米国の大学を中心としたオープンデータの現状 : 訪問調査から

畑埜, 晃平
九州大学附属図書館 : 准教授 | 理研AIP

出版情報：九州大学ライブラリーサイエンス専攻シンポジウム「オープンデータとデジタルヒューマニティーズ」, pp.1-, 2017-01-18. Department of Library Science, Graduate School of Integrated Frontier Sciences, Kyushu University
バージョン：
権利関係：

米国の大学を中心としたオープンデータの現状：訪問調査から

九大　附属図書館　/ 理研AIP（兼任）
畑埜晃平

資料（or 修正版）ダウンロード
https://goo.gl/GbPhYq

自己紹介

- 専門分野：機械学習・コンピュータサイエンス
  - 主に学習理論の研究に従事

- 2015　九大附属図書館　研究開発室に異動
  学術情報基盤の研究を開始
- 2016 理研　革新知能統合研究センター（AIP)
  計算論的学習理論チームリーダー　兼任

概要

1. オープンサイエンス，オープンデータの現状
   - 米，ヨーロッパ
   - 日本
2. 訪問調査の概要
   - ハーバード，UIUC，CDL，UCB
3. オープンデータ基盤
   - システム基盤
   - 人的基盤

- まとめ

オープン○○とは

- オープンアクセス(OA, BOAI)：
  誰もが，どのような目的に対しても，自由に使用／修正／共有が可能

- オープンデータ(OD)：
  オープンアクセス可能なデータ

- オープンサイエンス(*):
  - ≒オープンアクセス＋オープンデータ
  - 研究プロセス自体をオープン化する事により科学技術を発展させる試み

  (*)私個人の解釈

具体例：オープンデータ（のごく一部）

- 政府・自治体のオープンデータ
  - 福岡市のPM2.5 の計測データ

- 情報学研究データセット（NII）
  - Yahoo!データセット（知恵袋），楽天データセット（商品情報・レビュー等），etc.

