

鹿児島県大学図書館協議会講演会
2015年10月21日
鹿児島大学附属図書館ライブラリーホール

国際的動向を踏まえた オープンサイエンスの今後の展望と 機関リポジトリの未来

佐藤義則
東北学院大学

1

構成

1. はじめに: CSTI報告書
2. 学術コミュニケーションおよび学術そのものの変化
3. オープンサイエンスに向けた動き
4. データ公開, 共有の課題
5. 図書館および機関リポジトリとの関係

2

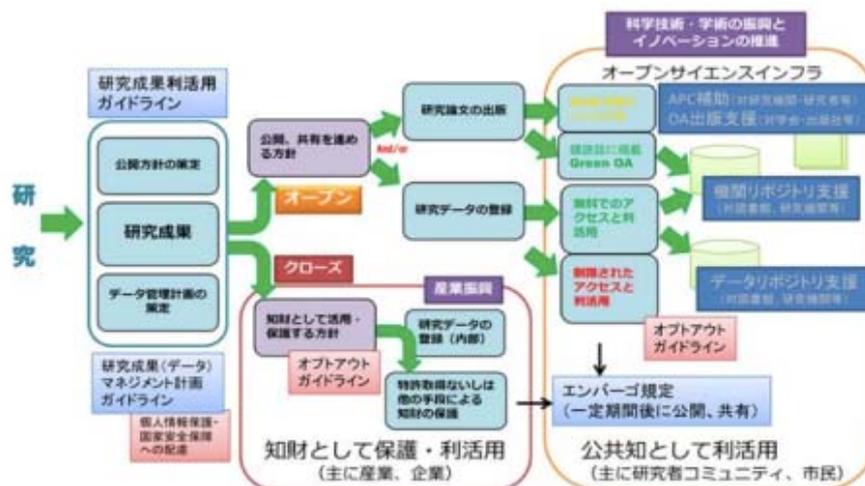
我が国におけるオープンサイエンス 推進のあり方について

～サイエンスの新たな飛躍の時代の幕開け～

2015年3月30日

[内閣府] 国際的動向を踏まえたオープンサイエンスに関する検討会

3



下記図表を参考に和訳、改定

Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 Version 1.0 11 December 2013 p.4
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oe-pilot-guide_en.pdf

国際的動向を踏まえたオープンサイエンスに関する検討会『我が国におけるオープンサイエンス推進のあり方について』内閣府、2015.3.30

<http://www8.cao.go.jp/cstp/sonota/openscience/index.html>

4

オープンサイエンス

- 「オープンアクセスとオープンデータを含む概念」
(p. 3)
- 対象：「公的研究資金による研究成果として得られた論文や研究データ(p. 15)
- 公的研究資金
 - 1) 競争的研究資金及び公募型の研究資金
 - 2) 国費が投入されている独立行政法人及び国立大学法人等の運営費交付金等

論文	原則公開
論文のエビデンスとしての研究データ	
その他研究成果としての研究データ	可能な範囲で公開

5

オープンサイエンス推進における 大学図書館の役割

- 図書館・機関リポジトリがオープンサイエンスの基盤(p. 14, p. 19の図)
- 「大学図書館職員等を中心としたデータ管理体制」(p. 21)
- 「論文や研究データの永続的, 長期的保存を担保するために...大学図書館... 等の協力を得ることが有効」(p. 21)

6

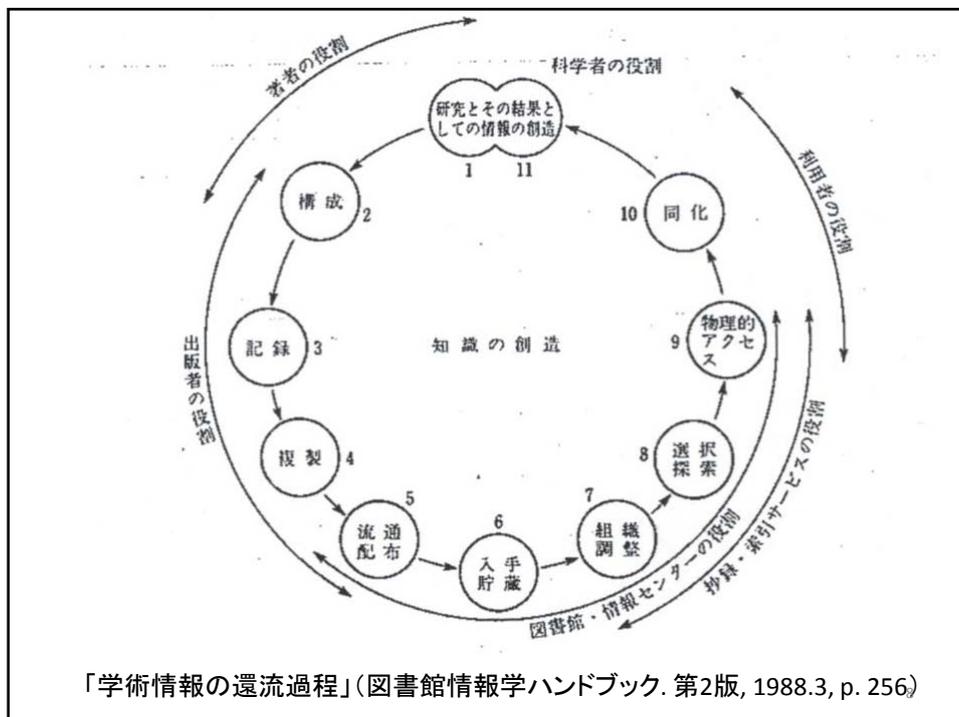
学術コミュニケーションの変化

- 学術コミュニケーションとは

「研究その他の学術的著作が生み出され、品質を評価され、学術コミュニティに広められ、そして将来の利用のために保存されるシステム」

- Association of Research Libraries. "Scholarly Communication." http://www.arl.org/focus-areas/scholarly-communication#.U4GZupR_vUQ

7



研究プロセス全般への関与 (パデュー大学図書館の例)

- ❖ 学術コミュニケーションとは、教育、研究、学問に関わる知識の創成から、流通、保存に至るまでの、研究を遂行し結果を共有するプロセスのことである。
- ❖ パデュー(大学)の図書館は伝統的に、他のたいていの学術図書館と同様に、図書、雑誌、その他の情報資源を収集し、現在および将来の学生、研究者、学者のためのコレクションを構築してきた。しかし、パデューの利用者は、学術コミュニケーション・サイクルの最初(例えば、データ管理やデータ・キュレーション)から、出版(公開)や流通(オープンアクセス・リポジトリや雑誌)を通じて、学術コミュニケーションに関連する追加のサービスを必要としており、パデューの図書館はこうした領域で支援を行うことができる。

