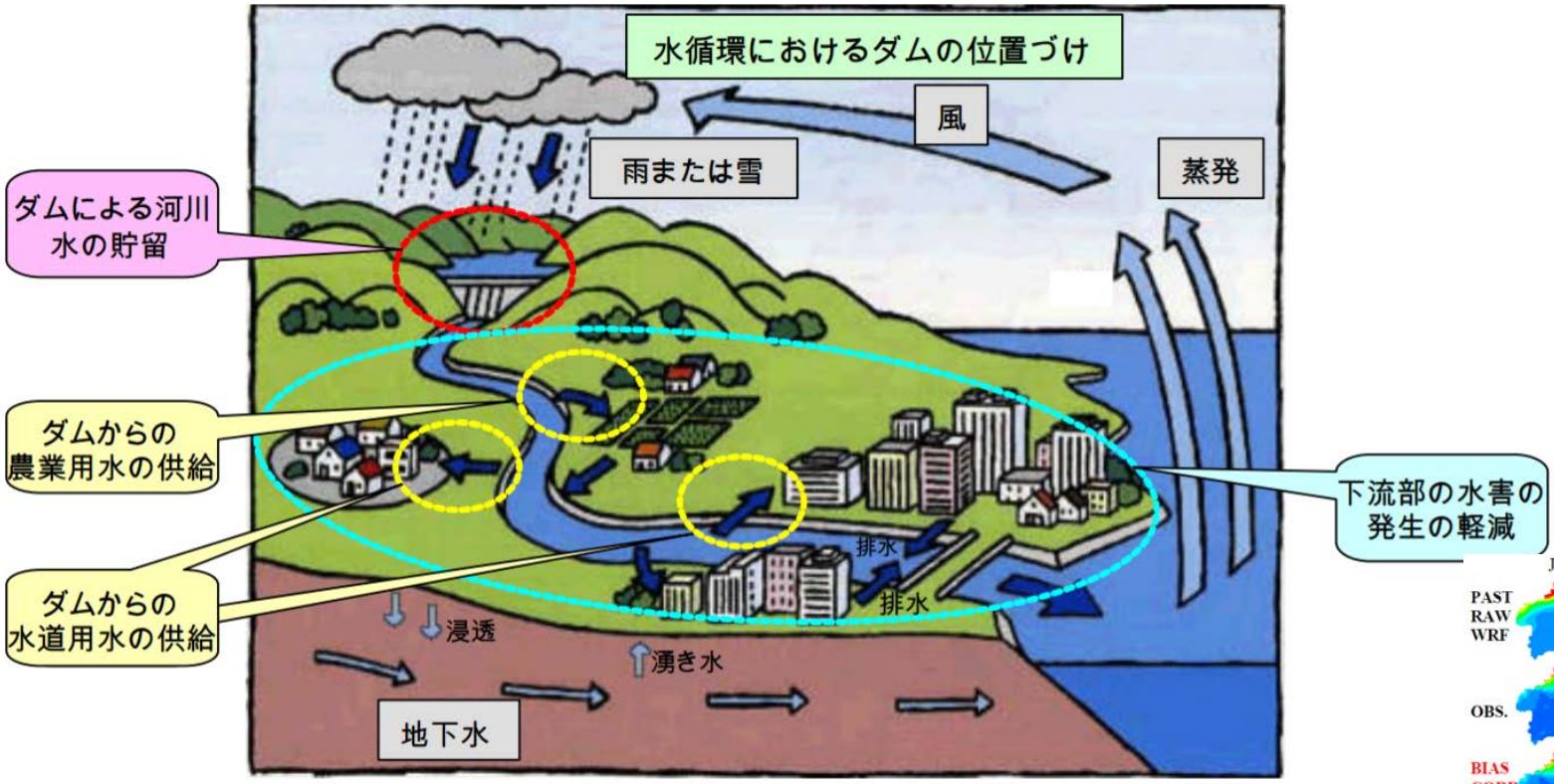




		FY23	FY24	FY24	FY26	FY27
人材育成イベントの参加者数		0	74	24	234	202
講義・講座受講者	学部	0	55	65	66	50
	修士	53	35	35	52	62
参画者し卒業、修了した学生数	学部	4	5	11	8	3
	修士	13	14	9	8	2
	博士	0	0	0	3	0

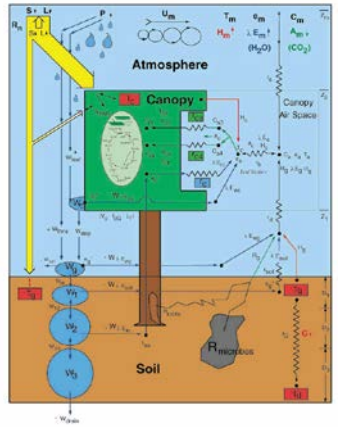
産学連携

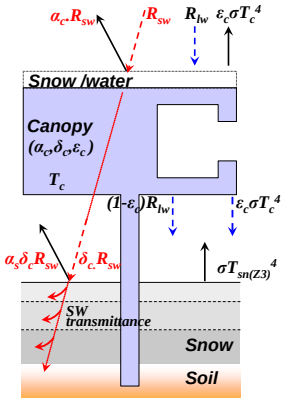
- 河川管理・ダム操作支援システム
  - 東京大学、日本工営、東京電力、中部電力、土木研究所と共同
  - 洪水や融雪時の、安全・最適な流水量管理へ