  サービスの高度化・新サービスの創出

- 日本古典籍データセット（国文研，CODH）
  - 源氏物語など７００点の本文，画像データ等

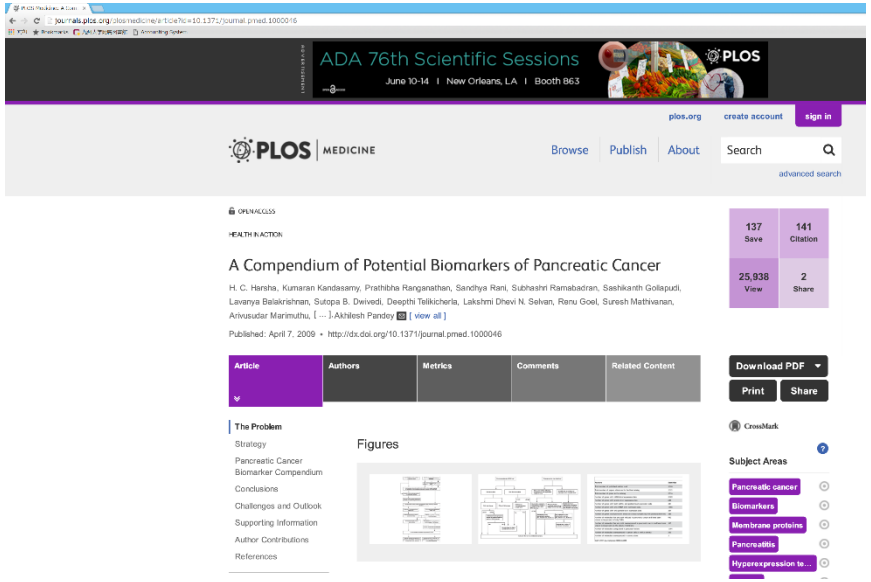
  デジタルヒューマニティーズ

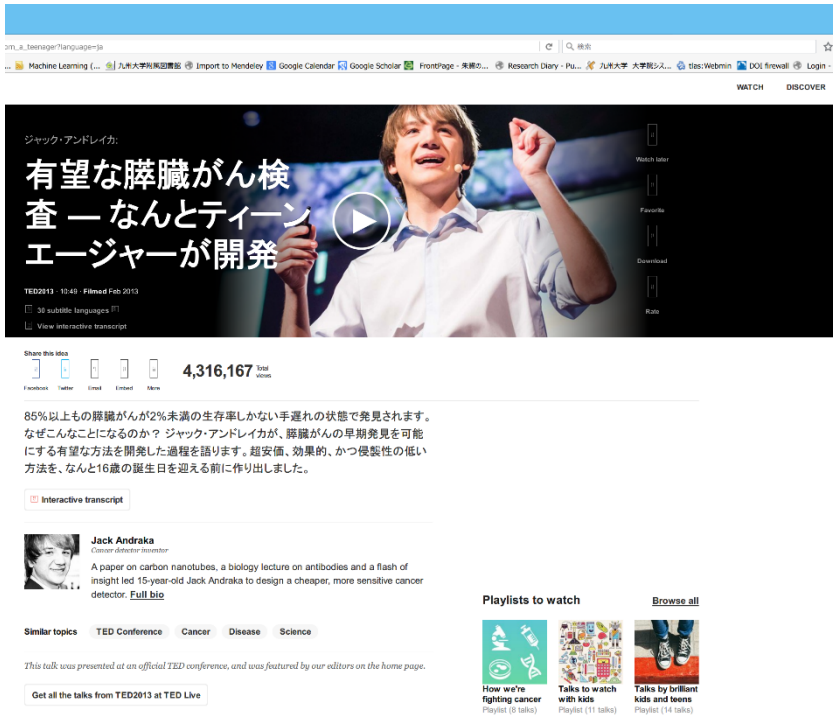
- 雑誌 Natureなど
  - （分野による）投稿時に研究データを一定の条件を満たすデータレポジトリ上での登録義務付け

  再現性・透明性の確保

具体例：オープンサイエンス（Cf. Citizen Science）

- 米国の１６歳の高校生が膵臓がんの早期検査法を発見
- OA雑誌から情報収集





https://www.ted.com/talks/jack_andraka_a_promising_test_for_pancreatic_cancer_from_a_teenager?language=ja

背景：オープン（研究）データに関する要求の高まり

研究者側：

- 研究データを２次利用したい
  ⇒分野特化型データリポジトリ（天文学，バイオ等）
- 結果を検証したい（透明性の確保）
- データを引用したい／されたい
- 関連：研究データの保存義務がある（九大は１０年）
  いずれにせよデータ保存基盤が必要

スポンサー（財団，政府）側:

- 資金（特に税金）によって援助された研究成果は広く公開すべき

オープンサイエンス・オープンデータの現状

- 米国
  - 2011 米国立科学財団（NSF）では研究申請書に「研究データ管理計画」（DMP）の欄を追加
    →米主要大学でデータレポジトリ構築加速
  - その他多くの財団でデータ公開の義務化

- EU
  - 2017- HORIZON2020(770億€規模の研究助成プロジェクト)の援助による研究データの原則オープン化 (open by default)
- 世界
  - 2013 G8サミット：オープンデータ合意
  - →G8諸国でオープンデータを推進する
  - 2013 Research Data Alliance(RDA)発足

- 日本
  - 2015 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」報告書 （日本の基本方針の取りまとめ）

訪問調査の目的

- 背景：日本において
  - 機関レポジトリの普及により，オープンアクセスのための基盤は整いつつある
    （OAの実現はこれから）
  - 研究データのオープン化について大学での取組が課題
- 目的：米国でのオープンデータに対する取組事例から，今後の方向性のヒントを得たい
- 対象：
  - 九大図書館の提携校であるイリノイ大学アーバナシャンペーン校（UIUC）
  - ハーバード大学
  - カリフォルニア大学デジタルライブラリ
- メンバー：宮本館長，石田准教授，畑埜

