

大学における 研究データ管理環境整備 の課題

京都大学 青木学聡

<https://orcid.org/0000-0002-5926-4903>

自己紹介

- 2016年頃から
 - 京都大学
 - 情報環境機構 IT企画室
 - 学術情報メディアセンター
 - アカデミックデータ・イノベーションユニット
 - 大学ICT推進協議会(AXIES)
 - 研究データマネジメント部会(主査)
- 「ICTによる大学での研究環境整備」が中心業務
 - 学内クラウドコンピューティング
 - 研究者情報の管理・流通
 - 研究データ管理

ICTによる研究支援、研究者支援の難しさ

- ICTとの向き合い方が様々
 - 「構成員全員が使う」という意味での「全学的」サービス設計は困難
(「〇〇なんてサービスいらない」から「〇〇が使えないなんて遅れている」まで)
- 大学への帰属意識が低い
 - 大学ロックインサービスを忌避(ex. @kyoto-u.ac.jp を 個人のGmail に転送)
 - 大学のサービスそのものが信頼されていない
 - 「全学サービス提供 = ガバナンス強化」への抵抗意識
- 「学外関係者」を取り込む必要性
 - 情報の利用が学内に閉じない
 - 他機関のIT資産を利用する場合も同様
- これらはそのまま「大学における研究データ管理基盤構築」の課題になる

研究スタイルによる研究データの とらえ方の違い

- 学説、手法の正当性を主張するための手段
 - 「新規性を主張する」データを持つことに意義がある
 - データ取得の正しさそのものが議論になる

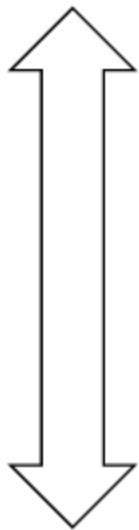
知財管理、研究公正の観点から研究データマネジメントを進めることが必要

- データを取得・収集すること、そのものが目的
 - 「他では取れない1点もの」のデータを持つことに意義がある
 - 公開、共有のためのノウハウの蓄積

オープンサイエンスの先進的事例として研究データマネジメントが実践されていることが多い

「研究データ管理」の多義性

義務的
形式的
：



自発的
多様性
：

- 「研究公正とコンプライアンス」のため
 - ポリシー先行で、具体的な組織的対応が追いついていない
 - 機関にとっては「リスク管理」の課題
- 「学術の発展と社会貢献」のため
(オープンサイエンス的取り組み)
 - 主にコミュニティレベルで多大な労力、時間をかけて醸成
 - 持続的な発展の問題
 - 機関にとっては「研究コンテンツ戦略」の課題

大学でのオープンサイエンスと研究データ管理

- 組織的な「研究データ管理」
2015年頃「研究公正の強化」が直近の契機
 - リスク管理としての「研究データ保存」としか認識されていない
- 組織的な「オープンサイエンス対応」
 - 「機関レポジトリ」、「オープンアクセス論文」の概念は浸透
 - オープンサイエンスの意義、義務については理解が不十分

組織(ICT支援部門)はどのように研究データ管理、オープンサイエンスをサポートすべきか

大学ICT推進協議会(AXIES)において

「学術機関における研究データ管理に関する提言」を作成
(近日公開)

- I. 研究データ管理における学術機関の役割
- II. 学術機関における研究データ管理の導入目的の実際
- III. 学術機関が管理・提供する研究データと利活用場面
- IV. 学術機関における研究データ管理を成り立たせる条件
- V. 学術機関における研究データ管理のための仕組み
- VI. 研究データ管理のためのデジタルプラットフォームの機能要件
- VII. 研究データ管理のための人材育成

学術機関が整備 するデジタルプラットフォームの具体例

- Edinburg University Research Data Service
<https://www.ed.ac.uk/information-services/research-support/research-data-service>
- 研究ステージ毎に利用できるICTを簡潔に紹介
1ページ1機能, 16ページ
<https://www.ed.ac.uk/files/atoms/files/rds-brochure-aug2018-online-version.pdf>
 - Before
 - During
 - After
 - (Training & Support)



A guide to the Research Data Service (1)