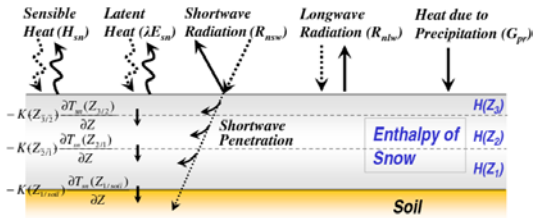
国土交通省 河川局 河川環境課「新しい時代のダム管理のあり方」

エネルギー水収支

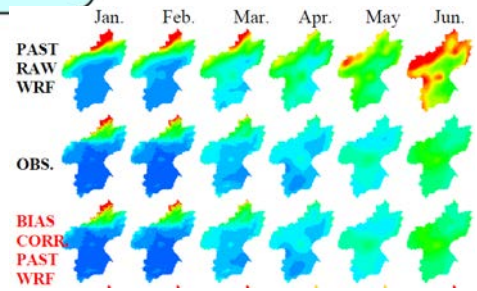




Snow over Forest region

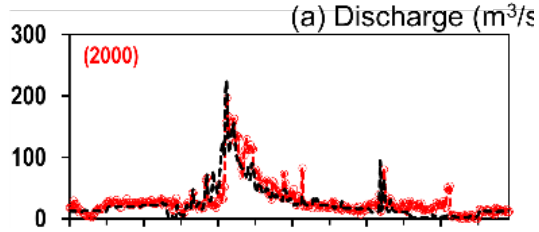


Snow over Bare Land



利根川上流での積雪面積

利根川上流での流出



社会やビジネスの宝としてのDIAS

持続可能な基盤に向けたビジネスモデルの構築へ

データ
インフラ
アプリケーション
研究者陣
実績

- 地球環境を中心とした、信頼できる多彩なデータ
- 科学技術外交、国際協力への貢献

- 運用体制
- 大規模ストレージ、計算リソース
- 充実したAPI、データの切り出し

- 国内外の社会課題解決に向けたソリューションの提供（水、生物多様性、気象ほか）

- 一流の国際的研究者陣
- 社会課題解決にコミットした研究陣
- 世界との緊密なコミュニティー

＋

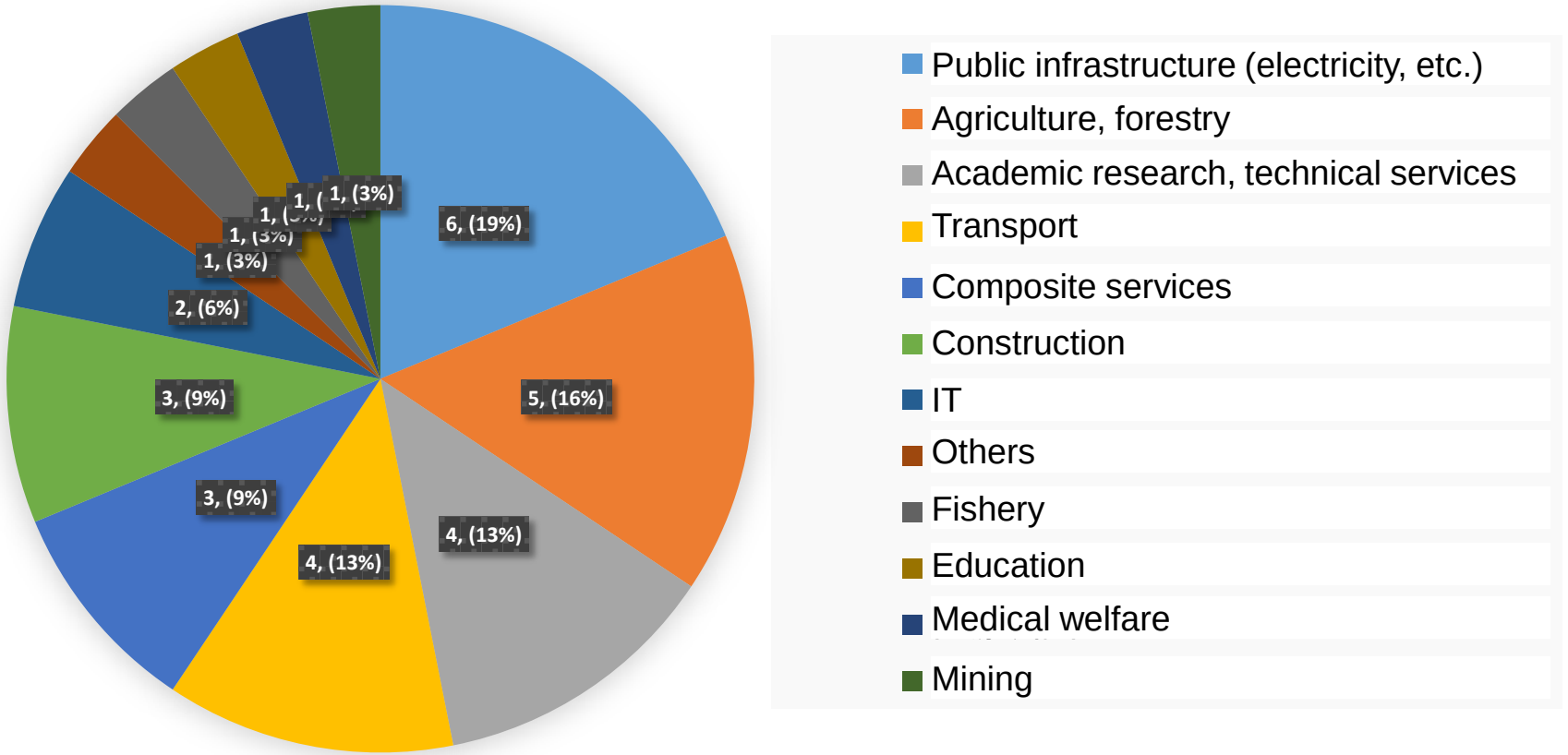
ビジネス

- RESTECを中心としたビジネス推進に向けた体制
- ビジネスコミュニティーとのつながり
- プログラムオフィスを中心としたマネージメント体制

Community Development
~ Approach to potential application developers ~

We communicate with potential application developers.

# of business entities with application ideas, which we contacted



しかし、万事が順調なわけではない

- 利用側と提供側のギャップ

オープンサイエンス

データ共有(Data Sharing)

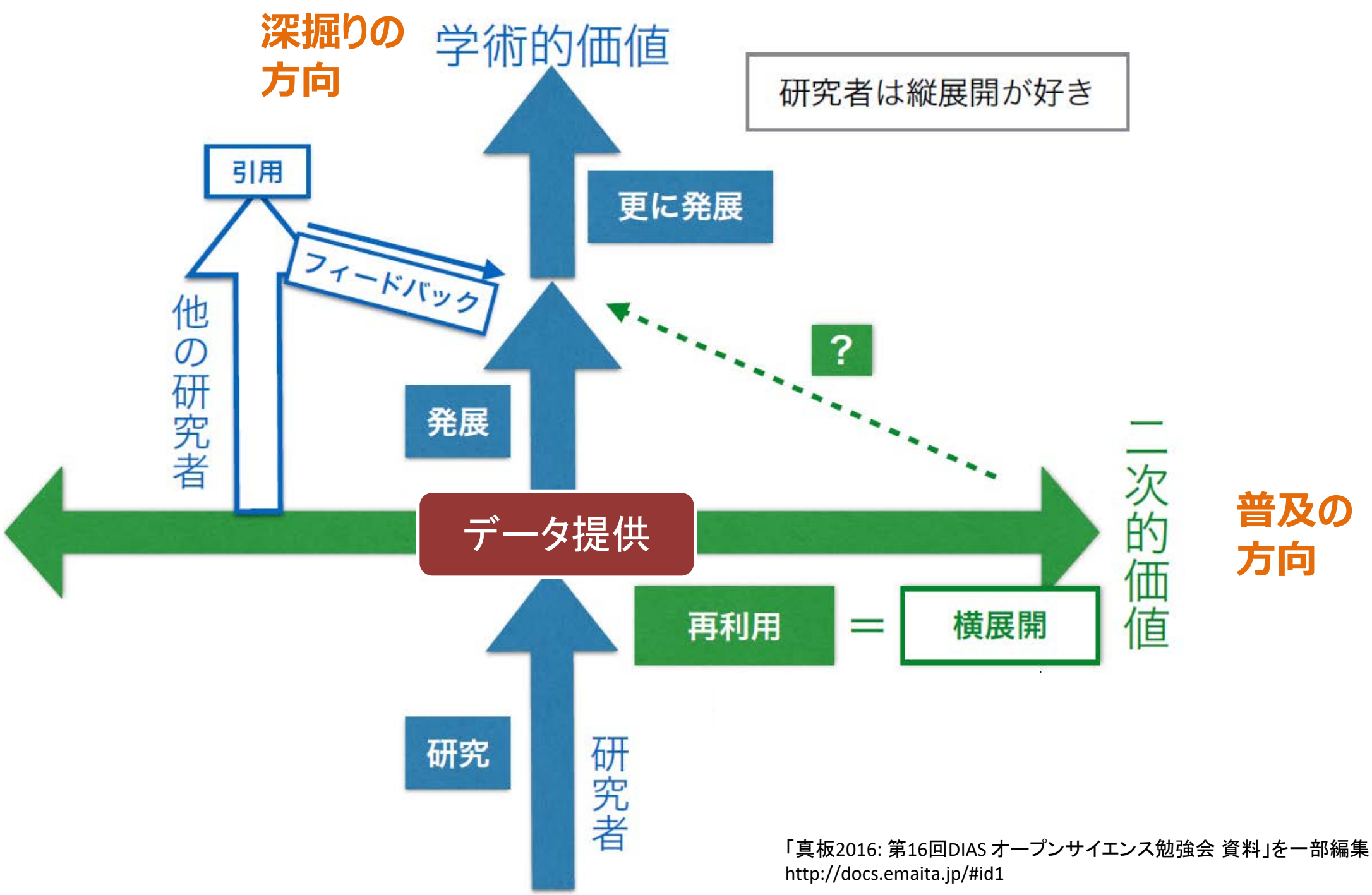
特定の個人・組織間でルールを決めて、
データを提供すること

データ提供者からみると、
データ利用者からの直接的なフィードバックを得やすい

データ公開(Open Data)

特別な制限なしで任意の他者に、
データを提供すること

データ利用者が誰か不明で、どこで使われるのかもわからず、フィードバックを得にくい



「真板2016: 第16回DIAS オープンサイエンス勉強会 資料」を一部編集
http://docs.emaita.jp/#id1

フィードバックという観点から考えると

- DOIやデータサイテーションは、データ提供のインセンティブになるか？

- 被引用回数の増加による効果：
  - 業績が主目的で、研究成果は手段というタイプに対しては、強いインセンティブになる

  - 自分の研究の深化が目的で、業績を気にしない（楽しく研究したい）タイプの場合、実はあまり関係がない
    - むしろ、（数は少なくても）自分と同等以上の能力を持つ研究者からの、直接的なフィードバックのほうがインセンティブになりえる

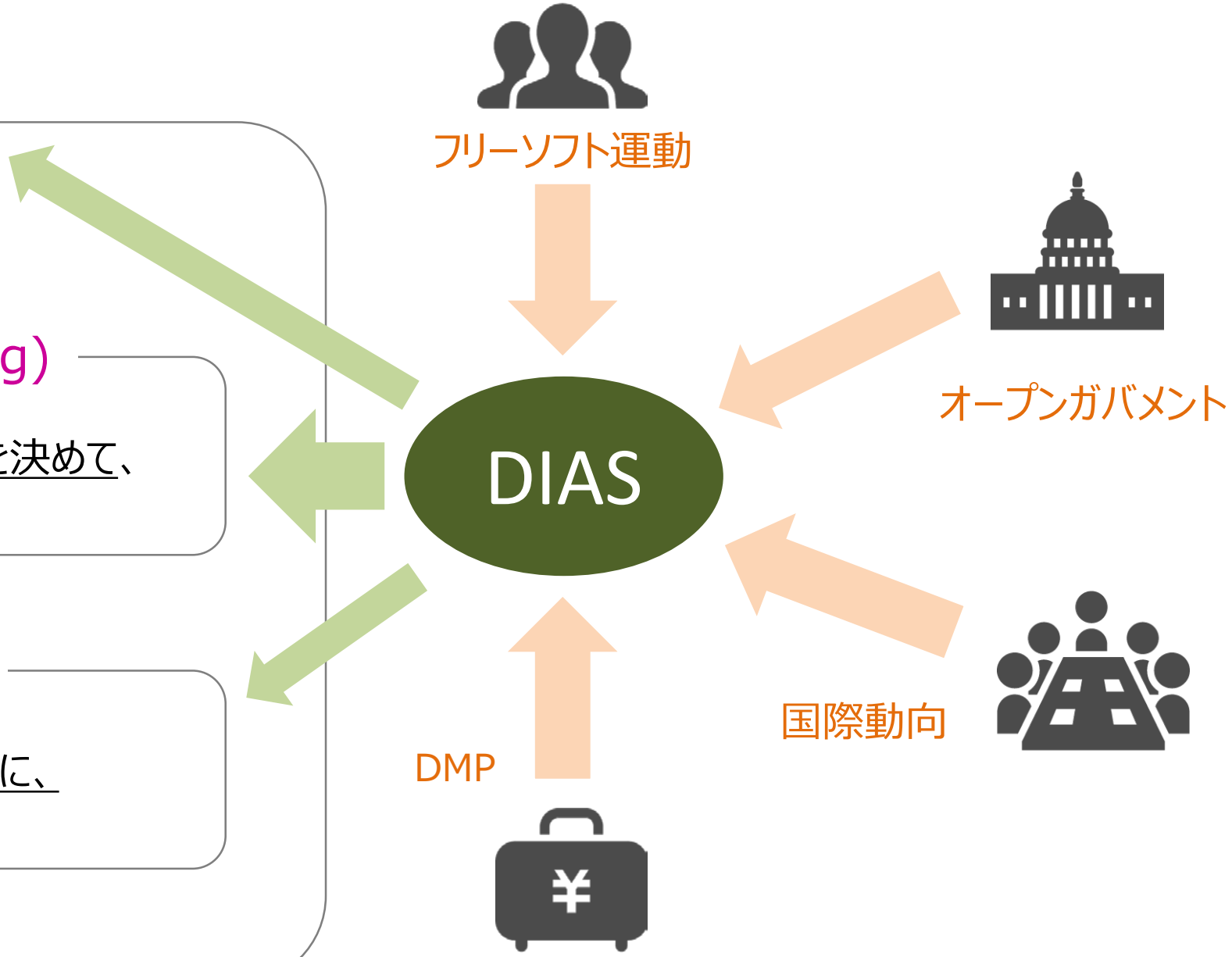
オープンサイエンス

データ共有(Data Sharing)