訪問旅程


サンフランシスコ近郊
（カリフォルニア大学
デジタルライブラリ）
アーバナ
（イリノイ大学
アーバナシャンペーン校）
ボストン
（ハーバード）

システム基盤編

そもそも研究データとは

- 研究データ
  - 研究成果を検証するための事実の（電子的）記録

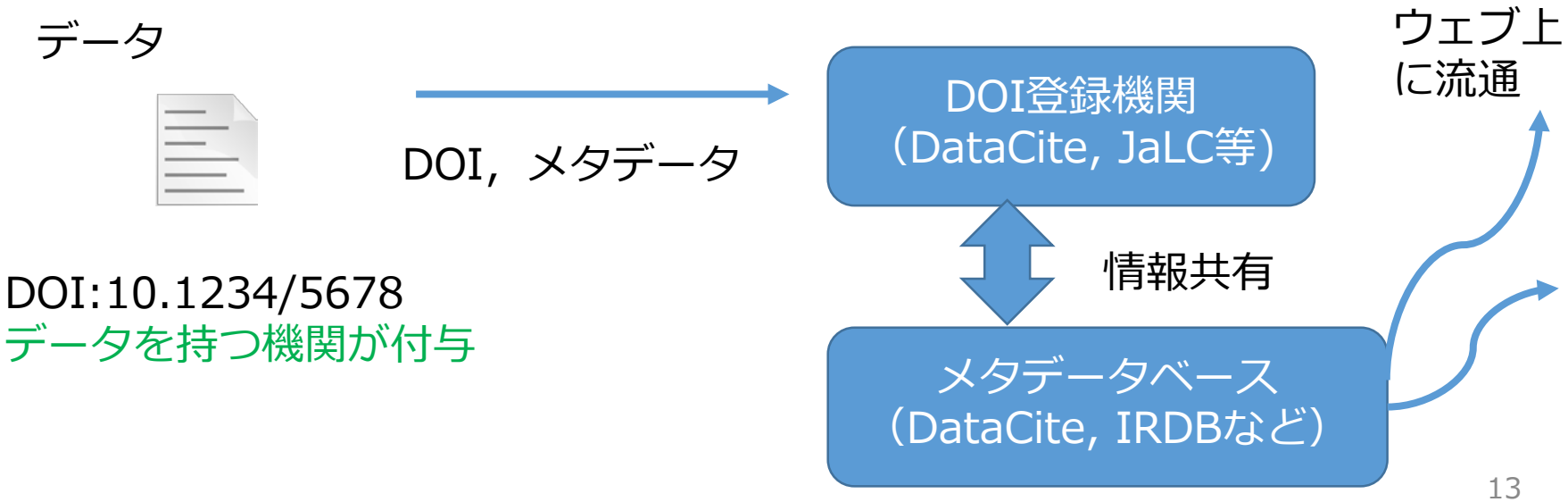
- メタデータ
  - データに付随する属性などの情報

    著者名，出版機関，分野…

前提知識：データ流通の仕組み

- DOI
  - 電子データに対する永続的な識別子　例：10.1234/5678
  - http://dx.doi.org/[DOI] を介してアクセス可能に

- DOI登録機関
  - DataCite, JaLC(Japan Link Center)など



前提知識：データレポジトリの要件

- 管理基盤
  - 財団等の指定するデータ管理計画，出版社の条件に対応
  - セキュリティ（権限，）
- 保存基盤
  - 大規模データも扱える（研究データ>>論文）
  - 分野に応じたメタデータを扱える
  - 外部クラウドとの連携
  - 頑健なバックアップ体制
- 流通基盤
  - 検索インターフェイスを持つ
  - DOI等の識別子を付与し，電子データをアクセス可能に
  - 外部データベースにメタデータを流通