出典: Purdue University Libraries. "Scholarly Communication at Purdue."
<https://www.lib.purdue.edu/scholarlyComm>

9

学術コミュニケーションの変化

- 変化の要因
 - － 技術の進展
 - ・ デジタル技術, インターネット, 電子ジャーナル, SNS...
 - － 研究者
 - － 図書館
 - － 政府機関(および資金提供団体)

※技術的要因とともに、政治的、社会経済的要因

※e-Science, e-Scholarship, e-Research, Cyber-Infrastructure

10

学術コミュニケーションの変化と 利害関係者(stakeholders)

- 研究者
 - 量的変化
 - 新たな研究分野の出現, 研究分野の細分化
 - 研究者数の増大(特に, 新興国)
 - 質的变化
 - デジタルな研究環境への進行に伴う一連の活動スタイルの変化
 - 検索, 収集, リーディング, 執筆, 連携, 観察, ノート作成, 翻訳, データ採取等におけるデジタル機器の利用
 - 学術的記録("Scholarly Record")の範囲の広がりおよび量的拡大
- 出版者
 - 1960年代以降の論文量の増大と巨大産業への成長
 - 1980年代以降のM&A(合併・統合)の進行と寡占化
 - 電子ジャーナルおよびビッグディール
- 図書館
 - 電子図書館
 - 機関リポジトリとオープンアクセス
- 政府機関(および資金提供団体)
 - オープンアクセスの義務化(研究資金の効率的運用)

11

農業ICT研究の流れ(農水プロジェクトを中心に)



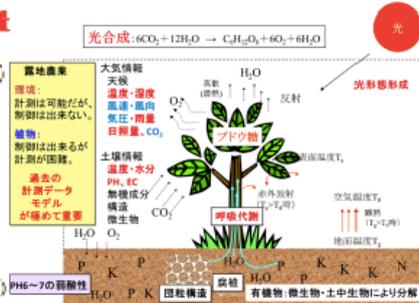
亀岡孝治「食・農分野におけるICTの利活用」

http://www.soumu.go.jp/main_content/000104520.pdf

12

農業センサーネットワークにおける計測項目

- **光**: エネルギー(光合成) **積算日射量**
- **温度**: エネルギー(代謝) **積算温度**
- **光**: 情報(光形態形成、気孔の開閉)
- **温度**: 情報(春化や休眠打破)
- 光合成のための物質吸収
 - **二酸化炭素**(大気中: 400ppm)
気孔(葉)は二酸化炭素の入口
 - **水**(土壌水分: 水ポテンシャル) 根
- 蒸散
 - 植物体温度の上昇を防ぐ
 - 養分吸収
(凝集力による蒸散流(気孔が出口))
- 蒸散作用に影響する要因
 - **温度**: ある温度での飽差(飽和蒸気圧と実際の蒸気圧との差)が大きいほど(乾燥しているほど)蒸散量は、蒸散作用は活発になる。
 - **風**: 風は蒸発を促進。気孔を通して蒸発した水蒸気を風が吹き払い、乾燥した空気を送り込むため、蒸散する量は風速の平方根に比例して増加。
 - **温度**: 気温が高くなるほど蒸散する量は、単位水蒸気量が大きい。
- 土壌
 - 物理的な視点。水分、水ポテンシャル。団粒構造
 - 化学的な視点。pH、土中ミネラル
 - 生物学的な視点。土壌微生物



亀岡孝治「食・農分野におけるICTの利活用」
http://www.soumu.go.jp/main_content/000104520.pdf

13

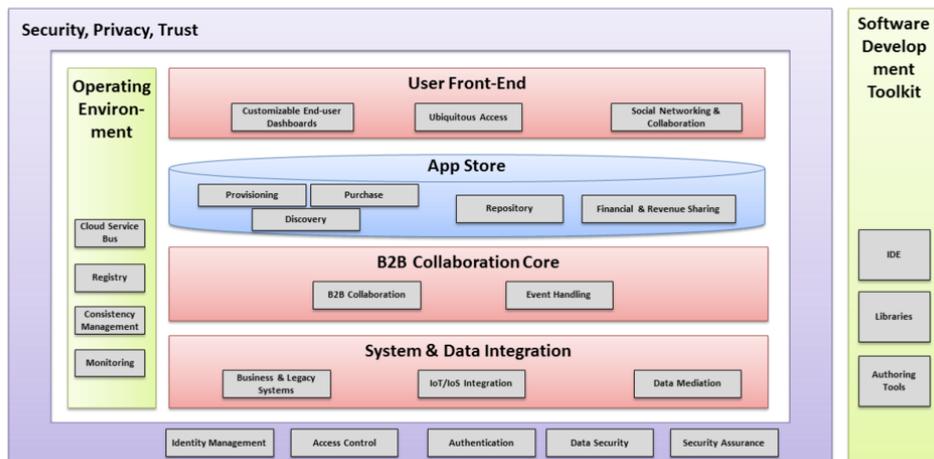
農業データに関する取り組み

- FAO (Food and Agriculture Organization of the United Nations)
 - Agrovoc: 農業オントロジー
 - VocBench: ウェブベースの、SKOS (Simple Knowledge Organization System)を用いたシソーラス、典拠リストの管理のための編集ツール
- EU
 - Finspace
 - FP7/2007-2013
 - FI-PPP (Future-Internet Public Private Partnership)

14

FIspace Modules

Seven major building blocks (called modules) constitute the FIspace platform. They are visualized in the figure below and briefly introduced in the text that follows.



<http://www.fispace.eu/>

15

学術論文へのオープンアクセス

- 米国2014年統合予算法修正案
 - 米国教育省、労働省、健康・人的サービス省に属し、年間の研究予算が1億ドルを超える機関に対し、パブリックアクセス方針を定めるよう要求(エンバーゴ期間: 12カ月)
- 欧州委員会 "Horizon 2020"
 - 同プログラム下の助成研究の成果に関連した査読論文は、1) 6ヶ月以内(社会科学, 人文学では12ヶ月以内)にセルフ・アーカイビング, 2) OAジャーナルまたはハイブリッドジャーナルに掲載, のいずれかの方法で公開することを要求
- フィンチレポート
 - ゴールドOA(ハイブリッドとpure OAジャーナルの両方)の重視 – 批判と影響
 - JISC APC
- SCOAP³
 - 図書館が従来「購読料」として支払っていたものを対象雑誌の「出版料」に振替えることで、世界中の誰もが無料でそれらの学術雑誌の論文を読むことができるオープンアクセス化の実現を指向