特定の個人・組織間でルールを決めて、
データを提供すること

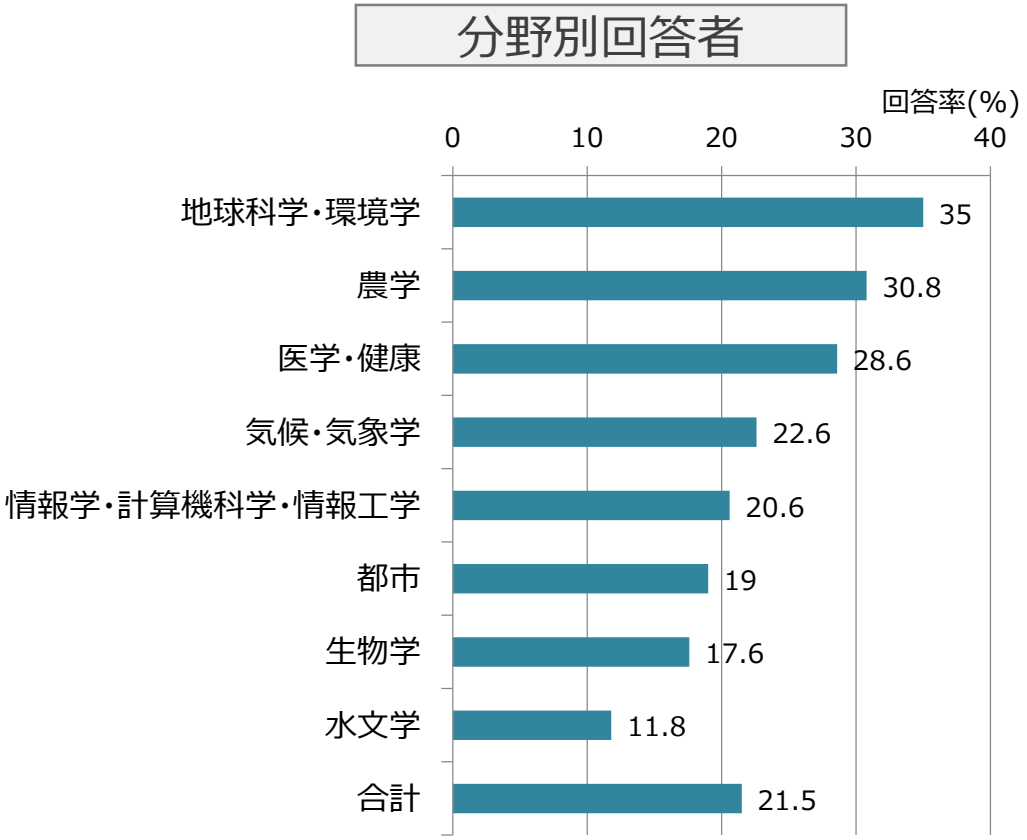
データ公開(Open Data)