※後述のデータリポジトリは上記の要件をおよそ満たしている

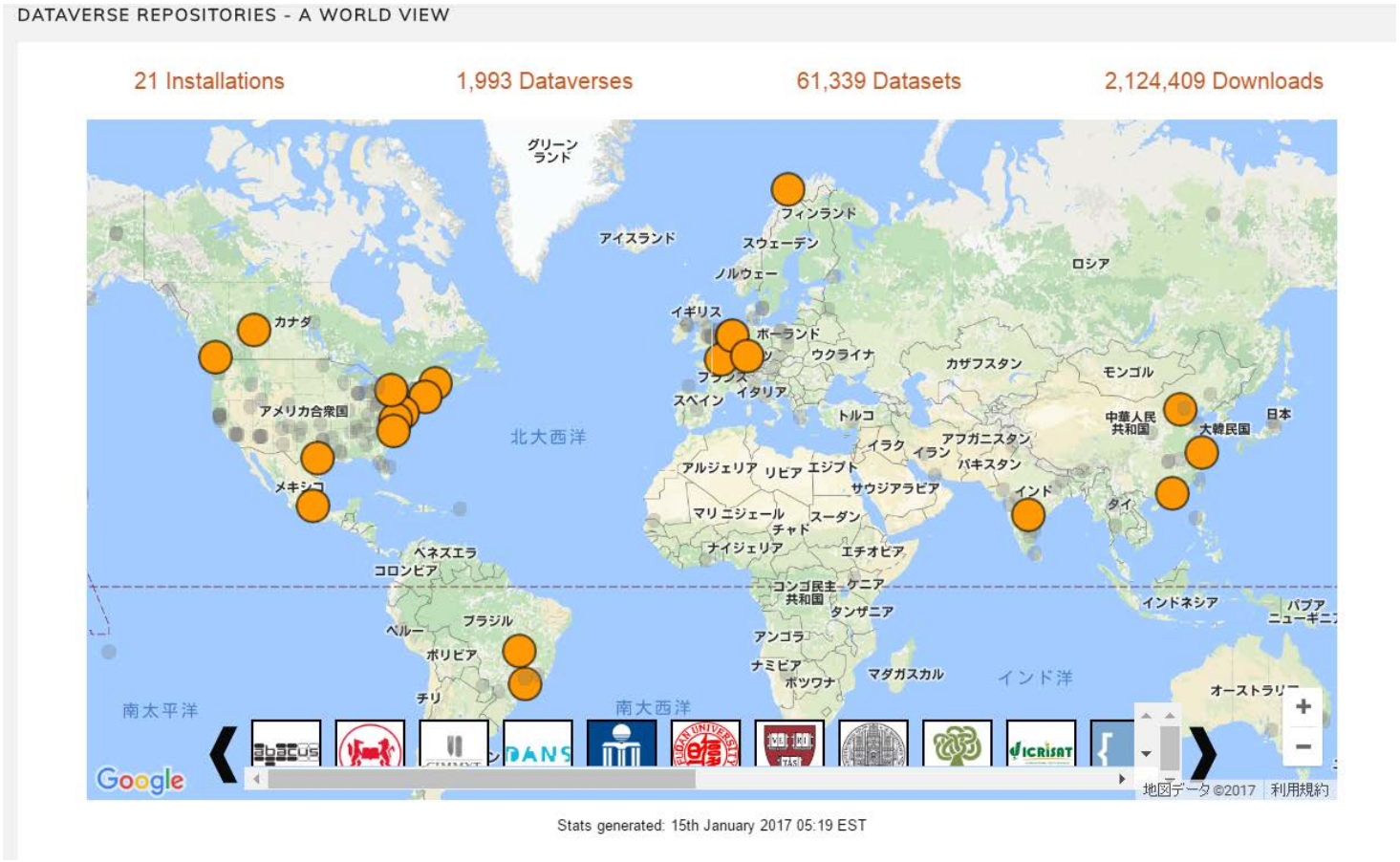
事例１：Dataverse（ハーバード大学）

- IQSS（Institute for Quantitative Social Science)で進められているのデータレポジトリ
  - オープンソース
  - Mercè Crosas and Gary Kingらのチームによって開発中

- 様々なレベルでのアクセス制限が可能（Datatag）
- LOCKSS (LOts o Copies Keep Stuff Safe) を用いてデータを複数の場所にバックアップ可能
- オープンアクセス雑誌のプラットフォームにも
- Dataverse内に複数の Dataverseを構成可能
  （例：大学 Dataverse-学部Dataverse）

Dataverse ネットワーク

- 約170の大学や組織でDataverseを運用(2017.3現在)

