

データキュレーションへの 期待と課題：自然科学から 人文科学まで

北本 朝展 (KITAMOTO Asanobu)

国立情報学研究所 (NII)

情報・システム研究機構

人文学オープンデータ共同利用センター (CODH)

<http://agora.ex.nii.ac.jp/~kitamoto/>

自己紹介



- 研究分野は**情報学**。画像処理や画像データベースの研究が出发点。
- データ駆動型サイエンスに発展し、**気象情報**、**地球環境情報**、**人文科学情報**などの分野で研究。
- 最近はオープンサイエンスにも関わる。

オープンサイエンスとは？

- 「オープン」という言葉を梃子にして、サイエンス（研究）の方向を変える。
- 「よりオープンに」という方向性を共有する活動を、一語で束ねると見える世界。
- 個々の活動ごとに「オープンサイエンス」の意味は異なり、単一の定義は困難。
- 大同団結？同床異夢？個々の活動を超える新しい目標を示せるかが問われる。

オープンサイエンスへの収束

透明性

オープンアクセス

共有

オープンピアレビュー

研究の再現性・
透明性・研究
データ保存

市民科学・クラウド
ファンディング

参加

オープ
ンサイ
エンス

オープンデータ

研究データ
データ出版
データリポジトリ

コラボレーション・オー
プンイノベーション

超学際研究

協働

メタ研究＝研究（システム）に関する研究

「オープン」の3つの側面

1. 他者が使える（再利用）

- ・ オープンデータやオープンアクセスなど。外部の人が研究結果を自分の目的に再利用できる。

2. 他者が検証できる（透明性）

- ・ オープンガバメントや研究再現性など。外部の人がエビデンスを検証し、正当性を判断できる。

3. 他者を受け入れる（参加）

- ・ オープンイノベーションや市民科学など。外部の人を招き入れ、共に価値を生み出す。

3種類の研究データ

研究資源データ

研究の入力となるデータ。観測データや評価用データセットなど。**再利用のためのオープン化**が求められる。

論文付属データ

研究の出力となるデータ。論文のエビデンスとなるデータなど。**透明性のためのオープン化**が求められる。

研究過程データ

研究の入力と出力の間で生み出されるデータ。日々の研究活動のエビデンスとなるデータなど。**研究不正防止（透明性）のためのクローズドな長期保存**が求められる。

事前に提出した講演概要

- 講演者自身の定義は、データの整理・編集・展示などを通してデータを活用しやすい形で広く公開する活動、というものである。
- この概念には、科学研究に関連するデータの循環を通して、オープンイノベーションや超学際的研究などを進め、さらに研究の透明性を高めることへの期待が込められている。
- 一方、これほど幅広い知識とスキルを有する概念に対して、それを実際に推進する人材をどう育成してどう評価するかが重大な課題として残っている。

人文科学データのキュレーション

人文学オープンデータ共同利 用センター（CODH）

- ・情報・システム研究機構 データサイエンス共同利用基盤施設が2016年4月1日に準備室を開設。2017年4月1日にセンター発足予定。 <http://codh.rois.ac.jp/>
- 1. 情報学・統計学の技術を用いて人文学の研究を行う。
- 2. 人文学のデータを用いて情報学・統計学の研究を行う。

CODH／NII／国文研の共同研究

人文学オープンデータ
共同利用センター
人文学データの研究者
や市民による利用を促
進するオープン化拠点。

歴史的典籍NW事業
日本の歴史的典籍30
万点をデジタル化し、
国際共同研究を推進す
る大型プロジェクト。

情報・システム研究機
構およびNII・統計数
理研究所が関わる。

人間・文化研究機構
国文学研究資料館が中
心的な役割を果たす。

情報学と人文学の協働により歴史的典籍の課題を解決。

研究者のためのオープンデータ

11月10日プレスリリース



日本古典籍データセット（国文研所蔵）

オープンデータとデジタルヒューマニ
ティーズ

日本古典籍データセット

- 2015年11月「国文研古典籍データセット」（350点）をNIIから公開。
- 2016年11月「**日本古典籍データセット**」（**700**点）をCODHから公開。
- 画像ファイルに加えて、書誌メタデータや専門家が付与したタグデータも同梱。
- 翻刻テキストは一部の古典籍のみに付属。
- ライセンスはCC BY-SA 4.0とする。