特別な制限なしで任意の他者に、
データを提供すること



地球環境情報分野の研究者（DIAS/GRENE-ei コミュニティ）に対して意識調査を実施

- 回答者数: 38/177
- 回答率: 21.5%

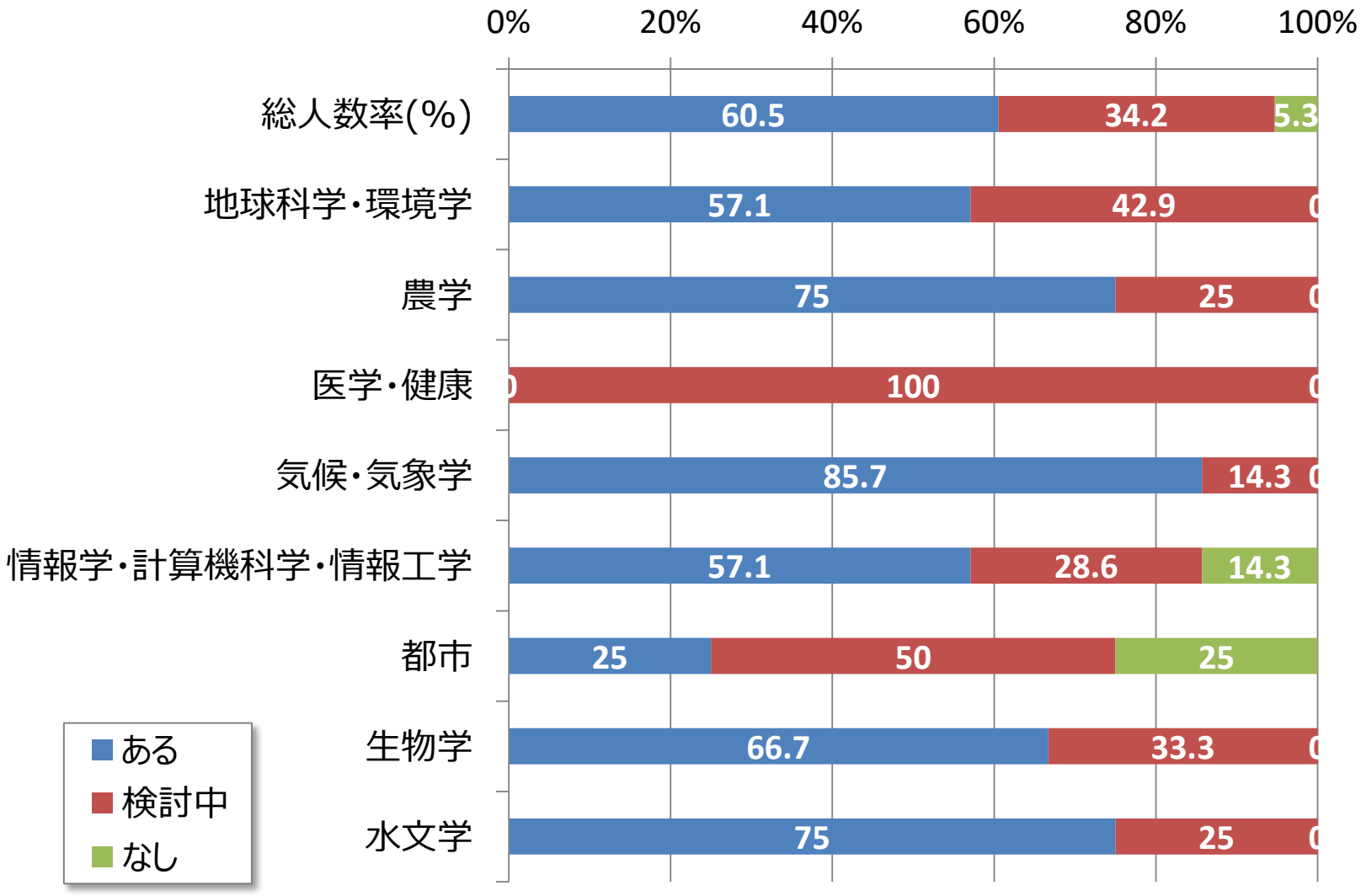


http://doi.org/10.1241/johokanri.59.514

素朴な疑問

- 研究者は、データの提供に反対している？

Q. 現在未提供のデータを、将来的に提供する予定はありますか？

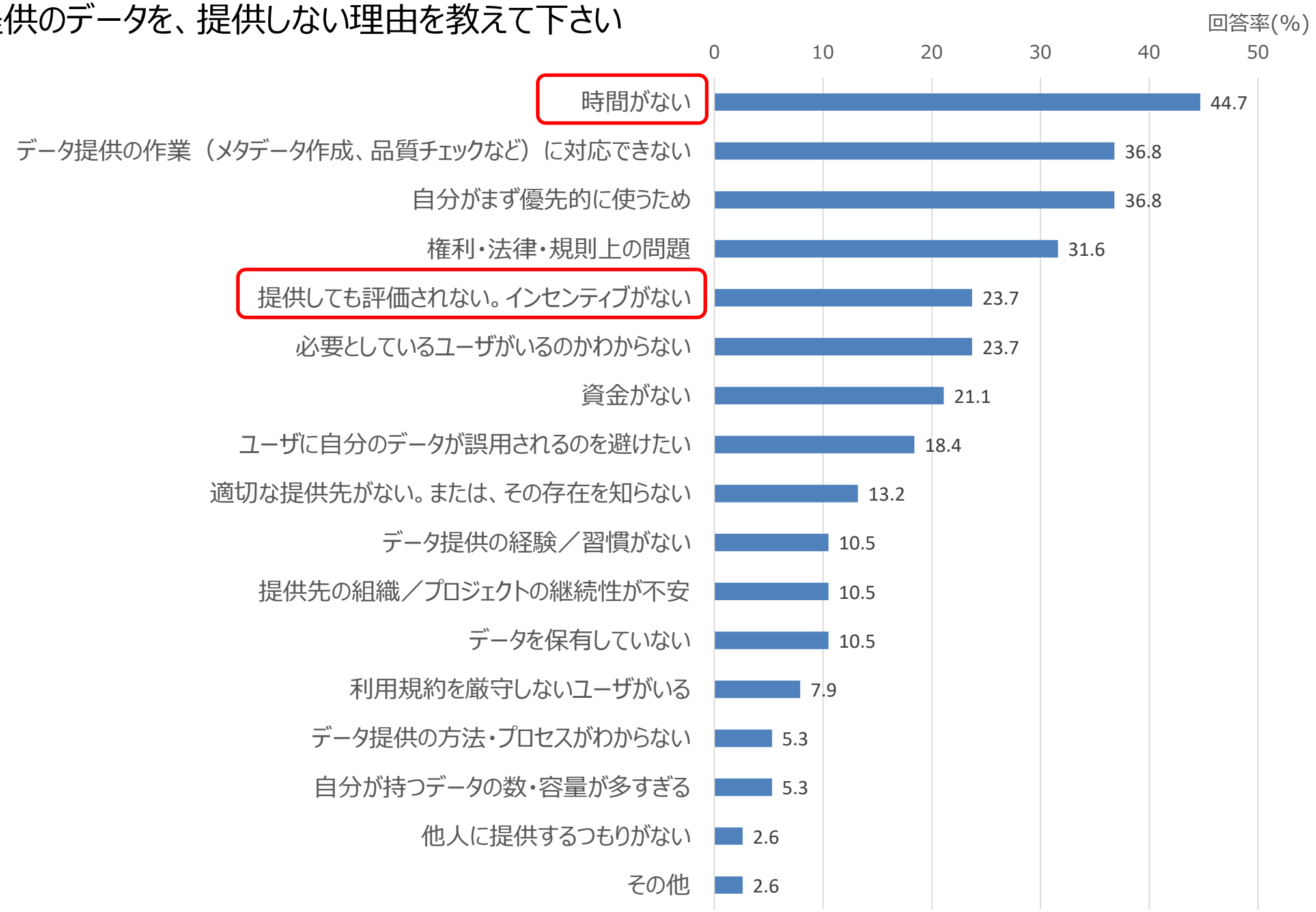