16

研究データ管理とオープンアクセス

- 研究データ共有
 - 「データ共有とは、他者が利用できるよう研究データを公開すること」(Borgman, 2012)
- 背景:
 - デジタル技術による、データの保存, 共有, 再利用の可能性の向上
 - データの量的増大

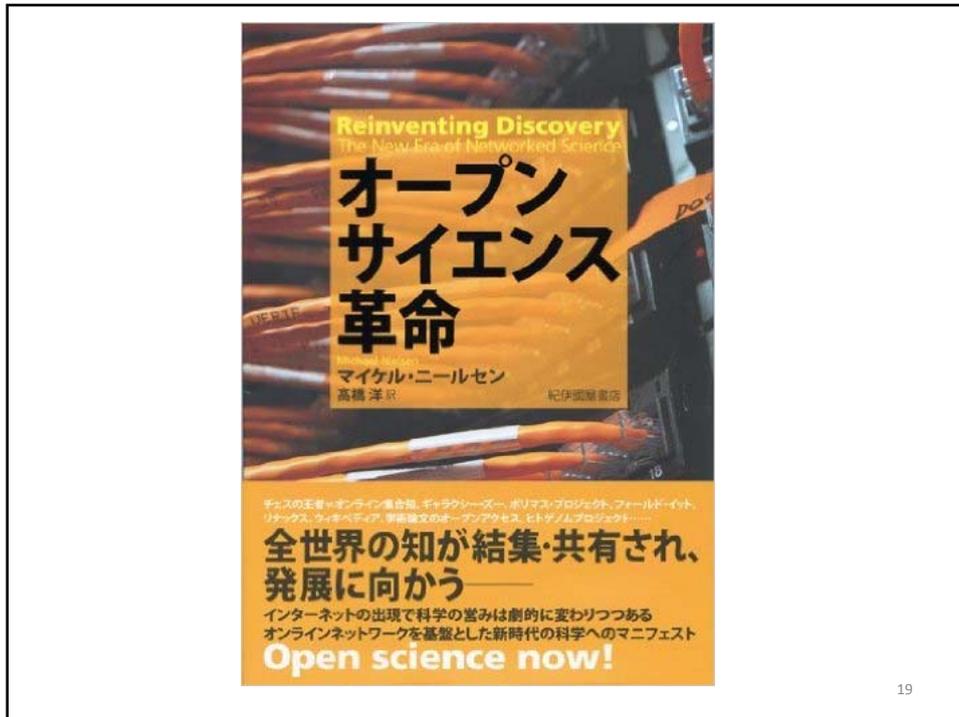
17

研究データの共有を後押しする原理

- 公的資金を受けた研究成果の還元
- 研究の再現または検証 (e.g. 「STAP細胞」問題)
- 既存のデータを用いた他者による新たな課題の研究 (メタ分析, マッシュアップ, 大量データのマイニング等)
- 研究と革新の進展: 「第四のパラダイム」としての計算科学による一連の新たな方法等

Borgman, Christine L. "The conundrum of sharing research data," *Journal of the American Society for Information Science and Technology*, 2012, vol.63, no.6, p.1059-1078.

18



19

研究データへのオープンアクセスの始まり

- World Data Center (WDC) system (1958 -)
 - 1957～1958年の国際地球物理学年の観測プログラムで収集されたデータのアーカイブと提供のために、ICSU(国際科学会議)によって設立
- CODATA (Committee on Data for Science and Technology; 科学技術データ委員会)(1966 -)
 - ICSU(国際科学会議)によって設立
 - 基礎定数についてのCODATAタスクグループ(1969 -)
 - 国際的に認められた基礎物理定数および関連した変換係数の組を定期的に提供

20

- ウェルカム・トラスト(Wellcome Trust)
 - 1997年, ゲノム配列に関する助成研究によって得られたデータ公開の要求を開始

- 米国国立衛生研究所(NIH)
 - 2003年2月, 研究データ共有に関する声明を発表。
 - ただし, 単年度の補助金申請額が50万ドルを超える申請者に対してのみ, 契約担当者との交渉を要求

- OECD 公的資金による研究データへのアクセスに関する原則及びガイドライン(OECD principles and guidelines for access to research data from public funding)
 - 2007年
 - 「公的資金を受けた研究のデータの共有とオープンアクセスは, 新たなデジタル技術とネットワークによる研究の可能性の最大化に役立つだけでなく, 研究への公的投資から得られる利益をより大きくする」

21

国際的な研究データ共有の例

- ヒトゲノム計画(Human Genome Project)
 - 人ゲノムの20,000以上の遺伝子配列を特定し染色体上にマッピングし, そしてDNAを構成する化学塩基の組み合わせ配列を明らかにする国際的取り組み。1990年に開始され, 計画は2003年に完了。世界中のさまざまな場所のDNA配列を保存する分散型データベースであるGenBankの構築
- タンパク質構造データバンク(Protein Data Bank), GEON(地球科学), 天文学における総観天空探査(Sloan Digital Sky Survey等)ほか

※omics(オーミクス;生物データの「すべて」を集約し, 解析する研究領域); ゲノミクス(genomics), トランスクリプトミクス(transcriptomics), プロテオミクス(proteomics;プロテオーム(ゲノムの各遺伝子に対応するすべてのタンパク質)の構造と機能に関する大規模研究), メタボロミクス(metabolomics; 細胞の活動によって生じる特異的な分子の網羅的な解析)等の総称

22

研究データ管理計画の要求

- 2011.2 NSF すべての研究資金申請における「データ管理計画 (Data management plan)」の提出要求
 - あくまでデータ管理計画であってデータ共有計画ではないが、共有を強く求めるとともに計画もピアレビューの対象とされているため強い影響力
 - 2003.2 NIH 研究データの共有に関する声明(単年度の補助金申請額が50万ドルを超える申請者に契約担当者との交渉を要求)
 - NSFにおいてもNIHよりも早い時期から資金提供契約でデータ共有が奨励されてきたが、強制力を伴わなかった
- 2011.6 全米人文科学基金 (National Endowment for Humanities) のデジタル人文学部門 NSFと同様の要求を発表
- 2013.2 米国大統領府科学技術政策局 (Office of Science and Technology Policy, Executive Office of the President) 研究開発支出額が年間100万ドルを超す政府機関に対し6ヶ月以内に連邦予算による研究成果(ピアレビュー出版物と機密研究以外の研究データ)へのパブリックアクセスを拡大するための計画案を提出するよう命令
- 2013.5 政府情報のオープンデータ化を義務付ける大統領令