画像公開でのIIIF活用

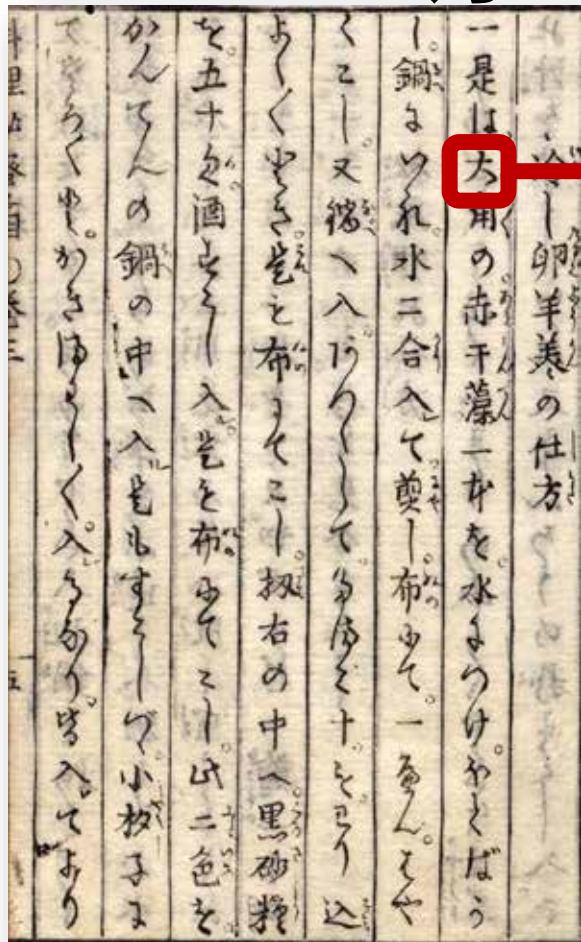


IIIF Curation Viewer

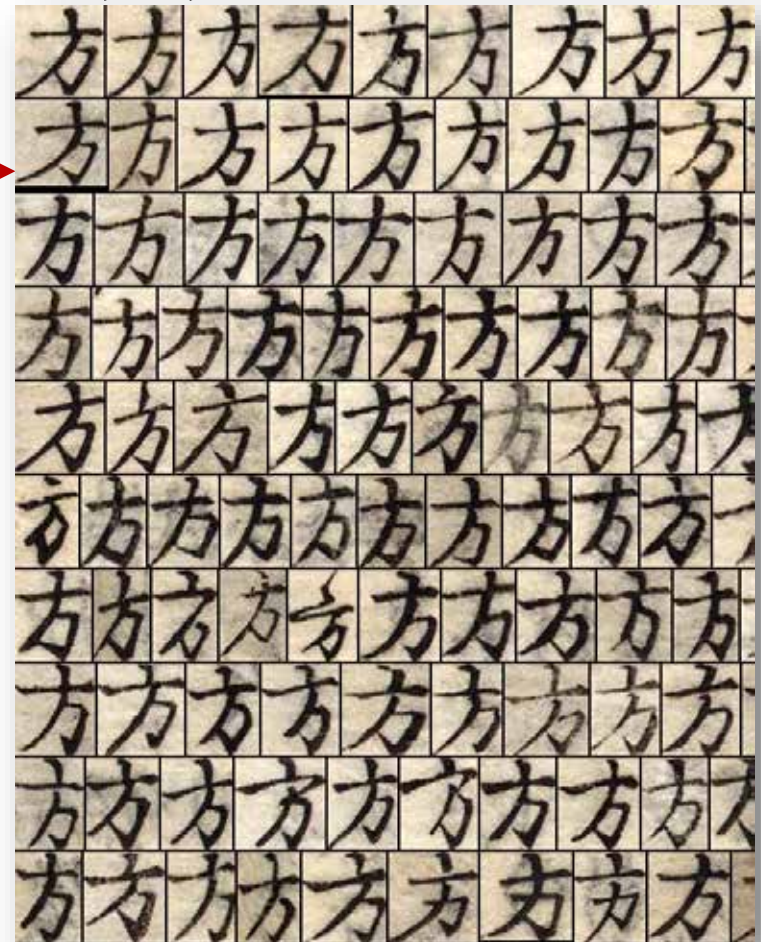
- IIIF (International Image Interoperability Framework)による多解像度の画像閲覧。
- 既存のビューアーに満足できず、独自のビューアーを構築。
- Core contributor: Jun HOMMA (@2SC1815J)

機械のためのオープンデータ

11月17日プレスリリース



日本古典籍データセット
(国文研所蔵)