- 94.7%の研究者が、未提供データの提供を考えている

素朴な疑問

- それなら、なぜ未提供のままなの？

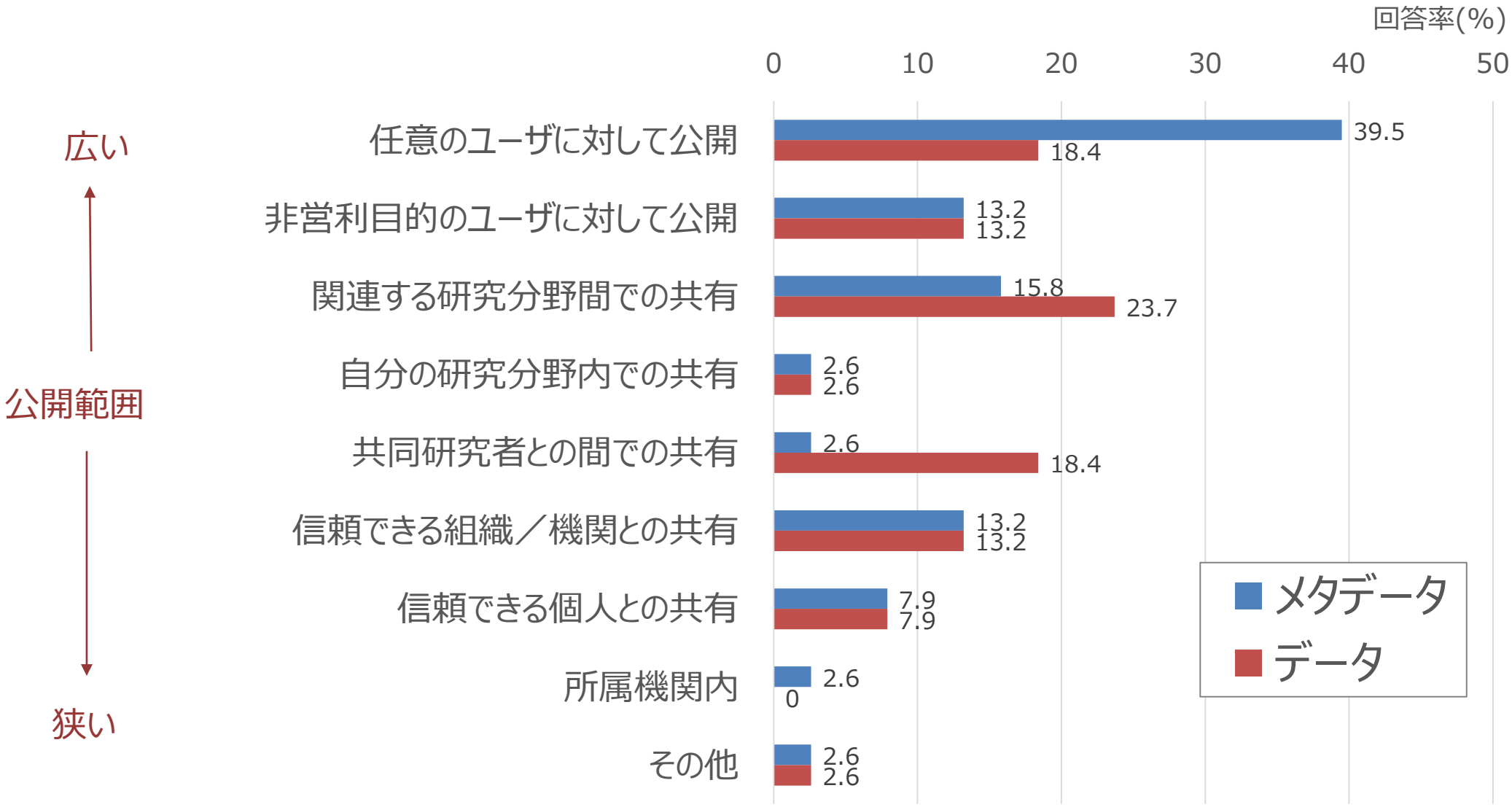
Q. 未提供のデータを、提供しない理由を教えて下さい



素朴な疑問

- 誰になら、提供しても良い？

Q. 賛同できるデータ/メタデータの提供範囲は？

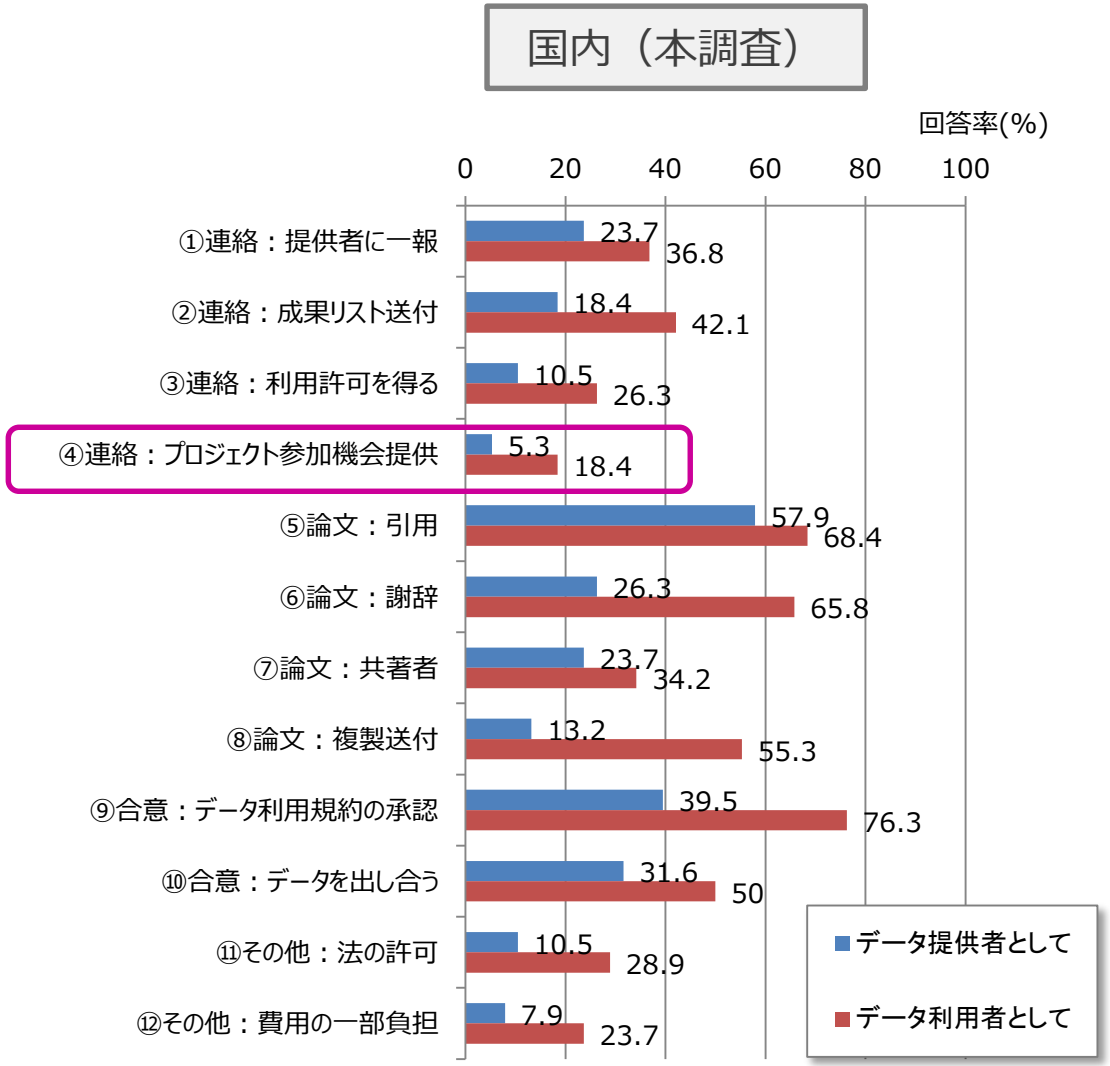


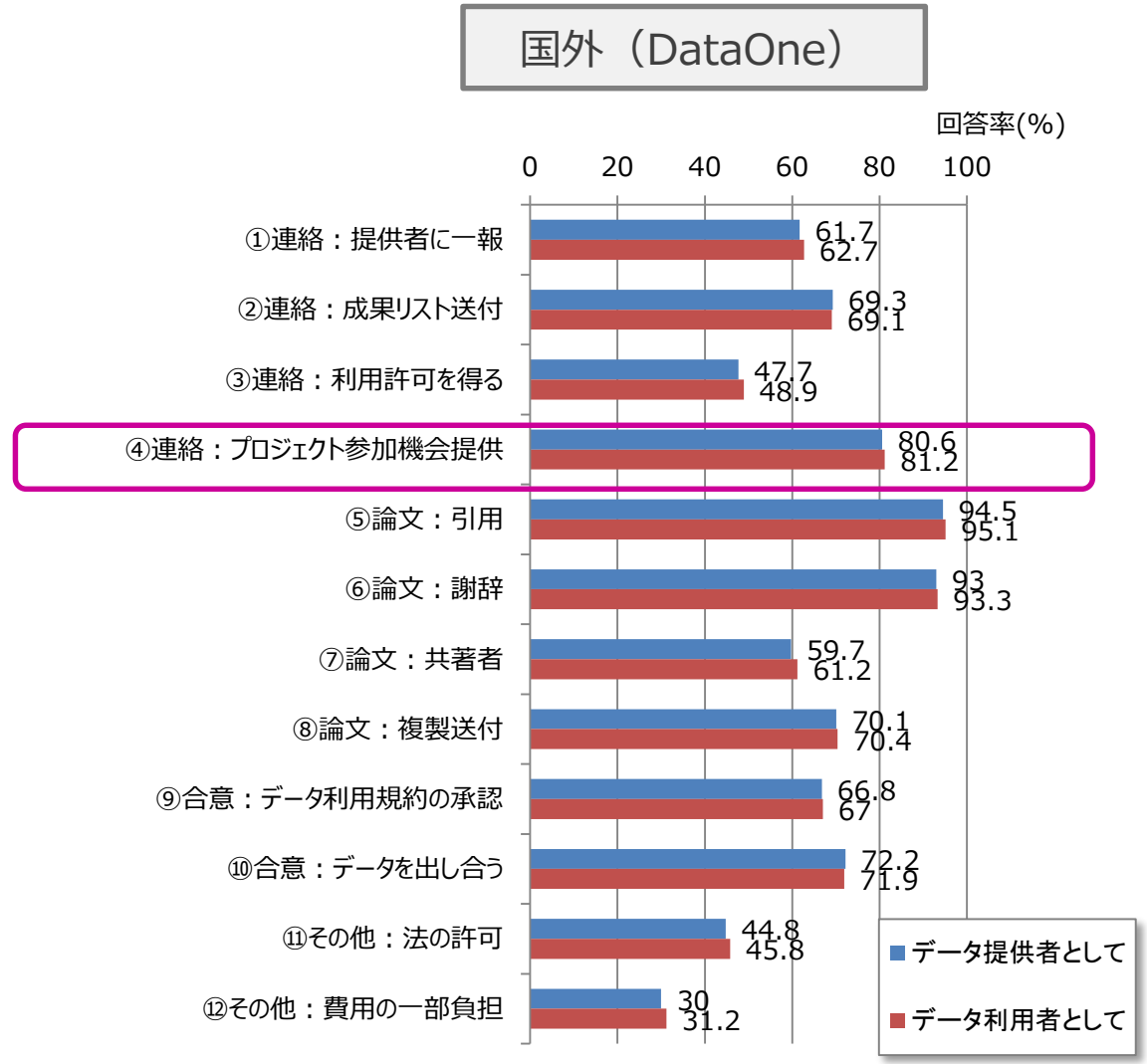
- メタデータに関しては公開しても良いが、
  データに関しては自分の分野に近い研究者と共有したいという声が最も多い