23

米国大学図書館関連の経緯

- Atkinsレポート (2003)
 - DARPA, NSF, NLM等による電子図書館の拡張の提言: アクセスと長期保存の提供(当初の10億円規模から30億円規模へ)
- PITAC [the President's Information Technology Advisory Committee]レポート(2005) *Computational Science: Ensuring America's Competitiveness*
- PITACレポートと同時期に、米国議会が連邦の資金提供機関に対して、研究の高コスト(データの重複)に対する疑問を提示
- ARL E-Science タスクフォース 開始(2006)
 - その後、現在進行中のE-Science ワーキンググループに

参考文献: James L. Mullins. "The policy and institutional framework." in J. M. Ray ed. *Research Data Management*. Purdue Univ. Press, 2013, pp. 25 – 44. ²⁴

米国大学図書館関連の経緯

- ARLによるレポート(2006) *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering*
 - データ管理における大学図書館・研究図書館の役割(サイバーインフラ, 計算機科学, あるいはeサイエンスの領域における連携を担う)の提言
 - 英国はじめ欧州へも影響
 - 三つの要素
 - データ管理に必要な技術的, 組織的能力の理解, モデル化, プロトタイプ形成のための研究開発
 - データ科学における新たな要員の開発のための訓練と教育プログラムの支援
 - 研究活動に変化をもたらす教育的取り組みの開発, 支援, 促進

25

米国大学図書館関連の経緯

- NSFによる提案招請(2007)「図書館とアーカイブの科学, サイバーインフラ, コンピュータと情報の科学, 各研究領域の専門知識の統合」
 - 予算額 100億円, 20億円×5プロジェクト
 - 広範な分野を対象とした連携を前提, 特に図書館情報学分野に焦点,
 - 大学図書館はコレクション整備, 貴重書のデジタル化等の資金提供を受けてきたが科学研究の重要なパートナーとしてのこれ程の規模は初めて
- 2008年夏, 2つのグラントが決定
 - DataONE(Data Observation Network for Earth, ニューメキシコ大学)
 - DC(Data Conservancy, ジョンズ・ホプキンス大学)

26

米国大学図書館関連の経緯

- IMLS (米国博物館・図書館サービス機構) による資金提供 (2007, 2010)
 - Data Curation Profiles Project
 - パデュー大学図書館とイリノイ大学アーバナ・シャンペーン校図書館情報学大学院の統合プロジェクトとして開始
 - 専門分野研究者のデータセットに対する考え方, 共有の状況, 共有の条件等について探求
- ARL E-Science インスティテュート (第1回) の開催
 - 図書館員を教育し, メンバー館の間の連携を生み出させるレベルまで引き上げる協調努力の促進
 - Webinar, 電話会議, イベント
 - 開催にあたり, ARL加盟館に支援金1館あたり5,000ドルを要請
- ARL SHARE (2013.6 -)
 - 米国大学協会 (AAU), 公立ランドグラント大学協会 (APLU), ARLによる、公的助成研究成果の保存, アクセス, 再利用の確保に向けた連携事業

27



UW, Berkeley, NYU collaborate in \$37.8M data science initiative

Posted by Anonymous | November 07, 2013

The University of Washington, the University of California, Berkeley, and New York University are partners in a new five-year, \$37.8 million award from the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation whose goal is to dramatically accelerate the growth of data-intensive discovery in a broad range of fields. [Watch a video of the announcement.](#)

The UW team, which includes more than a dozen faculty from across the campus, is led by [Ed Lazowska](#), Bill & Melinda Gates Chair in Computer Science & Engineering and Director of the UW eScience Institute. Berkeley's team is led by Nobel laureate astrophysicist Saul Perlmutter, and NYU's by neuroscientist and computer scientist Yann LeCun.

"All across our campus, the process of discovery will increasingly rely on researchers' ability to extract knowledge from vast amounts of data," Lazowska said. "In order to remain at the forefront, UW must be a leader in advancing the methodologies of data science, and in putting these methodologies to work in the broadest



UW core team (clockwise from lower left): Tom Daniel

28

<http://escience.washington.edu/blog/uw-berkeley-nyu-collaborate-378m-data-science-initiative>

国際的な取り組み

- G8 オープンデータ憲章 (G8 Science Ministers Statement [on Open Scientific Research Data]. 2013.6)
 - 原則1: 原則としてのオープンデータ
 - 原則2: 質と量
 - 原則3: 全ての者が利用できる
 - 原則4: 改善されたガバナンスのためのデータの公表
 - 原則5: 技術革新のためのデータの公表

<http://www.kantei.go.jp/jp/singi/it2/densi/dai4/sankou8.pdf>
- RDA (Research Data Alliance)
 - 2013年3月発足
 - “Data sharing without barriers”

29

機関リポジトリとは-1

- 単独あるいは複数の大学コミュニティの知的生産物を入手し保存する電子的コレクション
- 学術コミュニケーション改造の中心的構成要素。分散型出版構造での革新を促す
- 学術機関の質の具体的指標。機関の可視性、名声、価値を高める

Raym Crow, “The Case for Institutional Repositories: A SPARC Position Paper.”
ARL Bimonthly Report 223 (2002).
<http://www.arl.org/bm~doc/instrepo.pdf> ; (翻訳: 栗山正光「機関リポジトリ擁護論」http://www.tokiwa.ac.jp/~mtkuri/translations/case_for_ir_jptr.html)

30

機関リポジトリとは-2

- “機関およびそのコミュニティの構成員によって作成されたデジタル資料の管理運営と公開・提供・保存を、機関が責任を持って行う一連のサービス”
- 大学という機関においては、デジタル資料とは教員の研究成果や学生の学位論文、データセットを指す

Clifford A. Lynch, “Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age.” (2003)
<http://www.arl.org/newsitr/226/ir.html>

31

機関リポジトリに対する異なる視点



- 学術出版の変革
 - Open Access
- 機関内で生産される情報の収集と保存
- 多様な情報源のリンク

- 論文のリポジトリ
- デジタルリポジトリ
- eScience, eScholarship指向のリポジトリ

32

英国におけるデータポリシー

- RCUK「データポリシーに関する共通原則(Common Principles on Research Data Policy), 2011」: できる限り制約なしに適時のかつ責任ある方法でデータの公開を強く要求