- Before
 - DMPonline (<https://dmponline.dcc.ac.uk/>)による研究データ管理プラン作成支援
 - Finding & analysing data:
 - データ検索に関するコンサルティング
 - オンラインデータ解析サーバー(<http://stats.datalib.edina.ac.uk/sda/>) の紹介
- During
 - ネットワークストレージ
 - DataStore: SMB, NFS等で利用するネットワークドライブ.
 - DataSync: Dropbox のようなクラウド型ファイル共有サービス.
 - MS Office 365 Sharepoint の提供
 - バージョン管理: subversion & gitlab
 - Data Safe Haven:
sensitiveデータを扱うためのストレージ+仮想デスクトップ

A guide to the Research Data Service (2)

- After
 - DataShare (<http://datashare.is.ed.ac.uk/>): 機関データリポジトリ
 - PURE(Elsevire社の研究者総覧システム):
<http://www.research.ed.ac.uk/portal/>
研究者自身による研究成果公開の場を兼ねる
 - 外部データレポジトリの参照メタデータの登録
 - PURE自身での簡易なデータセット、文書公開
 - DataVault: 主にPUREの登録データを(メタデータごと)長期保存する

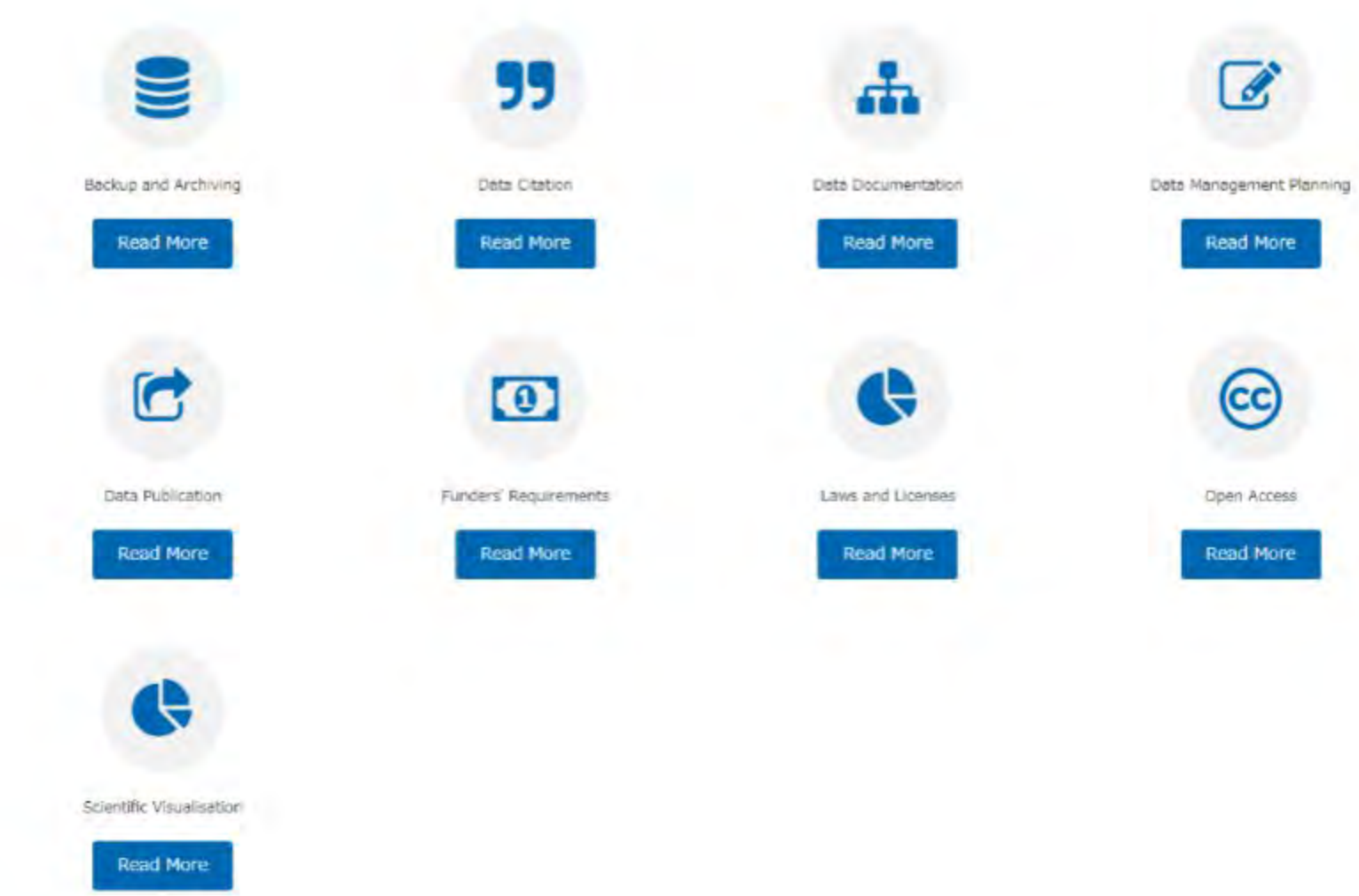
A guide to the Research Data Service (3)