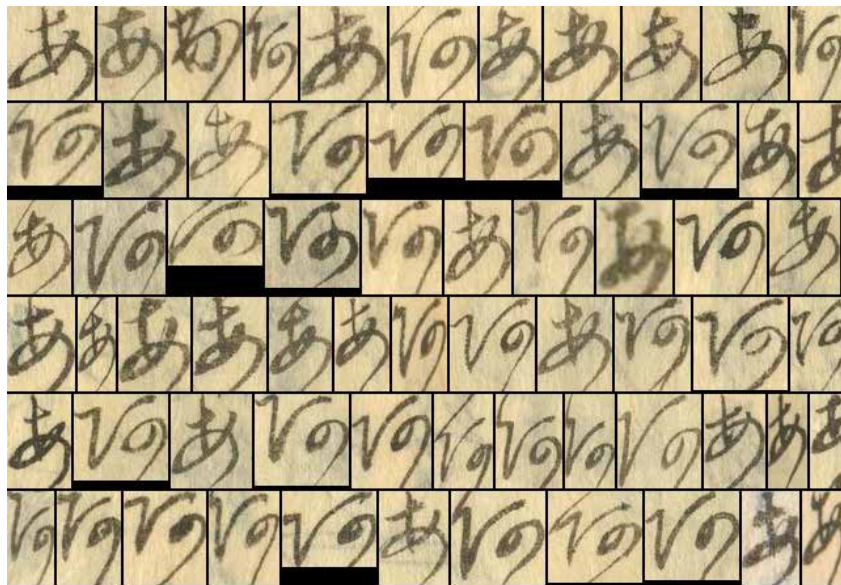
日本古典籍字形データセット
(国文研所蔵・CODH加工)

機械のための学習データ

文字種	文字数
し	3,929
に	3,147
の	2,908
て	2,398
り	2,193
を	2,021
か	1,910
く	1,739
き	1,715
も	1,463
1,521文字種	86,176文字

- 機械学習用のデータセットとして利用。
- **深層学習**ライブラリ Kerasを用いたCNNサンプルプログラムを配布。
- **座標情報**が存在するため、文字単位を越えた問題設定も可。
- **字母**の扱いは将来課題。

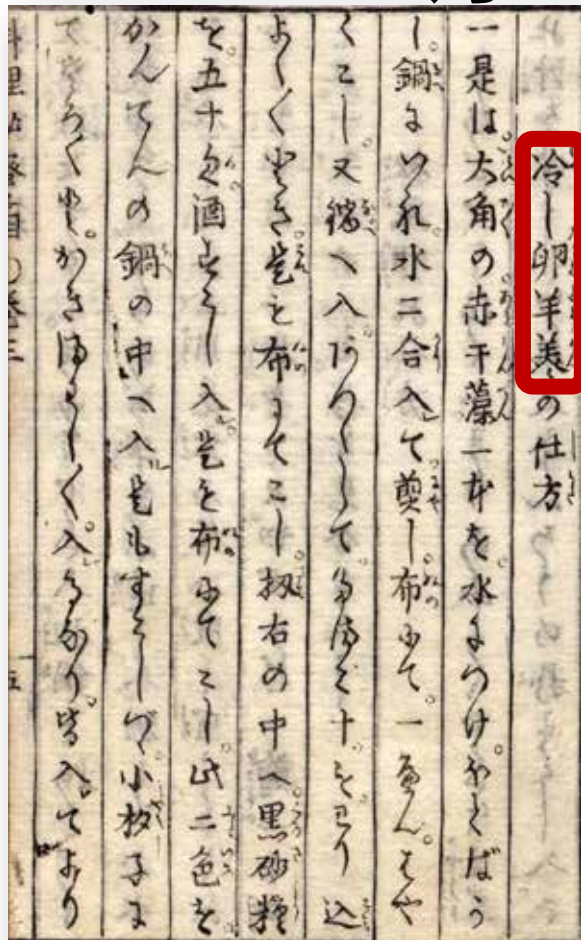
人間のための学習データ



- **字形のバリエーション**を目で見て確認できる。
- **学習アプリの素材データ**にもなる。
- **くずし字を読める人を増やす**ことが、データ利活用の好循環を生み出す。

市民のためのオープンデータ

11月24日プレスリリース



日本古典籍データセット
(国文研所蔵)

2017/01/18



江戸料理レシピデータセット
(CODH制作)

日本古典籍データセット
(国文研所蔵) を翻案

オープンデータとデジタルヒューマン
ティーズ

江戸料理レシピデータセット

- 
1. 江戸の料理本をデジタル化
 2. くずし字を翻刻
 3. 翻刻を現代語訳
 4. 現代語訳をレシピ化・公開
 5. クックパッドでもレシピ公開
 6. つくれぽで個人の経験を共有