● Full Coverage ● Partial Coverage ○ No Coverage

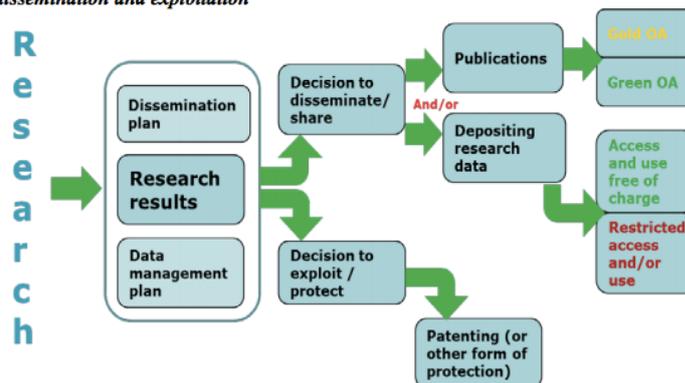
Research Funders	Policy Coverage			Policy Stipulations				Support Provided			
	Published outputs	Data	Time limits	Data plan	Access/sharing	Long-term curation	Monitoring	Guidance	Repository	Data centre	Costs
AHRC	●	●	●	●	●	●	○	●	○	●	●
BBSRC	●	●	●	●	●	●	●	●	●	●	●
CRUK	●	●	●	●	●	●	●	●	●	○	○
EPSRC	●	●	●	●	●	●	●	●	○	○	●
ESRC	●	●	●	●	●	●	●	●	●	●	●
MRC	●	●	●	●	●	●	○	●	●	○	●
NERC	●	●	●	●	●	●	●	●	●	●	●
STFC	●	●	●	●	●	●	●	●	●	●	●
Wellcome Trust	●	●	●	●	●	●	●	●	●	●	●

英国の資金提供団体の論文と研究データの公開方針. Digital Curation Centre.

“Overview of funders' data policies.” <http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies>

欧州委員会 “Horizon 2020” におけるオープンアクセス

Graph: Open access to scientific publication and research data in the wider context of dissemination and exploitation



より広い範囲の流通および展開の文脈における科学出版物と研究データへのオープンアクセス. European Commission. *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020*. Ver. 1.0 2013.12, 14 p.

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-³⁴

学術雑誌における研究データへの注目

- データジャーナルの創刊
 - Scientific Data (NPG; 2011 -), GigaScience (BioMed Central; 2012 -), Biodiversity Data Journal (Pensoft, 2013-), Geoscience Data journal (Wiley; 2014 -) 等 cf. MethodsX (Elsevier; 2014 -)
- Science: データに関する特集号 (2011.2 special issue “Dealing with Data”)
 - 「データの形成または分析に関連する」コンピュータプログラムの共有: 論文中での「データの利用可能性やキュレーションに関する特別な説明」
- 「進化と生態学の主要雑誌」新たなデータ・アーカイビングの方針 – 公開アーカイブへのデータのデポジットを要求または奨励; The American Naturalist, Evolution, Journal of Evolutionary Biology, Molecular Ecology, and Heredity
- 経済学その他でもデータへのアクセスを要求
 - 方式 フォーマル (Protein Data Bankへの登録へのデポジット, 論文には構造エントリ番号を記載), インフォーマル (資源へのリンク)

35

学術雑誌における研究データへの注目

- PLOS: 論文の一部としてData Availability Statementを要求
 - (データの公開方法 (リポジトリでの公開 [強く推奨], 関連情報ファイルの添付, 要請に応じて提供, サードパーティから提供) についての説明; なお, リポジトリでの公開の場合はDOIやアクセス番号の付加) を要求 (2014.3 -)
- “stub article” (stub; 切り株?; wikipediaでは展開が必要な短い記述のこと)
 - 「分野別リポジトリでの重要かつ新たなデータセットの寄託を公表し, その (データ) 生成に関わる背景情報を提供はするが, データの分析を行わず後続の出版物に委ねる (論文)」 (Lynch, Clifford A. “The shape of the scientific article in the developing cyberinfrastructure,” *CT Watch Quarterly*. 2007.8)

36

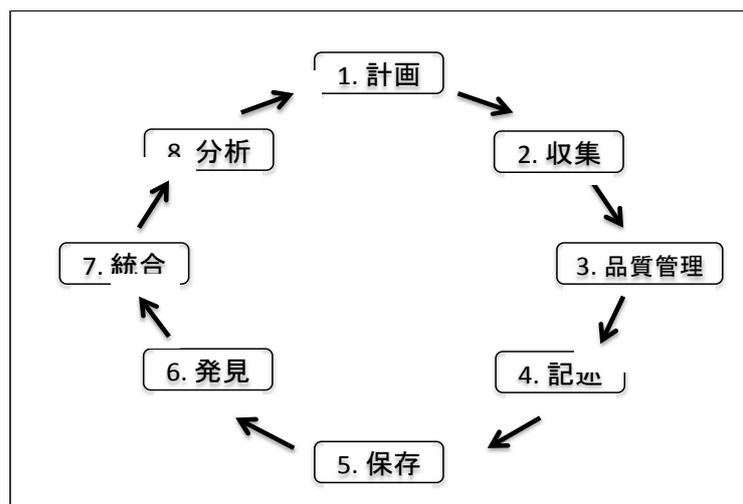
研究データ管理サービス (Research Data Management Services)

- 研究データ管理サービスとはデータのライフサイクル全体を対象として、「図書館がデータ管理に関連して研究者に提供するサービス、情報のサービス、および技術サービスを含む」
 - サービスの例.
 - データ管理計画あるいはメタデータ基準についての教員、スタッフ、あるいは学生との相談
 - 知見やデータセット引用へのレファレンス支援
 - データやデータセットのためのウェブ上の案内や発見支援
 - データリポジトリのための技術支援
 - リポジトリのためのデータセットの準備
 - データセットのリポジトリからの登録解除または除外
 - データセット用のメタデータの生成

Tenopir, C.; Birch, B.; Allard, S. *Academic Libraries and Research Data Services: Current Practices and Plans for the Future. An ACRL White Paper.* Association of College & Research Libraries. 2012, 54p.

<http://www.ala.org/acrl/sites/ala.org/acrl/files/content/publications/whitepapers/Tenopir>

データのライフサイクル



(<http://www.dataone.org/best-practices> をもとに作成)

38

データキュレーション

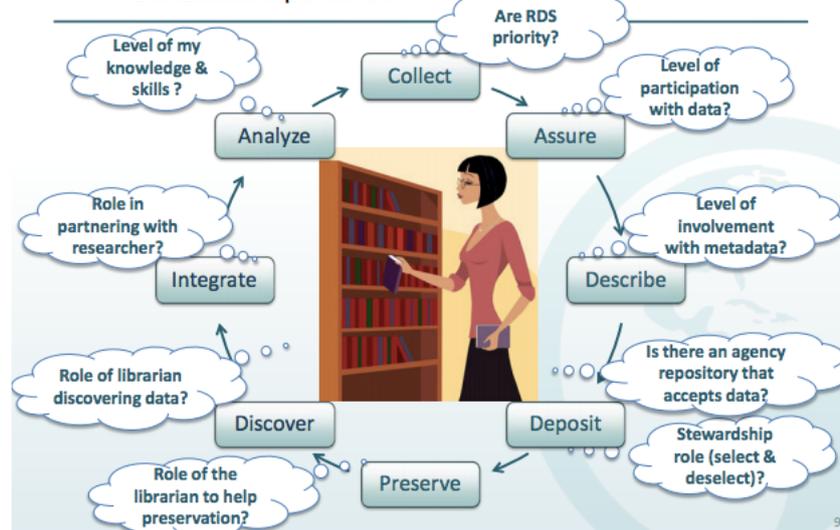
「科学，社会科学，人文学全般の研究教育活動に関係を持つ有益なデータのライフサイクル全般にわたる積極的で継続的なマネジメントである。データキュレーション活動は長期に渡って，データの発見，検索，品質維持，価値の向上再利用のための提供を可能にする。この新たな領域には，説明，アーカイビング，認証，管理，保存，検索，利用が含まれる。」

DCEP: Digital Curation Education Program.