素朴な疑問

- データ共有/データ公開するときの条件は？

Q.データ提供者/利用者として賛同できるデータ共有の条件（国際比較）



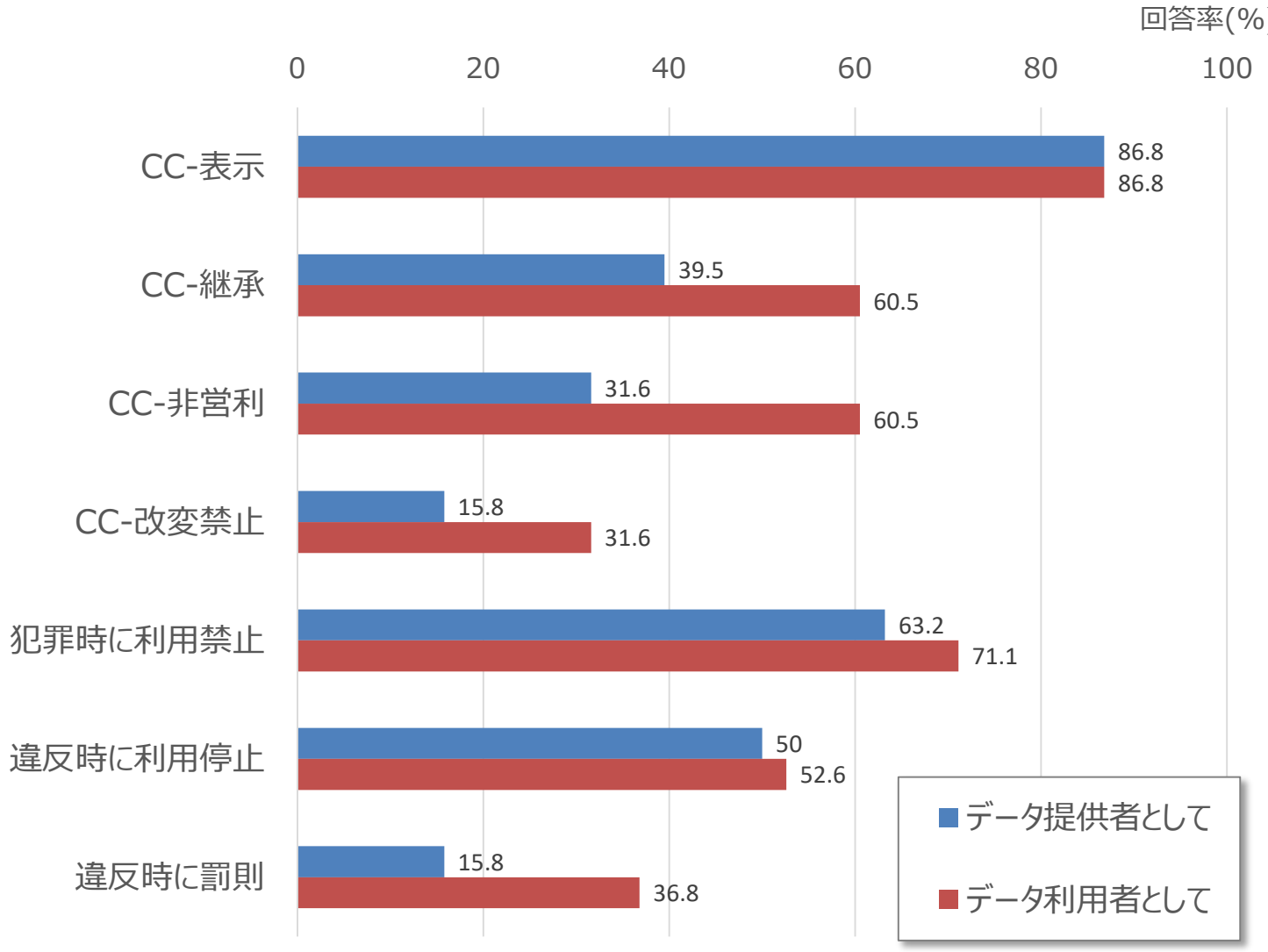


① 国内でも国外でも同様に、論文での引用や、データ利用規約の承認を必要な条件と見ている

② 国内の調査では、データ提供者として厳しい条件を要求せず、むしろデータ利用者として受け入れ可能な条件のほうが多い

③ 国際比較をしたとき、国外では、データ共有をコラボレーションのチャンスとみる一方、国内にその意識はほとんどない

Q.データ提供者/利用者として賛同できるデータ公開の条件は？



CC-表示を、９割近くが条件として挙げている

調査まとめ

Q.研究者は、データの提供に反対している？

A. そうではない。前向きに考えている。

Q. それなら、なぜ未提供のままなの？

A. 時間がないから。実際、作業が大変だし。

Q. 誰になら、提供しても良い？

A. 基本は、関係する分野の研究者。ニーズがあるなら、その他の人に、一部公開しても良い。

Q. データ共有する場合、条件は？

A. データ利用規約に従うこと。論文等で引用してくれること。

Q. データ公開する場合、条件は？

A. CCライセンスなら、CC-表示はマスト。

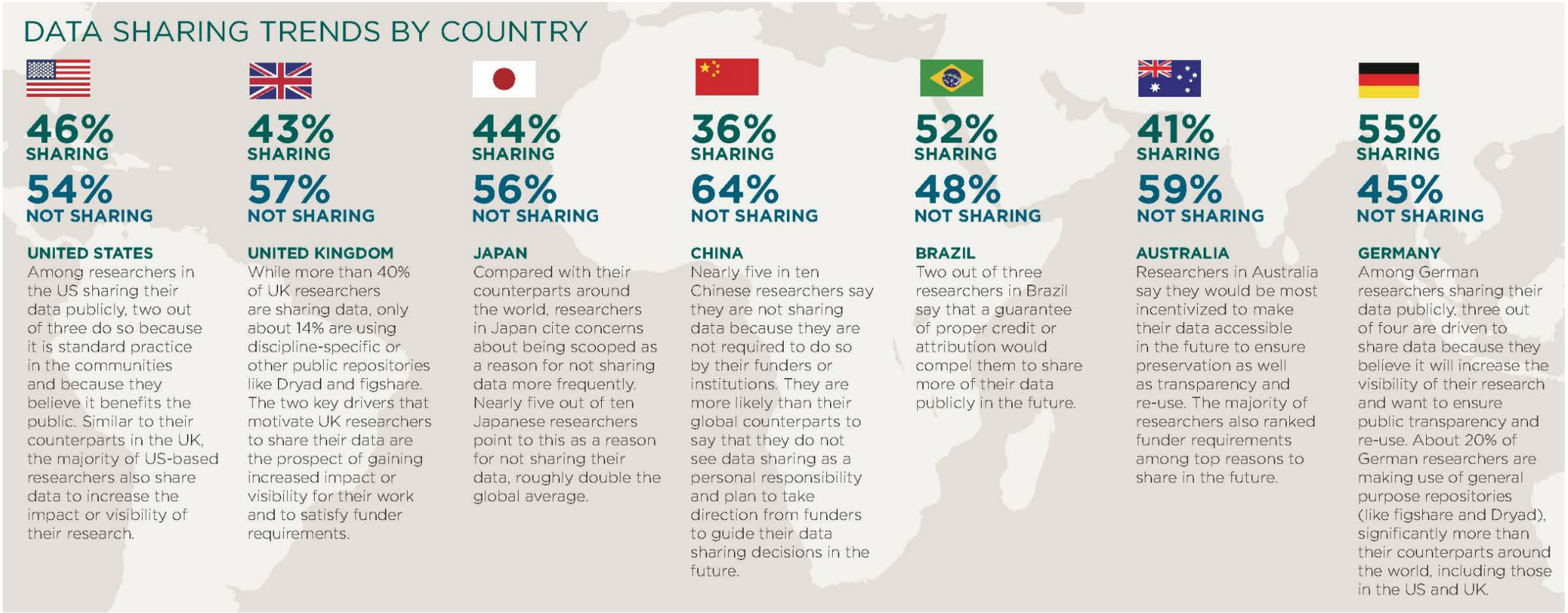
私たちは、これから何をすれば良いのか？

大前提

- 焦って、間違った考え・ルールを採用しないこと
  - 「日本が遅れている」論への疑問
  - 時間が解決することがある
  - チャレンジの継続が重要
  - オープンガバメントの文脈をオープンサイエンスに転用するのは間違い
  - DMPの理想と現実

「日本が遅れている」論への疑問

他国でも、試行錯誤の最中



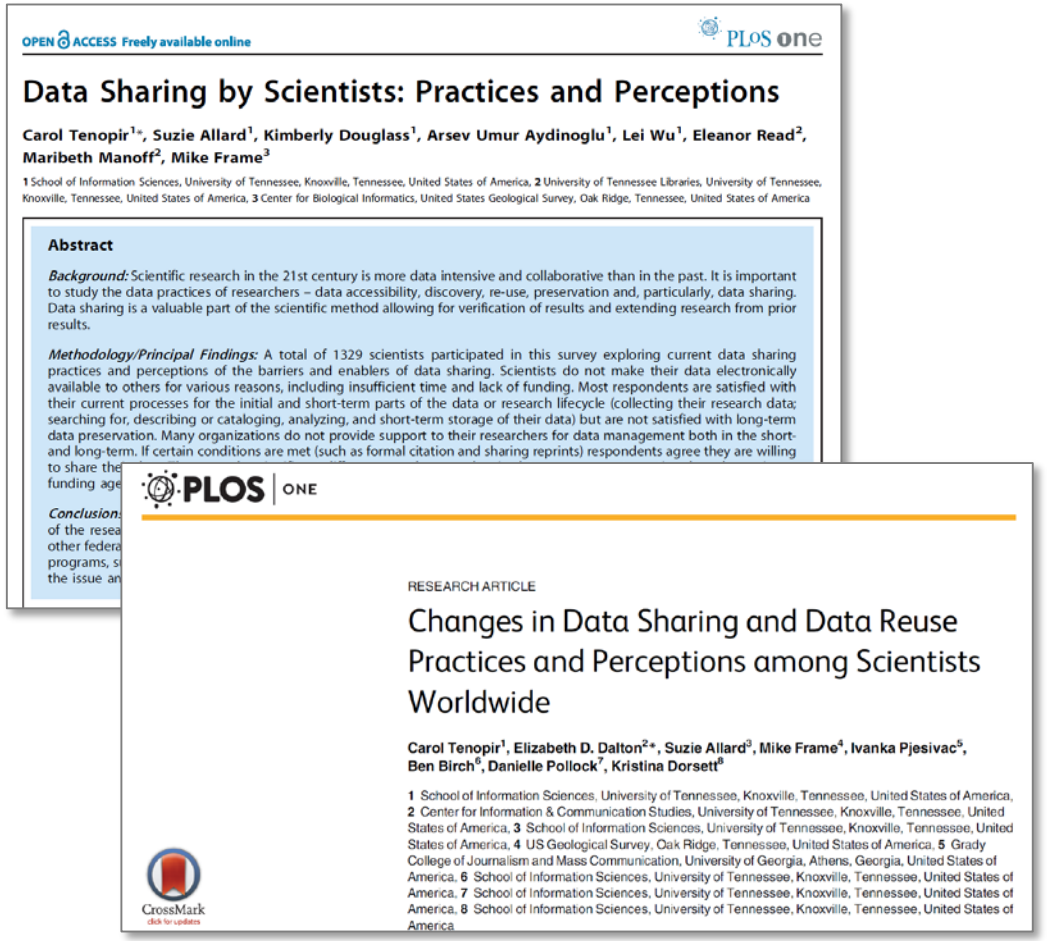
Ferguson 2014. "How and why researchers share data (and why they don't)".

時間/世代により意識は変化する

- DataONEの２つの調査
  - 2009/2010年の基礎調査
  - 2013/2014年のフォローアップ調査

- 数年を経て、
  - データ共有に対してポジティブな見解が増加。
  - 一方、データ共有に伴うリスクへの関心も増加
    - 間違った利用への不安など

- データ共有の障壁は、年代によって異なる
  - 若い世代は業績を気にする
  - シニア世代は資金を気にする



OPEN ACCESS Freely available online

PLoS one

Data Sharing by Scientists: Practices and Perceptions

Carol Tenopir[1]*, Suzie Allard[1], Kimberly Douglass[1], Arsev Umur Aydinoglu[1], Lei Wu[1], Eleanor Read[2], Maribeth Manoff[2], Mike Frame[3]

1 School of Information Sciences, University of Tennessee, Knoxville, Tennessee, United States of America, 2 University of Tennessee Libraries, University of Tennessee, Knoxville, Tennessee, United States of America, 3 Center for Biological Informatics, United States Geological Survey, Oak Ridge, Tennessee, United States of America

Abstract

Background: Scientific research in the 21st century is more data intensive and collaborative than in the past. It is important to study the data practices of researchers – data accessibility, discovery, re-use, preservation and, particularly, data sharing. Data sharing is a valuable part of the scientific method allowing for verification of results and extending research from prior results.