協力：合同会社AMANE

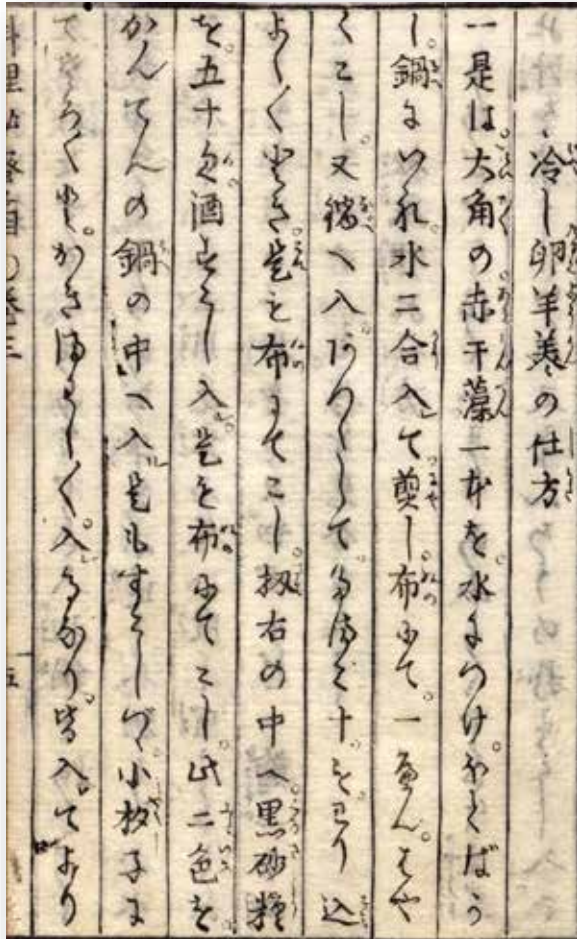
1. 江戸の料理本をデジタル化



- 『**万宝料理秘密箱**（まんぼうりょうりひみつばこ）』
- 初版は1785（天明5）年。
- 素材としての卵一種で100の料理のバリエーションを楽しむ「**卵百珍**」。
- 「**百珍物**」：食べるための食事から、食を楽しむ生活への変化という時代背景。

日本古典籍データセット
（国文研所蔵）

2. くずし字を翻刻

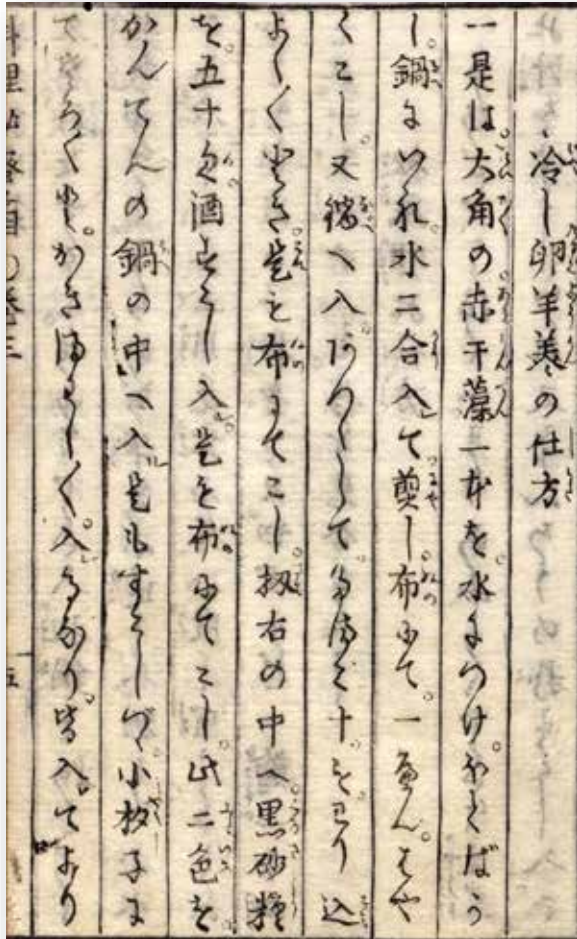


日本古典籍データセット
(国文研所蔵)

1	是は 大角の 赤干藻一本を 水につけ ほとばかし
2	鍋にいれ 水二合入して 煎し 布にて 一へん はやくこし 又鍋へ入レ あつくして
3	たまご十ウを わり込よくよくとき 是も布にてこし
4	扱右の中へ 黒砂糖を 五十匁 酒すこし入ル 是も布にてこし
5	此二色を かんてんの鍋の中へ入ル
6	是もすこしづつ 小杓子にて そろそろと かきまわしかきまわし 入れるなり
7	皆入れてより 又葛粉をすこし 水にてとき入レ
8	扱鍋をぬき 早く折敷にても うちあげ 平めに延し 入レ物ともに 水に入レ 冷し遣ふ

江戸料理レシピデータセット (CODH制作)

3. 翻刻を現代語訳



日本古典籍データセット
(国文研所蔵)