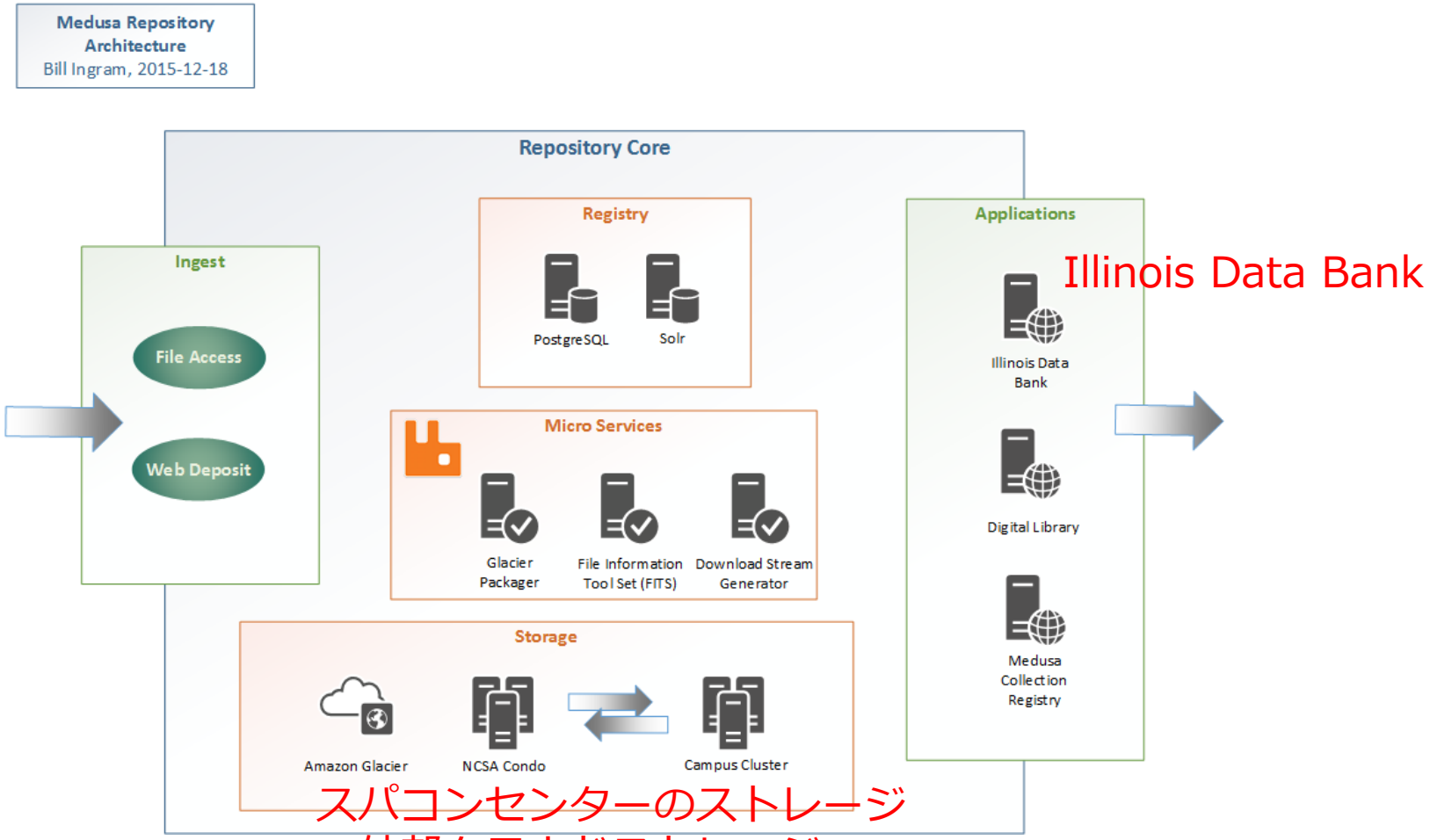


http://dataverse.org/

事例２：Illinois Data Bank（UIUC）

- UIUCで開発されたデータリポジトリ(2016-)
  - 図書館はインターフェイスを担当
  - 保存基盤などは既存の図書館のデータ保存システムやスパコンセンターNCSAのシステムを使用
- 学内の dropbox ライクなシステムも使える
- 学内でのバックアップの他にAmazon のクラウドも利用