Methodology/Principal Findings: A total of 1329 scientists participated in this survey exploring current data sharing practices and perceptions of the barriers and enablers of data sharing. Scientists do not make their data electronically available to others for various reasons, including insufficient time and lack of funding. Most respondents are satisfied with their current processes for the initial and short-term parts of the data or research lifecycle (collecting their research data; searching for, describing or cataloging, analyzing, and short-term storage of their data) but are not satisfied with long-term data preservation. Many organizations do not provide support to their researchers for data management both in the short- and long-term. If certain conditions are met (such as formal citation and sharing reprints) respondents agree they are willing to share the

funding age

Conclusion

of the resea

other federa

programs, s

the issue ar

PLOS | ONE

RESEARCH ARTICLE

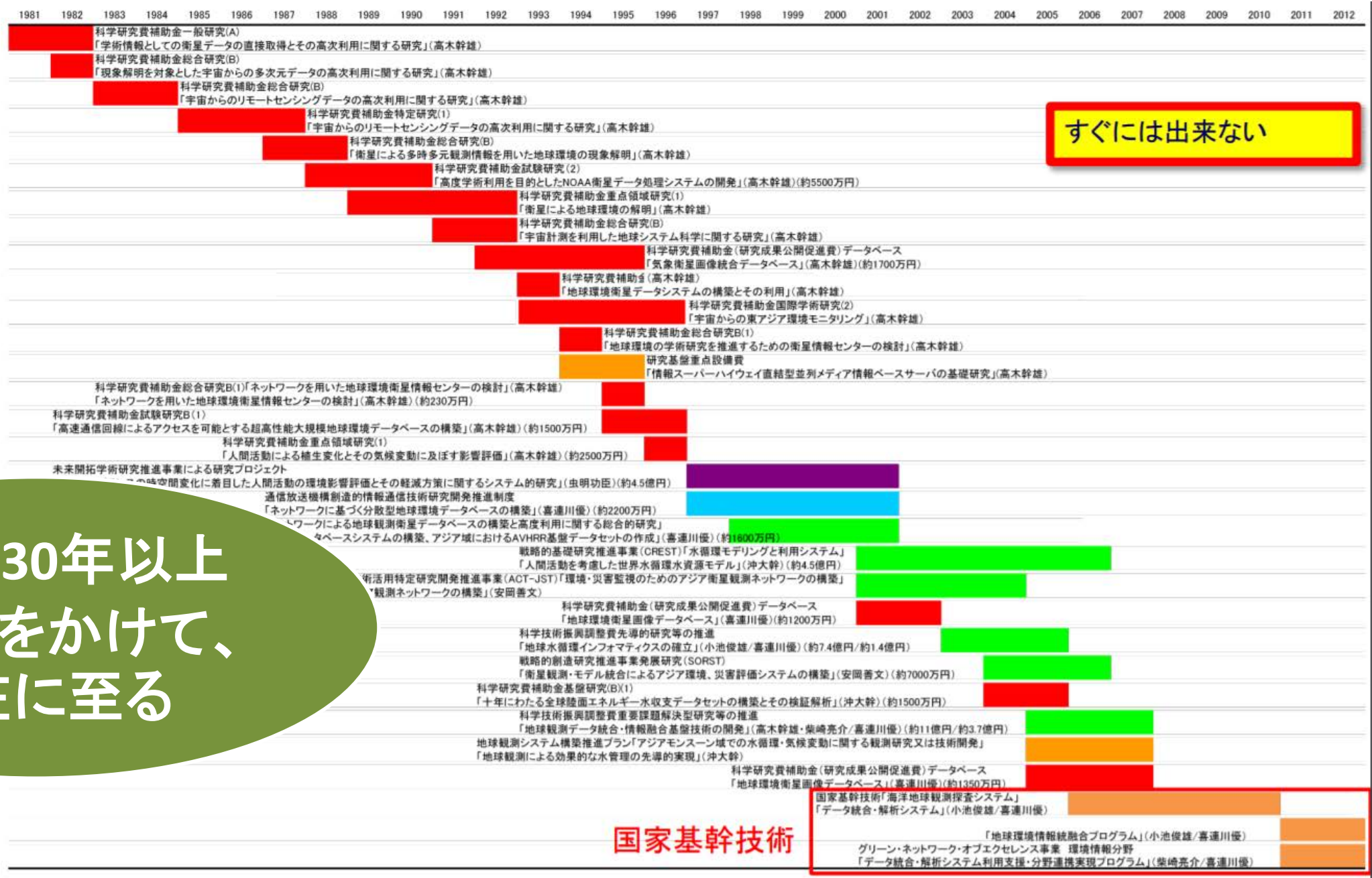
Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide

Carol Tenopir[1], Elizabeth D. Dalton[2]*, Suzie Allard[3], Mike Frame[4], Ivanka Pjesivac[5], Ben Birch[6], Danielle Pollock[7], Kristina Dorsett[8]

1 School of Information Sciences, University of Tennessee, Knoxville, Tennessee, United States of America, 2 Center for Information & Communication Studies, University of Tennessee, Knoxville, Tennessee, United States of America, 3 School of Information Sciences, University of Tennessee, Knoxville, Tennessee, United States of America, 4 US Geological Survey, Oak Ridge, Tennessee, United States of America, 5 Grady College of Journalism and Mass Communication, University of Georgia, Athens, Georgia, United States of America, 6 School of Information Sciences, University of Tennessee, Knoxville, Tennessee, United States of America, 7 School of Information Sciences, University of Tennessee, Knoxville, Tennessee, United States of America, 8 School of Information Sciences, University of Tennessee, Knoxville, Tennessee, United States of America

CrossMark

継続は力



すぐには出来ない

DIASも30年以上の年月をかけて、現在に至る

オープンガバメントの文脈をオープンサイエンスに転用するのは良くない

- オープンガバメント: 公共データのオープン化
  - Open by default
  - 公共データは、一般知識があれば理解できる。
  - ユーザの数が多い。
  - 政府は、市民にデータ公開を要求されたら従う立場（Public servant）

- オープンサイエンス: 研究データのオープン化
  - Ideally open
  - 研究データを扱うには専門知識が必要。
  - ユーザが少ない。
  - 公的資金で作成したデータがオープン？⇒研究者は、Public servantなんですか？

DMP（Data Management Plan）の理想と現実

- 欧米のDMPも、理念の一方で、ドキュメントの中身を見ると、項目や書いてある内容も結構バラバラ。
- DMP義務化の前に、ちゃんと調査して、実効性・有効性を検証したほうがいいのかも。。。

USGS CDR/ECV DMP

Products of research

The USGS will develop science-quality, applications-ready, time-series of key terrestrial variables and produce them on an operational basis using historical, current, and future Landsat observations. The terrestrial variables will follow the guidelines established through the Global Climate Observing System (GCOS) and include Climate Data Records (CDRs) that represent geophysical transformations of Landsat data (e.g., calibrated radiances, inter-calibration of Landsat instrument radiances, surface reflectance and surface temperature), and Essential Climate Variables (ECVs) that represent specific geophysical and biophysical land properties. CDRs and ECVs offer a framework for producing long-term Landsat datasets suited for monitoring, characterizing and understanding land surface change over time (Strategic Plan for Developing Landsat-scale Climate Data Records and Essential Climate Variables).
Therefore, the USGS Land Remote Sensing Program sets the following goals for the next five years:

1. Generate a surface reflectance CDR from calibrated Landsat 4-7 and future Landsat missions as a first step in transforming Landsat data into a time series for terrestrial monitoring.
2. Conduct research to develop and implement a technical approach to develop terrestrial ECV products of dynamic surface water extent (SWE), burned area (BA) and snow covered area (SCA).
3. Link this work to USGS terrestrial monitoring activities involving scientific assessments and decision support.

Nominal CDR production will generate surface reflectance products for every Landsat acquisition over the continental U.S. (CONUS) and Alaska from 1984 to the present. There are approximately 1,176,019 scenes archived at EROS from Landsat 7 Enhanced Thematic Mapper Plus (ETM+) and 603,863 scenes from Landsat 5 Thematic Mapper (TM). The total volume currently estimated for storing the raw Landsat pixel values, surface reflectance values, and ancillary quality information is 4.49 petabytes (PB).

Data format

The surface reflectance CDR derived from Landsats 4-7 Thematic Mapper (TM) and Enhanced Thematic Mapper Plus (ETM+) is produced using Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) software baselined at EROS and currently running under version 2.2.1. Landsat 8 Operational Land Imager (OLI) surface reflectance is generated with a different algorithm to take advantage of that instrument's unique characteristics. All Landsat surface reflectance products share the basic specifications listed below.
Projection: Universal Transverse Mercator (UTM)
Format: Exelis Visualization (ENVI) binary (.img)
Pixel Size: 30-meter (m)
Temporal Coverage Landsat 4 TM: July 1982 to December 1993
Temporal Coverage Landsat 5 TM: March 1984 to May 2012
Temporal Coverage Landsat 7 ETM+: July 1999 to within one week of present
Temporal Coverage Landsat 8 OLI: April 2013 to within one week of present

Data Management Plan

1 Types of data

The project will collect and analyze the following data:

- Conductivity and temperature from moorings and shipboard CTD surveys
- Horizontal currents from Lowered ADCP and moorings.
- Horizontal currents from shipboard sonar
- Fine and micro-scale velocity from the WHOI High Resolution Profiler
- Fine and micro-scale temperature from fast-response thermistors (χpods)

2 Data and Metadata Standards

Data will be shared in matlab MAT file format and/or as netCDF files. Data quality will be in accord with published uncertainty ranges for each instrument and within error bars for standard processing techniques. These PIs have experience with this mix of data types from previous collaborative efforts. Data responsibilities include:

PI	Responsibility
A. Thurnherr	LADCP-CTD analysis.
L. St. Laurent and E. Shroyer	HRP microstructure analysis
S. Jachec	Ongoing model output prediction
J. Moum, J. Nash	χpod microstructure data
M. Alford, J. Nash, J. MacKinnon	Mooring data

3 Data access and sharing

All field data collected under this program will be made available as per NSF guidelines within 2 years of collection via published manuscripts, publicly available final reports to NSF, and data archiving with NODC. Recognizing that any individual PI server may become unavailable over time, data will be made available by PI website locations and also by specific request to any colleague. Models codes to be employed are all public domain. Published peer-reviewed manuscripts will document the simulations and forcing sufficiently.

4 Data archiving and preservation

Aside from the LADCP-shipboard CTD profiles, there are currently no established standards for archiving or data from many of the fine- and micro-scale sensors used in the proposed work. This is a concern of the Climate Process Team on Ocean Mixing, of which many PIs are members. We propose to work with the CPT to evolve formats for data and metadata suitable for archiving both sensor and (critically) model output from the experiment. Field data will be provided to NODC upon project completion. Ultimate archival formats will be determined in consultation with NODC and with the CPT. Adequate archiving is anticipated to be an expensive, time-consuming task. All PIs have included funds for this effort in their budgets.