2017/01/18

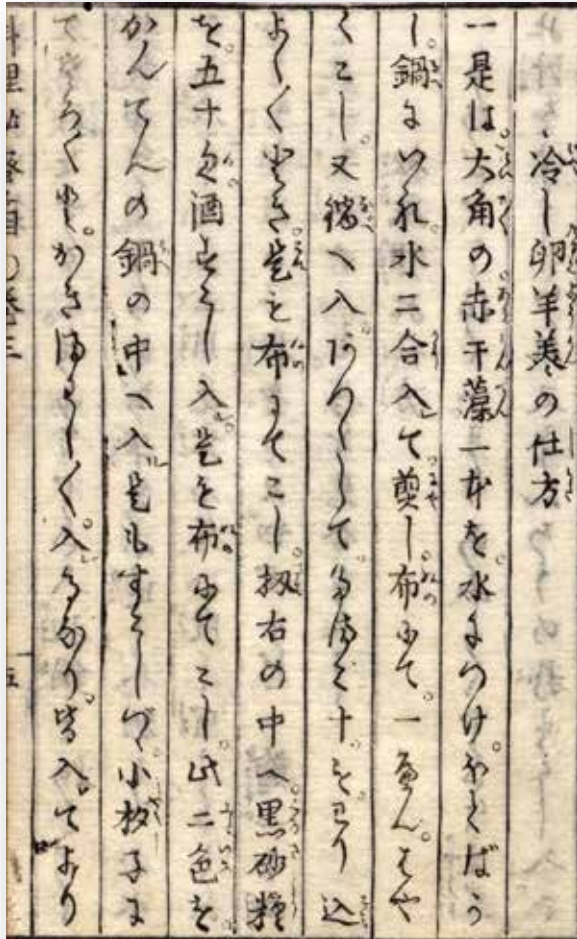
1	大きな赤寒天を1本水に付けてふやかす。
2	鍋に寒天と水2合（360cc）を入れて煮溶かす。
3	②を一度布で素早く漉し、再び鍋に入れて熱する。
4	生卵10個をよく溶き、布で漉す。
5	④の中に黒砂糖50匁（200g）と酒少しを入れ、布で漉す。
6	⑤を寒天の鍋に入れる。小さな杓子で少しずつそろそろと混ぜながら入れる。
7	⑤を全て鍋の中に入れたら、葛粉を水で溶き、鍋に入れる。
8	鍋を火から上げ、素早く中身を容器（折敷）に広げ、平たく延ばし、容器ともに水で冷やす。

江戸料理レシピデータセット（CODH制作）

オープンデータとデジタルヒューマニ
ティーズ

21

4. 現代語訳をレシピ化



日本古典籍データセット
(国文研所蔵)

1	寒天を水につけて、ふやかします。
2	生卵をよく溶きます。
3	溶いた生卵を布でこします。
4	黒砂糖と酒を入れ、溶かします。
5	4を3に入れ、再びこします。
6	鍋に寒天と水（180cc）を入れて煮とかします。
7	6を布などでこし、再び鍋に入れて熱します。
8	7の熱した寒天の中に、5の卵液を少しずつ入れます。
9	全て入れ終わったら、水でといた片栗粉を鍋に入れてさっと混ぜ合わせます。
10	鍋を火からあげ、中身を容器に入れます。
11	冷蔵庫で、2時間程度冷やします。

江戸料理レシピデータセット（CODH制作）

文字と写真による調理手順



江戸料理レシピデータセット (CODH制作)

オープンデータとデジタルヒューマニ
ティーズ

現代の状況に合わせたレシピ

- **材料の違い**：「葛粉」を「片栗粉」へと、入手しやすい材料に変える。
- **道具の違い**：「容器ごと水で冷やす」を「冷蔵庫で2時間程度冷やす」へと、現代の道具に合わせて変える。
- **分量の違い**：「卵10個」を「卵5個」へと現代の生活にあった分量に変える。

現代の文化を反映したレシピに「翻訳」

過去のことば・道具・文化

- **ト活卵**：蒲鉾を藁の太さ程に切る。
- **小豆餅卵**：ゆで卵を水飴の中に入れ、表面にくるりと塗る。
- **鶉卵**：鶉か鳩か□か雲雀か鴨か、その他どのような鳥にしても、鳥肉を丁寧によく叩く。
- **家主貞良卵**：上には、行灯の火皿を乗せ、この中に灰をうすく引き、火を入れて焼く。