- Training
 - MANTRA(<http://datalib.edina.ac.uk/mantra/>):
 - 研究者(学生含む)向けのデータマネジメントe-learningテキスト
 - MOOC
 - <https://www.coursera.org/learn/data-management>
にてオンライン学習コースを開講

特に, 若手研究者,
大学院生に対する情報リテラシ教育として普及に注力

その他の事例: Göttingen eResearch Alliance Research Data Management

<http://www.eresearch.uni-goettingen.de/content/research-data-management-0>



トップダウン・ボトムアップ間調整の 必要性

- ポリシーの制定は組織の在り方を決定
するために必要
- 広範なケースに対応するため抽象的な表現
にとどまらざるを得ない
- 成文化されると「面倒」「締め付け」
ととられる

ポリシーから研究現場へのアプローチ

「あるべき」論
prescriptive

- ポリシー・基本方針
- 手順書
- ガイドライン
- チェックリスト
- 教育プログラム、教材

「どう改善する」
descriptive

- 習熟度モデル
- ルーブリック

研究データ管理のための ループリック・成熟度モデル

- 研究者個人に対するもの
 - California Digital Library が提案(2018)
- 組織に対するもの
 - Australian National Data Service が提案(2018)
 - 他いくつか先行事例あり(後述)

California Digital Library による Research Data Management Rubric

- John A Borghi, Stephen Abrams, Daniella Lowenberg, Stephanie Simms, John Chodacki, "Support Your Data: A Research Data Management Guide for Researchers",
<https://doi.org/10.3897/rio.4.e26439>
- <https://uc3.cdlib.org/2018/01/11/support-your-data/>
- 主に研究者個人が実践する研究データ管理の成熟度評価
- 6つの評価項目 - 研究データのライフサイクルにそった
「Planning your project」 「Organizing your data」 「Saving and backing up your data」 「Getting your data ready for analysis」 「Analyzing your data and handling the outputs」 「Sharing and publishing your data」
- 4つの評価レベル - 「Ad hoc」 「One-Time」 「Active and Informative」 「Optimized for re-use」

RDM Rubric

<https://uc3.cdlib.org/2018/01/11/support-your-data/>

	Ad Hoc	One-Time	Active and Informative	Optimized for Re-Use
Planning your project	When it comes to my data, I have a "way of doing things" but no standard or documented plans.	I create some formal plans about how I will manage my data at the start of a project, but I generally don't refer back to them.	I develop detailed plans about how I will manage my data that I actively revisit and revise over the course of a project.	I have created plans for managing my data that are designed to streamline its future use by myself or others.
Organizing your data	I don't follow a consistent approach for keeping my data organized, so it often takes time to find things.	I have an approach for organizing my data, but I only put it into action after my project is complete.	I have an approach for organizing my data that I implement prospectively, but it not necessarily standardized.	I organize my data so that others can navigate, understand, and use it without me being present.
Saving and backing up your data	I decide what data is important while I am working on it and typically save it in a single location.	I know what data needs to be saved and I back it up after I'm done working on it to reduce the risk of loss.	I have a system for regularly saving important data while I am working on it. I have multiple backups.	I save my data in a manner and location designed maximize opportunities for re-use by myself and others.
Getting your data ready for analysis	I don't have a standardized or well documented process for preparing my data for analysis.	I have thought about how I will need to prepare my data, but I handle each case in a different manner.	My process for preparing data is standardized and well documented.	I prepare my data in such a way as to facilitate use by both myself and others in the future.
Analyzing your data and handling the outputs	I often have to redo my analyses or examine their products to determine what procedures or parameters were applied.	After I finish my analysis, I document the specific parameters, procedures, and protocols applied.	I regularly document the specifics of both my analysis workflow and decision making process while I am analyzing my data.	I have ensured that the specifics of my analysis workflow and decision making process can be understood and put into action by others.
Sharing and publishing your data	I share the results of my research, but generally I do not share the underlying data.	I share my data only when I'm required to do so or in response to direct requests from other researchers.	I regularly share the data that underlies my results and conclusions in a form that enables use by others.	Because of my excellent data management practices, I am able to efficiently share my data whenever I need to with whomever I need to.