<http://cirss.lis.illinois.edu/CollMeta/dcep.html>

39

Figure 2. The librarian ponders whether she has the background, skills, and education to provide RDS.

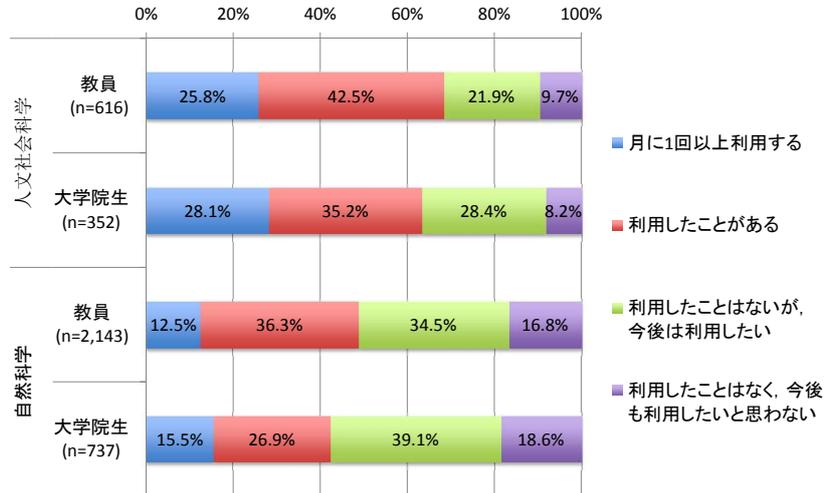


Tenopir, C.; Birch, B.; Allard, S. *Academic Libraries and Research Data Services: Current Practices and Plans for the Future. An ACRL White Paper.* Association of College & Research Libraries. 2012, 54p.

<http://www.ala.org/acrl/sites/ala.org/acrl/files/content/publications/whitepapers/Teno>

40

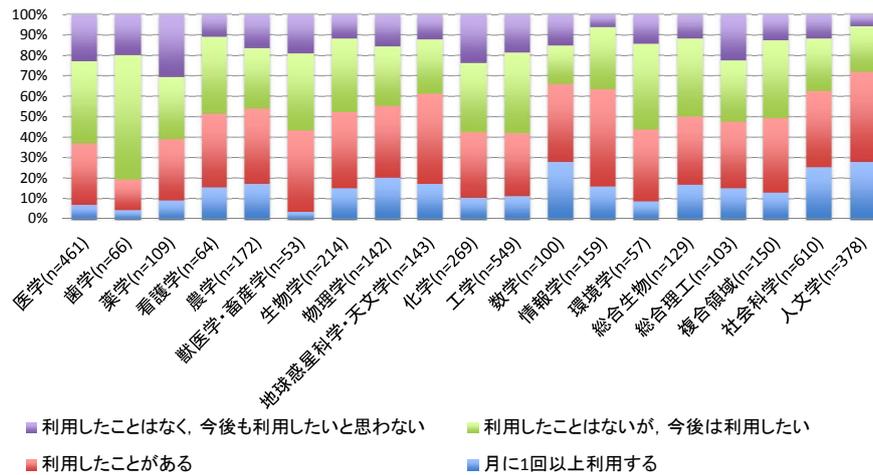
共有リポジトリやウェブ等で公開された研究データの利用 (SCREAL調査2014)



<http://www.screal.jp/>

41

共有リポジトリやウェブ等で公開された研究データの利用 (SCREAL調査2014)



<http://www.screal.jp/>

42

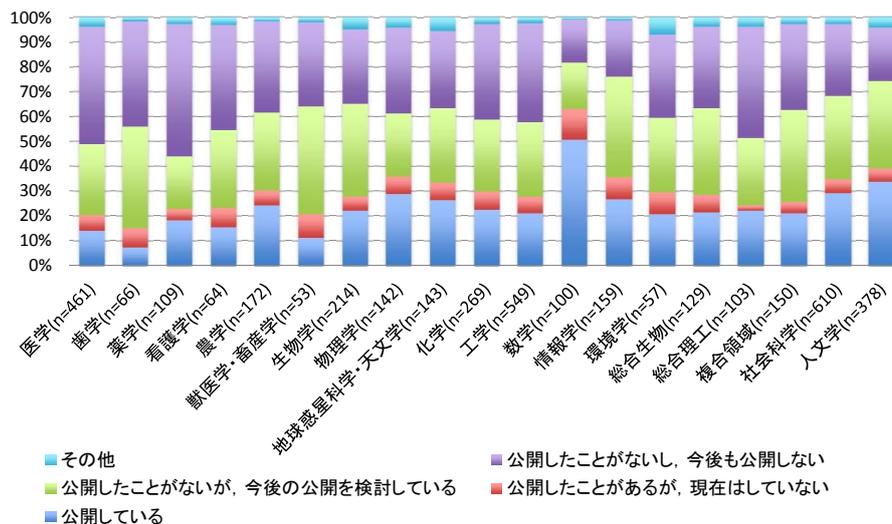
自分の研究データの共有リポジトリやウェブでの公開(SCREAL調査2014)