江戸時代の文化を感じさせる記述から、過去の生活への想像を膨らませる。

江戸料理レシピデータセット

データ概要	説明
原本画像データ	日本古典籍データセットで公開する画像です。くずし字を読み、かつ江戸時代の日本語や料理法を知っていれば料理が作れます。
翻刻テキストデータ	原本画像のくずし字をテキスト化したデータです。江戸時代の日本語や料理法を知っていれば料理が作れます。
現代語訳データ	翻刻テキストデータの内容を現代の日本語に翻訳したデータです。江戸時代の料理法を知っていれば料理が作れます。
現代レシピデータ	現代語訳データの内容を、現代の道具や食材でも作れるものに実装し、食材の分量や写真を加えてより具体化したデータです。手順に従えば料理が作れます。

1. **翻刻**：全**107**点
2. **現代語訳**：翻刻107点中**20**点
3. **レシピ化**：現代語訳20点中**5**点

オープンデータ (CC BY-SA)
としてウェブサイトで公開。

<http://codh.rois.ac.jp/edo-cooking/>

クックパッドでもレシピ公開

クックパッドと日本家政学会 食文化研究部会が運営する「クックパッド江戸ご飯」に参加。



多くの人々が既に馴染んでいるウェブサービス（アプリ）からも、江戸料理のレシピを活用。



<http://cookpad.com/recipe/4153357>

極めて大きな反応



<https://twitter.com/caille2006/status/802575840819089409>

2017/01/18



1074 リツイート

<https://twitter.com/jouhouken/status/801693251052781568>

作って食べてみた

0テレ ニュース・情報 ドラマ パラエディ・音楽 スポーツ 映画 アニメ ミニ番組 イベント 通販 動画 プレゼント

0テレNEWS24 2016年(平成28年) 12月5日(火) 今日の実況 仙台

人は、ふれあって育つ。 **UR賃貸住宅** 社宅も **UR** 礼金 ¥0 手数料 ¥0 更新料 ¥0 保証人 不要

0テレNEWS24 ストレイニュース every. NEWS ZERO ガンキョウ 山田孝之 ドキュメント 深層NEWS Daily Planet 0771 NEWS 24

トップ 社会 エンタメ 国際 政治 経済 スポーツ 投稿 ブログ 特集 地方 天気 地震 防災

トップ > 社会 > 「江戸ご飯レシピ」再現、作って食べてみた

0テレNEWS24 「江戸ご飯レシピ」再現、作って食べてみた 2016年11月30日 21:05

ツイートする シェアする

 江戸ご飯レシピ Hot Word

全文

11月28日、「江戸ご飯レシピ」という言葉を含んだツイートが大きく伸びました。

毎日「寄り添うガウチョパンツ」 2,149円(税込) soulberry 2016 Hello Autumn

「いま見るべき」「いま知るべき」厳選ニュースをメルマガでお届け

Q ニュース検索 検索

カレンダー検索

★ アクセスランキング 社会

- 1 豊中市妊婦刺殺「経歴はなく事故」と否認
- 2 電車で女性の髪切り「アクションで見る」
- 3 千葉大医学主3人の署名公表 医師も逮捕
- 4 女性の押画術「ばらまくぞ」航空自衛官逮捕
- 5 重慶の重婚スタッフ偽りけいしつ 黒澤連

<http://www.news24.jp/articles/2016/11/30/07347892.html>

オープンデータとデジタルヒューマニ
ティーズ

得られた教訓

- 「データをクックパッドでも公開する」と説明すると、**人々の顔がぱっと輝く**。
- 仮説：クックパッドという現代と江戸料理という過去の**ギャップ**に**意外感**あり。
- 仮説：馴染みのプラットフォームに入ると、「**自分たちのデータ**」とを感じる。
- **市民のためのオープンデータ**：どこで公開するかで、インパクトが変わる。