例：Illinois Data Bank



*図の出典：Thomas Habing氏(UIUC Library) の資料より

事例３：カリフォルニア大学デジタルライブラリ

- EZID
  - DataCite を通じ，DOIを発行するためのツール（ウェブ，API）
  - データリポジトリにおいて標準的なツール
- DMPTool
  - データ管理計画作成支援ツール
  - 同じく標準的
- Merritt
  - カリフォルニア大学のデータリポジトリ
  - EZID，DMPToolなどのツール群の組合せにより構成
- DASH
  - Meritt を含むデータリポジトリ
  - インターフェイスを強化

事例のまとめ・考察

- 個々の大学で独自開発
- いずれの大学も独自の開発チーム（数人）を持つ
  - 継続的な開発が可能
- ＋動機づけ：様々な財団がオープンデータを課す

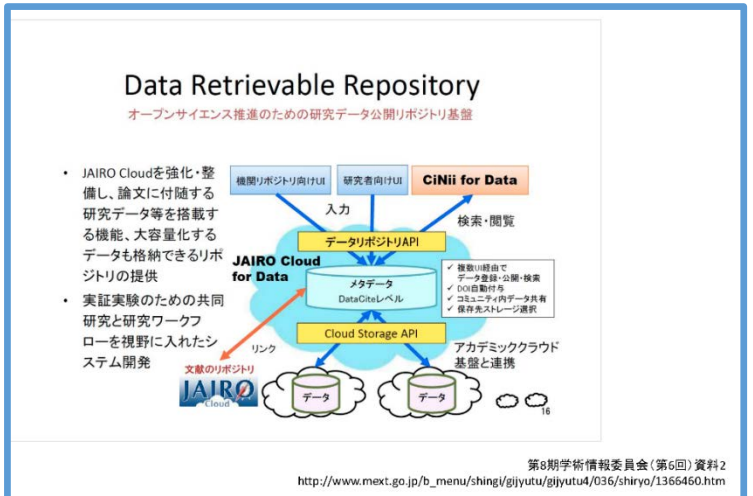
  一方，日本では資金・人材・動機づけが不足

- 研究室・学科単位では Dataverse は導入可能かも
  - 大学単位でも実績あり（データ保存基盤の課題？）
  - DataCite を通じたDOI付与には対応しているが，

    日本のJaLCには未対応
  - 個人単位では figshare, zenoodo? で十分

日本の大学における データリポジトリ確立のために

- 懸案事項：データ保存基盤
  - 学内計算機センター等データインフラの専門家との連携が不可欠
- 自前で開発
  - 持続的な開発チーム・資金・覚悟（？）が必要
- もしくは，共用データレポジトリ（？）への結集



第8期学術情報委員会（第6回）資料2
http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu4/036/shiryo/1366460.htm

人的支援体制編

人的基盤

- 研究データのオープン化のためにはシステム基盤だけではなく，人的サポートも不可欠

- データキュレーション(data curation)
  - GLIS, UIUCにおける定義

    the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education
  - データを流通・管理するための**様々な**サポート
- データキュレータ
  - データキュレーションに関わる人材

DataVerseにおける人的支援 (ハーバードの場合)

- 研究者が自分でシステムにデータセットを登録
  - メターデータも作成
  - ファイルタイプ等の制限もなし
- 登録されたメタデータ等を定期的にチェックし、スタッフが追加・修正
    - 従事しているスタッフは2，3名程度
    - 研究者に聞きながら、メタデータについては追加・修正を行う
- １０名，4-5名のデータキュレータ

Research Data Service (UCIC)

- 専属スタッフは2名(システム構築を除く)
- Heidi J. Imker, ph.D
  - Director, Research Data Service, Associate Professor, University Library
  - バイオケミカルの研究者からの転身
  - 研究者自体にデータそのものを扱っていた経験はあり
- Elizabeth Wickes
  - Data Curation Specialist, Research Data Service
  - 社会学出身
  - イリノイ大学iSchoolに通学
  - プログラミングスキル等は自分で勉強（Python ）

Data Curation Specialistの仕事

- システム設計
  - メターデータに関する勉強
  - メタデータの設計
- ワークショップの開催
  - Introduction to Data Management
  - Documentation and Organization for Data and Processes
  - Making Research Data Public: Why, What, and How
  - Data management/Python drop-in hours
- コンサルテーション
  - データ管理計画の作成支援
- システム立ち上げ後
  - データセット登録のサポート