<http://www.screal.jp/>

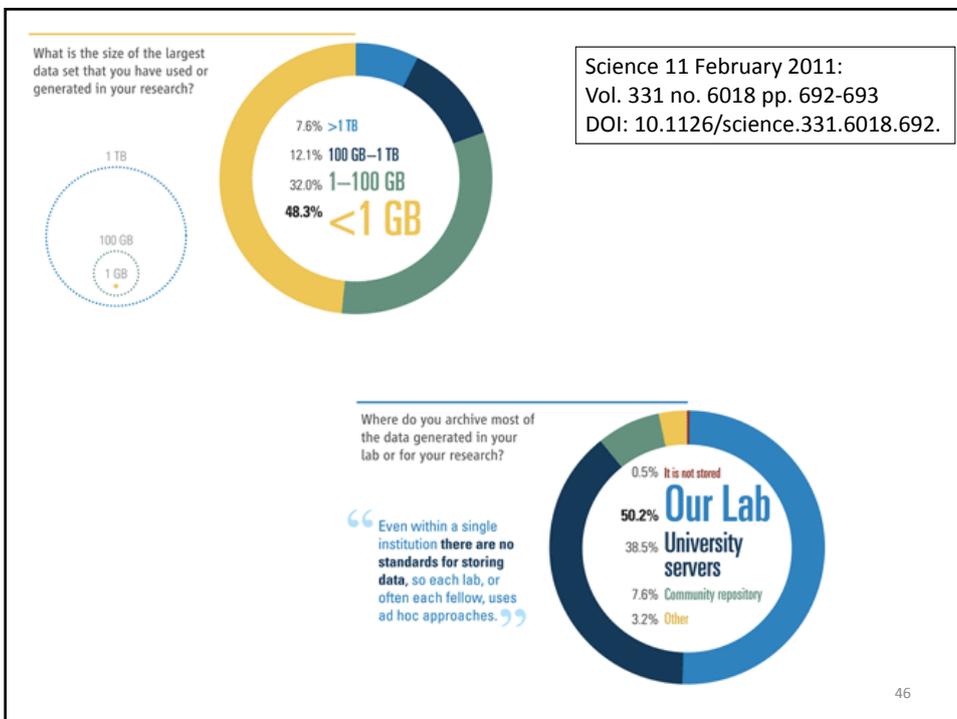
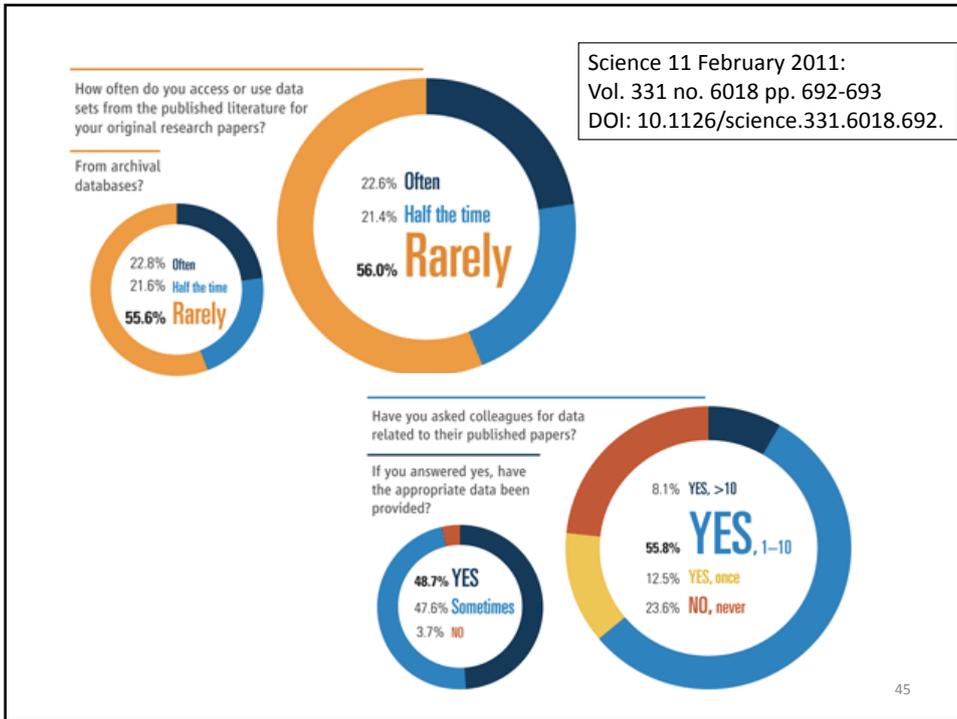
43

自分の研究データの共有リポジトリやウェブでの公開(SCREAL調査2014)



<http://www.screal.jp/>

44



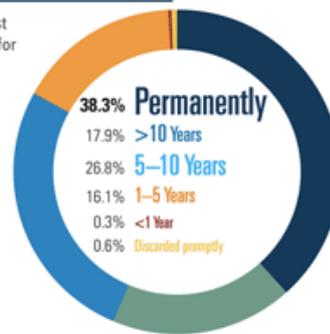
Is there sufficient funding for your lab or research group for data curation?

“There are many tales of early archaeologists burning wood from the ruins to make coffee. If we fail to curate the environmental archives we collect from nature at public expense, we essentially repeat those mistakes.”



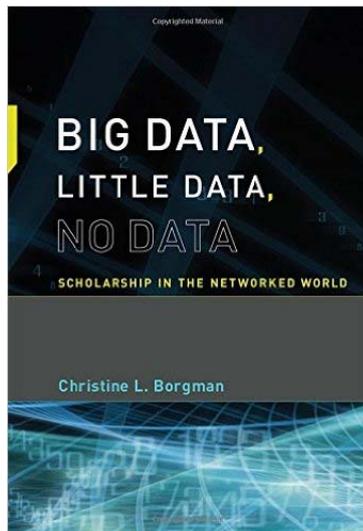
Science 11 February 2011:
Vol. 331 no. 6018 pp. 692-693
DOI: 10.1126/science.331.6018.692.

For how long do you store most data generated in your lab or for your research associated with your publications?



47

データ公開, 共有の課題



Christine L. Borgman.
Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press, 2015.1

48

データ公開, 共有の課題

1) データとは何かはあいまい

- 多様なデータ, データセット
 - 社会科学のサーベイ調査や, 人文学のテキスト・コーパスを含む
 - NSFは研究データの定義を回避
 - データ管理計画において, 説明をFAQの最初に置くことでデータの定義を回避
- 「Q. データ管理計画の対象となるデータの構成要素は何か?」
「A. データを構成する要素は, ピアレビューとプログラム管理の過程を通じて関係コミュニティによって決定されることになる。データ, 出版物, サンプル, 物的コレクション, ソフトウェア, モデルが含まれるが, これらに限定されるわけではない」
- したがって, 何を保存するか, 公開するかもあいまい