自然科学データのキュレーション

デジタル台風

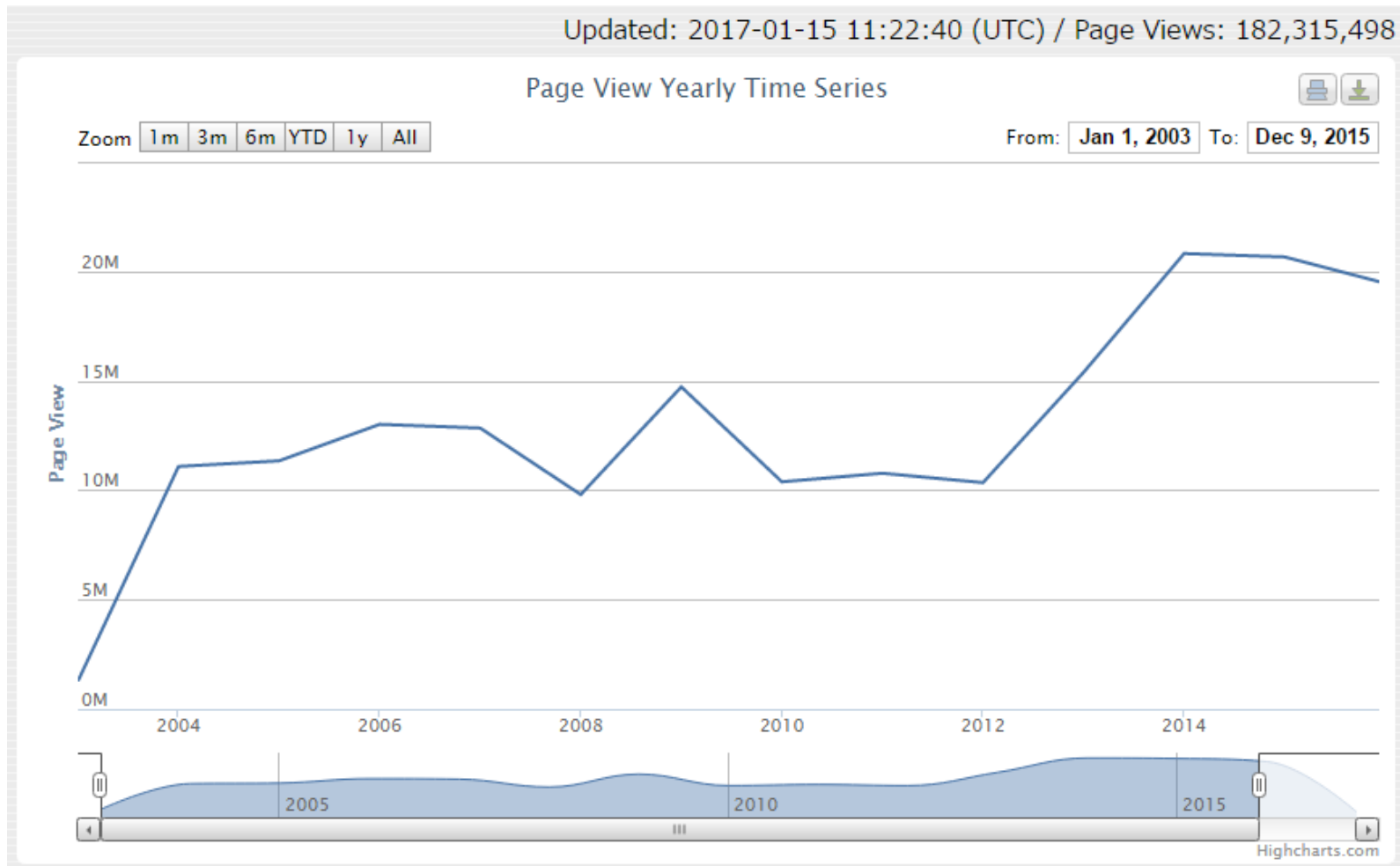
<http://agora.ex.nii.ac.jp/digital-typhoon/>

- 台風に関する「**あらゆる情報**」を集約。
- **最新状況の把握と過去アーカイブの検索**がシームレスに接続。
- 利用実績は**年間約2000万ページビュー**。
- **研究者のみならず、一般市民も多数利用。**

アクセス実績推移

<http://agora.ex.nii.ac.jp/digital-typhoon/access/yearly/>

Updated: 2017-01-15 11:22:40 (UTC) / Page Views: 182,315,498



ストリームからメディアへ

- **センサからの観測データ**：時々刻々生成され、その意味もダイナミックに変わる。
- **データの目利き**：流れ行くデータの中から、価値あるデータを拾い出して見せる。
- **現在の文脈との結合**：過去のデータと紐づけると、新たな価値が見いだせる。
- **キュレーション駆動開発**：データの価値を展示するシステムを即興的に開発。

[このページの著作権](#)
[記事の再利用について](#)
[お問い合わせ](#)
[お問い合わせ](#)
[お問い合わせ](#)
[お問い合わせ](#)
[お問い合わせ](#)
[お問い合わせ](#)
[お問い合わせ](#)
[お問い合わせ](#)

デジタル台風：2013年台風30号（ハイエン） HAIYAN

ホーム Earth デジタル台風 ブログ 2013年 台風30号 English

台風201330号（HAIYAN）

台風の名称：ハイエン（HAIYAN） うみつゆめ（中国）

- デジタル台風：台風画像と台風経路図（QRコード／ケータイ画面表示用）
- デジタル台風：ニュース・トピックス
- 気象庁(PMA)：台風情報・予想経路図 / 台風進路予報図 [Google Maps/グーグルマップ版]
- 中央気象台気象観測センター(JWC)：台風情報・予想経路図（?）
- 気象庁気象研究所：気象庁防災関係Web（?）
- 気象庁Web：この台風の経路とトラックバック（気象庁Web）
- ライフーン：この台風へのトラックバック（トラックバックURL）
- 台風空想：この台風に関する空想（台風空想）
- ツイフーン：この台風に関するつぶやき（ツイフーン @tweephonr）
- 台風なう！：この台風に関する問合せ（台風なう！ / @typhoonnau）