DATA MANAGEMENT PLAN

We will comply with the open access data policy as described in *Division of Ocean Sciences Data and Sample Policy, NSF 04-004*. All data resulting from this project will be published and made openly available to the research community and the public through several sites. Basic oceanographic (salinity and temperature, chl *a*, DOC) and MET data obtained from the cruise will be archived in the NOAA PMEL database that is set up for our cruise (http://saga.pmel.noaa.gov/data_servers.html), as well as other appropriate data banks including the National Oceanographic Data Center (NODC) and the Biological and Chemical Oceanography Data Management Office (BCO-DMO; http://bco-dmo.org/), within six months after completion of our cruise. Other data will be archived and made available at appropriate sites once data are published or after the completion of this project, either at one of these sites or at the main project website that will be developed and maintained at www.esf.edu/chemistry/kieber/kieber.htm. Aerosol and bubble measurements will be described and made available at the Russell group website at http://aerosols.ucsd.edu/. A metadata listing will also be available through these websites or through other appropriate sites as they become available.

The following Table gives the main data we anticipate collecting during this project, plus the appropriate database where data will be submitted. We will also make use of our websites (see above) for sharing these data with colleagues.

Table 1. Data archiving plan.

Data	Responsible PI, Institution	Database
Temperature and salinity	Kieber, SUNY-ESF	PMEL, NODC, BCO-DMO
Chlorophyll	Kieber, SUNY-ESF	PMEL, NODC, BCO-DMO
MET Data	Keene, UVA	PMEL, NODC, BCO-DMO
Irradiance and attenuation coefficients	Kieber, SUNY-ESF	PMEL, NODC, BCO-DMO
CAA and CHO	Kieber, SUNY-ESF	Kieber website
Dissolved organic carbon (DOC)	Kieber, SUNY-ESF	PMEL, NODC, BCO-DMO
CDOM absorbance spectra	Kieber, SUNY-ESF	PMEL, NODC, BCO-DMO
FTIR organic functional groups	Russell, UCSD	Russell website
SMPS, OPS, APS particle distributions	Russell, UCSD	Russell website
HR-TOF-AMS organic/inorganic mass	Russell, UCSD	Russell website
Bubble size distributions, Surface tension	Russell/Stokes/Deane, UCSD	Russell website
Whitecap coverage	Russell/Stokes/Deane UCSD	Russell website
Ionic composition of bulk aerosol	Keene, UVA	PMEL, NODC, BCO-DMO
Ionic composition of size-resolved aerosol	Keene, UVA	PMEL, NODC, BCO-DMO
Generator operating conditions	Keene, UVA	PMEL, NODC, BCO-DMO

私たちは、これから何をすれば良いのか？

- すぐに結果を出そうと焦らない
- やるべき事を継続
- やること（やったこと）は検証する

- 図書館・情報学の分野では、文献管理で得たノウハウが研究データに転用できるか、悩んでいる人が多いと予想。

- そこで、参考までに、文献の世界とデータの世界を、少し比較してみました

文献と研究データのボリューム感

国立国会図書館

蔵書数（平成26年度）

- 総計: 41,074,863

蔵書数

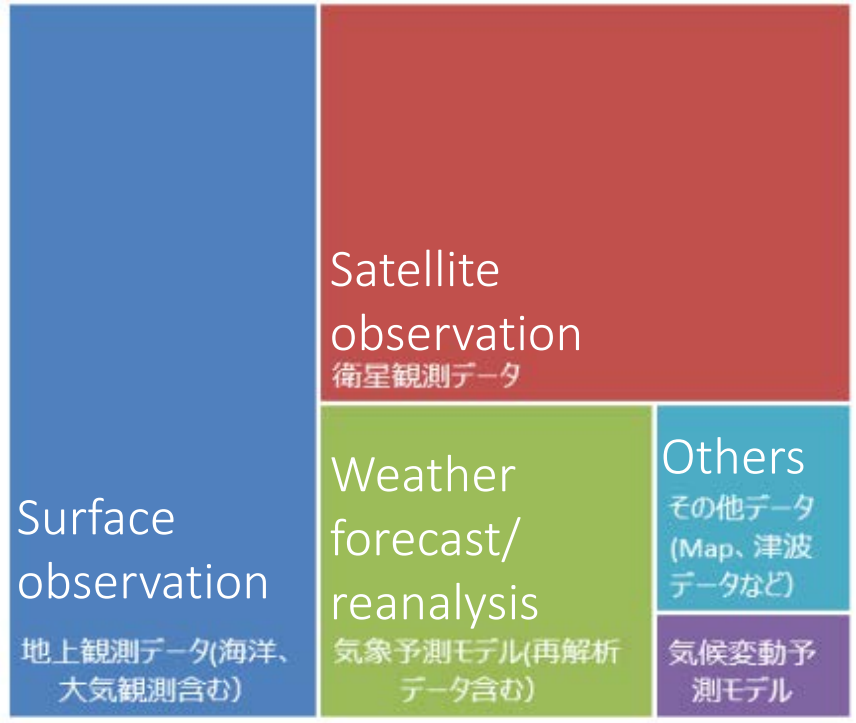
	蔵書数	年間受入点数
総計	4,107万4,863点	85万4,539点
図書	1,053万4,602点	21万19点
雑誌・新聞	1,650万1,384点	57万6,061点
マイクロ資料	910万9,994点	1万7,453点
録音資料	70万5,118点	1万1,909点
光ディスク（CD-ROM等）	12万8,083点	7,325点
地図	55万7,900点	6,144点
博士論文	58万9,696点	2,836点
文書類	37万5,163点	1万433点

http://www.ndl.go.jp/jp/aboutus/outline/numerically.html

DIAS

DIASの公開データ（平成27年度）

- ファイル数（非公開を除く）：58,759,211
- 容量: 約500TB



文献と研究データのサイズ感

世界一大きい本

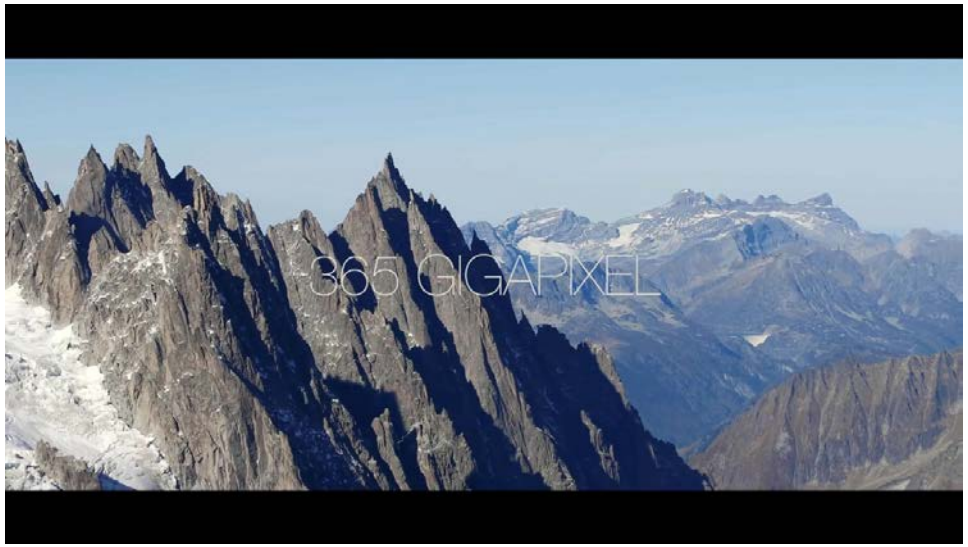
「The Collection, Obama and Pluralism」
5472ページ
厚さ約34cm

http://www.antarafoto.com/peristiwa/v1289293212/buku-tertebal-di-dunia

誰でも（子供も）持ち運び可能

世界一大きい画像データ

「Mont Blanc panorama」
3650億画素
46TB



http://www.telegraph.co.uk/travel/ski/news/365-gigapixel-Mont-Blanc-panorama-becomes-the-worlds-largest-ever-photograph/

普通のPCに入らない
（扱える人が限られる）

文献の著者・読者⇒データ提供者・利用者

- 最近の「オープンサイエンス」の議論は、「オープンアクセス」、「オープンイノベーション」などの抽象論や、「DOI」や「メタデータ」などの技術論が目立つ。
- そこには、「データ提供者」と「データ利用者」という、重要なプレイヤーに対する視点が抜け落ちている

> 図書館は、時代を超越しようとする著者と読者との出会いを支えてゆく重大な役割と責任を果たしていかなければならない
>
> 飯島　昇藏（早稲田大学図書館長）「時代を越えて生きるために ――著者，読者および図書館の責任」

文献の世界の著者・読者の関係が、研究データの世界のデータ提供者・データ利用者の関係。

データ提供者をリスペクトし、データ利用者のニーズを汲みながら、
データ提供者とデータ利用者をつなぐ場をどう作っていくか、そこから考えてみましょう。

まとめ

- 科学の進展には、普及と深化の 2 軸があるが、現在のオープンサイエンスの議論は、普及方向に偏りがち。
- 研究者も、自身のデータの提供について反対しているわけではなく、色々と考えている。
- 焦って、データのオープン化を強要・義務化するのは、良くない。
- 重要なのは、データ提供者とデータ利用者をつなぐ場をどう設計するか、国・FA・情報学・図書館などはそこにどう関わるか、を出発点にすべき。