49

データとは何か

- 人工物, または観察記録, Michael Bucklandの簡潔な表現を用いるとせいぜい「不確かな証拠」(Buckland, 1991)
 - 見る人の目にのみ存在するのかもしれない; 観察記録, 人工物, あるいは記録がデータを構成することを認めること自体が学術活動である
 - 最も広く引用される定義 (National Academies of Science Report, p.15)
 - 「データは, 対象物, アイデア, 状態, 状況, または他の要素を記述する事実, 数字, 文字列, シンボルである」
 - より最新の実用的定義(国際学会の資料) - [共有の議論に有用]
- 本文書で用いる「データ」という用語は, 広く包括的であることを意図している。デジタルによる文献の表現(テキスト, 音声, 静止画, 動画, モデル, ゲーム, あるいはシミュレーションを含む)に加えて, 分光学, ゲノム配列, 電子顕微鏡法のデータといったさまざまなタイプの実験室データ, リモートセンシング, 地理空間, 社会経済のデータといった観測データ, そして人間または機械によって生成または編集された他の形式のデータといった, 一般に有用とするためにコンピュータの機器やソフトウェアの活用を必要とするデータやデータベースの形式もあてはまる。(Uhlir & Cohen, 2011)

50

Open Dataの定義

- 1) だれでもが自由に利用, 再利用, 再配布できるなら, そのデータはオープンである。最大でも専門分野に限られ, 帰属表示および/または継承の要件がある。
Open Data Commons. (2013). <http://opendatacommons.org/>
- 2) 知的なオープン性の基準に見合うデータ。データは, 入手可能, 利用可能, 評価可能, 理解可能でなければならない。
Boulton, G., et al. (2012). "Science as an open enterprise." The Royal Society. <http://royalsociety.org/policy/projects/science-public-enterprise/report/>
- 3) オープンであること, 柔軟性, 透明性, 法律遵守, 知的所有権保護, 公的責任, 専門家意識, 相互運用性, 品質, セキュリティ, 効率, 説明責任, 持続可能性。
Organization for Economic Cooperation and Development. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. <http://www.oecd.org/dataoecd/9/61/38500813.pdf>

51

データ公開, 共有の課題

2) 学術論文との違い

- ✓ 権利関係
 - 著作権
 - その他(プライバシー, 特許, 法的責任, ...)
- ✓ 公開のインセンティブ

※雑誌論文は読者への供給のためにパッケージ化されているが, データは学術活動のプロセスから解放するのが難しい。データの公開には,

多くの場合, 研究実施と公表物の執筆に対するかなりの投資が求められる。データは, 研究生活の過程全般を通して蓄積される貴重な資産であり, 仮に公開する場合でも注意深く行われるべきである

52

学術論文のオープンアクセス

- 「オープンアクセス文献は、デジタルでオンラインの、無料の、そしてたいの著作権やライセンスの制限から自由なものである」
- 含意されていること
 - a. 出版者といった別の団体へ所有権を移転しない限りおよび移転するまでは、著者が自らの著作の著作権者であること
 - b. 研究者は基本的に研究論文の執筆に対し報酬を受けることはない。研究者は自らの論文を収益を失うことなく広く頒布することができる

※Green OAもGold OAも、上記の原則に依拠

※研究データの場合は、両方ともあてはまらない

53

By "open access" to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.

Budapest Open Access Initiative (2002)
<http://www.budapestopenaccessinitiative.org/read>

54

知識コモンズの経済学

		減少性(枯渇性)/競争性	
		低	高
排除可能性	困難	公共財 一般的知識 パブリックドメインのデータ	共有資源 図書館 データ・アーカイブ
	容易	有料財またはクラブ財 購読方式の雑誌 購読方式のデータ	私的財 印刷体の図書 生データまたは競争性のデータ

Christine Borgman, "Data Scholarship in the Humanities," <http://works.bepress.com/borgman/341/> ;
 C. Hess & E. Ostrom (Eds.), Understanding knowledge as a commons: From theory to practice. MIT Press, 2007, p. 9.

55

データ公開, 共有の課題

3) データの利用は文脈依存

- ✓ メタデータと説明が不可欠
 - しかし, 作成には時間を要する
 - 誰に向けて説明? (説明を理解できるレベル)
- ✓ データの利用には特別な機器やソフトウェアなどが
必要になる場合も
- ✓ 誤用, 誤解釈, 法的責任 --- 「データは資産でも債務
でもある」
- 「“実在しない場所からの視点”などない。知識は常に,
場所, 時間, 状況, 慣習, 理解に根付いている。単一の
知識ではなく, 複数の知識が存在する」

Van House, N.A. (2004). "Science and technology studies and information studies." in Cronin, B. (Ed.). *Annual Review of Information Science and Technology*. Medford, NJ, Information Today, 38, 3-86.

56

データ公開, 共有の課題

4) データ出版という単純化の問題

- 学術出版の三つの機能:
 - i. 正当化 (legitimization); ピアレビューを通じて
 - ii. 頒布 (dissemination)
 - iii. アクセス, 保存, キュレーション
- データ自体が査読を受けているわけではない
- データ出版 (data publication) というメタファーは, 国際統計, 国勢調査等の狭い意味でしか成立しない
- 雑誌論文とデータセットの間の1対1のマッピング
 - 発見を容易にはするが, 関係はしばしば多対多

57

データ公開, 共有の課題

5) 保存とアクセスのバランス

- 保存に最も有効な形式は, アクセスには向かない (i.e. 高解像度 vs 低解像度)
- デジタル保存: ダークアーカイブ等, 研究コミュニティ, 大学コンソーシアム, あるいは国全体でのインフラ投資が不可欠

6) 保存のコストは学術論文の場合よりも高づく

58

データ公開, 共有の課題

「ビッグデータ, リトルデータ, そしてノーデータ」の概念さえも, きちんと理解されないままである。より良いデータ管理, 共有, 認定, 帰属を促進する取り組みはかなり志向されているが, 利害関係者はその出発点, 最終目標, そしてそこに至る道筋に賛成していない。データと呼ばれるものは何かについての合意が欠けており, すべての領域にまたがるデータ・スカラーシップの多様性に対応する共有, 公開, デPOSIT, 認定, 帰属表示, 引用, アクセス維持の政策を確立するのは困難なままである。データへのアクセス維持は困難で費用のかかる作業であり, コストと便益は利害関係者間に偏って分布する。何を保持するか, そしてその理由に関する疑問は, 誰が, どのように, なぜ, 誰に対して, どれくらいの期間, データを保持していくかという疑問と分離できない。個々の研究者, 学生, 図書館員, 文書館員, 研究担当者, あるいは雑誌編集者は, この巨象の難問にせいぜい蟻の目を持つに過ぎないのかもしれない。」

Christine L. Borgman. *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press, 2015.1

59

図書館および機関リポジトリとの関係

- 1) なぜ研究データを扱うのか(例. University Record)
- 2) 「データを確実にアーカイブすることがオープンアクセスへの第一歩」(例. DataVerse, SciDrive)
- 3) 知識インフラの構築
 - 知識インフラの特性
 - 障害が起きるまで, 存在に気づきにくい
 - コストや努力が見えにくい
 - 「集合的な」取り組みの必要性
 - 識別子管理(DOI, ARK, PURL等)

60