RDMルーブリック(和訳)

https://github.com/kyoto-u/rdm_rubric

	その都度	1度のみ	常に見直す	再利用のために最適化
データマネージメント計画	データを入手した段階で、自分なりに考える。標準化や文書化はしていない。	最初に計画を作成するが、通常は、途中で見直すことはしない。	詳細な計画を作成し、プロジェクト期間を通じ参照と見直しを行う。	自分あるいは他人が将来スムーズにデータを利用できるように計画している。
データの体系的整理	一貫した手法に従っていない。そのため、しばしば探し出すのに時間を要する。	手法はあるが、それを適用するのは、プロジェクト終了時のみ。	先を見越した手法を導入しているが、必ずしも標準化されてはいない。	他人が自分でデータを検索、内容を理解し、利用できる。
データの保存とバックアップ	作業中にどのデータが重要であるかを決め、通常はこれを特定の1か所に保存する。	どのデータを保存すべきかを理解している。損失のリスクを緩和するため、作業後にバックアップをとる。	作業中でも定期的にデータを保存するシステムを用いている。複数のバックアップを保持している。	再利用する機会を最大限生かせるよう考慮した方法と場所にデータを保存している。
データを解析可能な状態にする	標準化もしくは十分に文書化された手続きはない。	解析に適した形式を検討し整備するが、事例ごとに異なる方法をとっている。	データ処理方法は標準化され、かつ十分に文書化されている。	将来も自分自身及び第3者が利用しやすい方法で、データ処理を行っている。
データ解析と成果の取扱い	解析や実験がどのような手順やパラメータにより実施されたか確認するために、これらをやり直すことが多々ある。	解析作業終了後、パラメータの詳細や解析手順、プロトコルを文書化している。	データ解析の際、決められた方法で解析のワークフローや判断のプロセスの詳細をいつも記録している。	自身が行った解析のワークフローや判断プロセスの詳細が他人によっても実行できることが確認できている。
データの共有と公開	研究成果を共有しているが、その根拠となるデータについては共有できていない。	リクエストがあった場合にのみ研究データの共有を行っている。	自身の研究成果や根拠となるデータは、第三者が利用できる形式で共有できる。	優れたデータマネージメントを実践し、自身のデータをいつでもだれとでも効率的に共有

ワークショップによるルーブリック作成、相互評価

- 2018年10月の京都大学
アカデミック・
データワークショップにて実施
- 参加者10名
- Google Docs
にあらかじめ質問事項を記載した文
書を配置、各人が同時に回答を書
きこむ
- 他人の回答から、「どのように回答
すればよいか、記述精度はどの程度
か」がリアルタイムで認識
できるので、「手が止まらない」
- 回答時間は約30分、そのあと結果を
見直しディスカッション
- ディスカッションの概要は、



Australian National Data Service (ANDS) による Capability maturity model table

<https://www.ands.org.au/guides/creating-a-data-management-framework>

組織での研究データ管理評価のためのルーブリック

5つの評価項目 – 「Policies and procedures」 「IT Infrastructure」 「Support Services」 「Managing Metadata」 「Managing Research Data」

5つの評価レベル – 「Process is disorganised & ad hoc」 「Process is under development」 「Process is standardised, communicated」 「Process is managed, Measured」 「Focus is on continuous improvement」