Data Management Consultation

- Walk through the research project and data management planning process
- Help determine what data can or cannot be made accessible
- Help determine what data can reasonably be preserved
- Identify resources available on campus
- Identify resources available elsewhere

From "Research Data & the Art of Simultaneously Meeting Needs and Picking Battles" (Heidi Imker, PhD, Director, Research Data Service, Associate Professor, University Library, University of Illinois at Urbana-Champaign)

RESEARCH DATA SERVICE

Data Management Workshops

	February Series (4 PM – 5 PM)	April Series (1 PM – 2 PM)
Introduction to Data Management	Feb 10th	Apr 6th
Data and Process Documentation and Organization	Feb 17th	Apr 13th
Data Publication and Sharing: Why, What, and How	Feb 24th	Apr 20th

@ Main Library but willing to travel!

From "Research Data & the Art of Simultaneously Meeting Needs and Picking Battles" (Heidi Imker, PhD, Director, Research Data Service, Associate Professor, University Library, University of Illinois at Urbana-Champaign)

Data Curation Specialistに必要な知識・スキル

- メタデータ等の知識
- プログラミング等のITスキル
  - ただし、本格的なものではなく、基礎的なものでよい
  - データに関するよろず相談
    - 相談例：８００Ｍｂのｃｓｖファイルを Excel で開けない
- 分野におけるデータ使用方法の理解
- 自発的に行動すること
  - Data Curation Specialistとして何ができるかを探っている状態
  - データのオープン化、オープン化されたデータの利活用だけでなく、研究全体のプロセスを応援したい

UC Berkeley図書館の事例

- CDLのシステムと利用者(研究者)を結ぶサポートをするスタッフ
  - 専属は各キャンパス(10キャンパス)に1名程度
  - その他、パートタイムや兼任が数名いるところもある
- Jamie V. Wittenberg
  - Research Data Management, Service Design Analyst
  - Research IT / Library

The iSchool at Illinois

- Specialization in Data Curation
  - Data curation is the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation.
- コース概要
  - MS/LIS院生すべての必修科目(2科目)
  - 専門のための必修科目(3科目)
  - 選択科目(2科目)

Specialization in Data Curation; Course List

REQUIRED COURSES

Required of All MS/LIS Students

- LIS501 Information Organization and Access
- LIS502 (2 hours only) Libraries, Information and Society

Required for the Specialization

- LIS531 Foundations of Data Curation
- LIS562 Metadata in Theory and Practice
- LIS586 Digital Preservation

Specialization in Data Curation; Course List

Recommended Electives

Students must take two (2) courses from this list; taking four (4) is advised.

- LIS452 Foundations of Information Processing in LIS
- LIS453 Systems Analysis and Management
- LIS490DB Introduction to Databases
- LIS560 Digital Libraries
- LIS561 Information Modeling
- LIS590OD Ontology Development
- LIS590RO Representing and Organizing Information Resources

人的基盤まとめ

- データに関するよろず相談
- アウトリーチ活動
- “まずデータを増やすことが重要”
  - データレポジトリの充実＞メタデータの“正確さ”
- 大学におけるデータレポジトリの意義
  - interdisciplinary （分野横断性）
  - 分野毎のデータレポジトリと競合するものではない
  - 分野毎のデータリポジトリですくいきれないデータを保護

→図書館

まとめ

- オープンデータのためのシステム基盤・人的基盤の調査（ハーバード，UIUC，UC）

- システム基盤
  - データリポジトリとしての要件を整理

- 人的基盤
  - データキュレータの実際

謝辞

石田栄美

Sue Kriegsman Mercè Crosas Sonia Barbosa Franziska Frey

Peter Suber Kris Markman Amy Deschenes Steven Brzozowski

Sue Gilroy Ramona Islam Mikel Burt other staffs in Harvard

Allen Renear Linda Smith Stephen Downie Bill Mischo Tim Cole, Mary Schlembach Christie Wiley Kelli Trei Alex Cabada Ali Krogman Thomas Habing John Wilkin Kyle Rimkus Aaron McCollough

Heidi Imker Rebecca Bryant Elizabeth Wickes other staffs in RDS

Perry Willett Stephen Abrams Lisa Schiff Jamie Wittenberg other staffs in UCB