ANDS Capability maturity model table

	Level 1 Initial	Level 2 Development	Level 3 Defined	Level 4 Managed	Level 5 Optimised
	Process is disorganised & ad hoc	Process is under development	Process is standardised, communicated	Process is managed, measured	Focus is on continuous improvement
Institutional policies & procedures	Policies & procedures may be undeveloped, not up to date, and/or inconsistent.	Policies & procedures are developed & harmonised.	Policies & procedures are promulgated & absorbed into behaviours.	Policies & procedures accepted as part of the culture & subject to audit.	Policies & procedures are subject to review & improvement.
IT infrastructure	IT infrastructure provision is patchy, disorganised & poorly publicised.	Funds are invested in technology & skills. Responsibilities are defined. Processes are established, defined & documented.	Management shows active support. Facilities are well-defined & communicated, standardised & integrated.	Funding adapted to need. Management is actively engaged. Documentation kept up to date.	Concerted efforts to maintain, update & publicise infrastructure. Metrics & feedback used to optimise services.
Support services	Training is ad hoc, curation & preservation services are disorganised, data management planning is unsupported & other services inconsistent & poorly publicised	Investment in skills. Services identified & staffed. Responsibilities are defined. Documentation & training developed.	Active participation in training with widespread availability of support services.	Widespread take-up of services. Curation & preservation acknowledged as critical to the institutional mission.	Customer feedback used extensively to update & improve services.

ANDS Capability maturity model table (続き)

	Level 1 Initial	Level 2 Development	Level 3 Defined	Level 4 Managed	Level 5 Optimised
	Process is disorganised & ad hoc	Process is under development	Process is standardised, communicated	Process is managed, measured	Focus is on continuous improvement
Managing metadata	Metadata management is ad-hoc, chaotic & understood by only a few and without established standards.	Responsibilities are defined & skills developed. Processes are established & documented. Metadata applied to key datasets & shared externally.	Processes are standardised & integrated. Metadata created for new datasets & shared externally, to ensure data is findable and accessible.	Metadata quality metrics are collected. All datasets described in machine-readable format & metadata shared. Metadata aligns with the FAIR data principles.	Continuous improvement applied to processes & capabilities.
Managing Research Data	Data is stored in ad-hoc facilities. Only data custodians know where the data is stored and accessible only to the researcher or small group of researchers, Standard formats are not applied and the potential for reusability is limited	Institutional data storage facilities are being developed. Data standards established	Well defined data storage facilities are established. Data managed using standard open formats widely used. Some connectivity between systems permitting limited reuse and sharing.	Data routinely stored in established repositories in machine-readable formats using open standards in established facilities according to FAIR principles.	Continual improvements to maintain update & publicise infrastructure. Metrics & feedback used to optimise services.

組織のためのRDMルーブリック/成熟度モデルに関する議論

- Jian Qin, Kevin Crowston, Arden Kirkland, "Pursuing Best Performance in Research Data Management by Using the Capability Maturity Model and Rubrics", Journal of eScience Librarianship, Vol. 6, No. 2 (2017) e11113 <https://doi.org/10.7191/jeslib.2017.1113>
- RDA "fair data maturity model wg"
<https://www.rd-alliance.org/groups/fair-data-maturity-model-wg>
- Angus Whyte, Magdalena Getler (DCC), "Building the capabilities for a professional RDM support service", Danish National Forum for Research Data Management,
<https://slideplayer.com/slide/14755415/>
- 池内有為, 「研究データ管理 (RDM) の目的地と現在地」, 情報の科学と技術, 69 巻 3 号 (2019), 125-127
https://doi.org/10.18919/jkg.69.3_125

大学でのオープンサイエンス・研究 データ管理支援環境の整備方針

- 執行部による「強い動機付け」
 - 継続性の担保
 - 明確な上位ポリシーと豊富なベストプラクティス
- 標準的なフレームワークの提示と活用
 - フレームワークと個別手法とのギャップの認識と改善
 - 2者間の実施手順でコンフリクトが生じた場合の調停方法
 - 新規プロジェクト、若いPI、未知の研究領域をターゲットに「より良い習慣付け」