2015

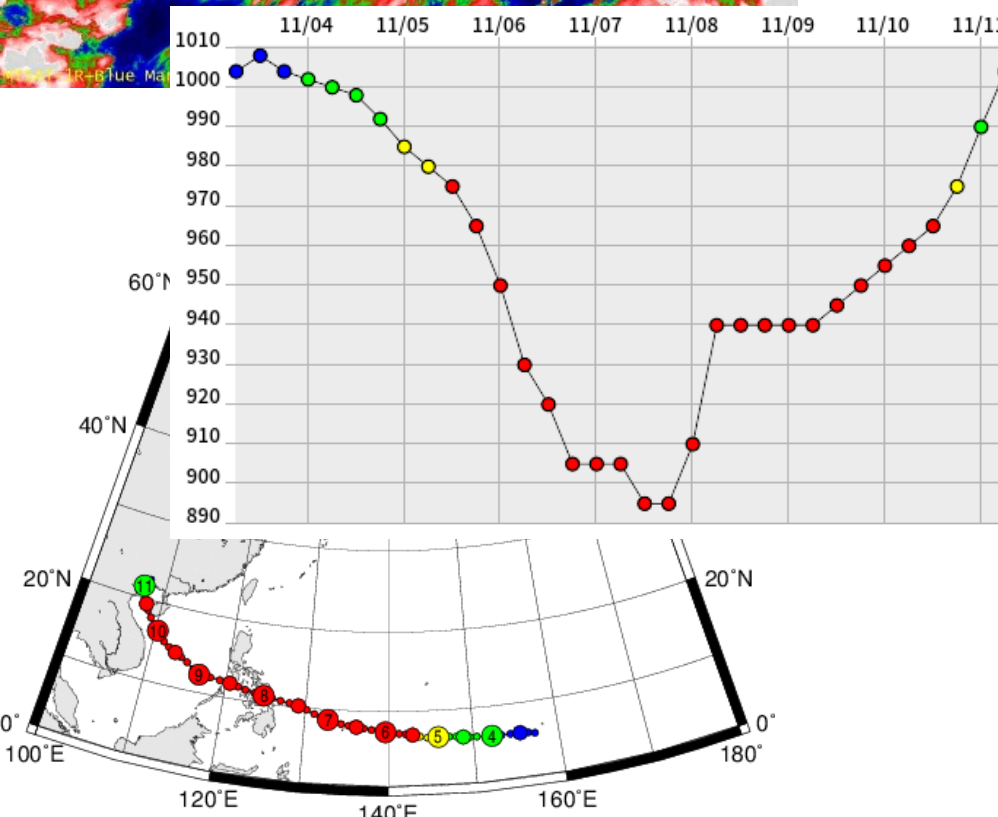
- 2015年台風25号（サンディ）(SANDY)
- 2015年台風24号（コップ）(COPPU)
- 2015年台風23号（オースティン）(AUSTIN)
- 2015年台風22号（ハンジク）(HANJIKU)
- 2015年台風21号（ドゥラン）(DURAN)
- 2015年台風20号（ウツパ）(UTUPA)
- 2015年台風19号（ヴァムコ）(VAMCO)
- 2015年台風18号（ステラ）(STELLA)
- 2015年台風17号（アタナ）(ATANA)
- 2015年台風16号（コンゴ）(CONGO)
- 2015年台風15号（モズベ）(MOZUBE)
- 2015年台風14号（ワシントン）(WASHINGTON)
- 2015年台風13号（ハリス）(HARRIS)
- 2015年台風12号（フィニー）(FINNEY)
- 2015年台風11号（リンダ）(LINDA)
- 2015年台風10号（オースティン）(AUSTIN)
- 2015年台風9号（ハリソン）(HARRISON)
- 2015年台風8号（クリス）(CHRIST)
- 2015年台風7号（ジョージ）(GEORGE)
- 2015年台風6号（ヘレン）(HELEN)
- 2015年台風5号（イレーヌ）(IRENE)
- 2015年台風4号（ケイト）(KATE)
- 2015年台風3号（フラン）(FRANK)
- 2015年台風2号（ミカエル）(MICHAEL)
- 2015年台風1号（アリス）(ALICE)

2013年12月28日 23:00 JST

台風30号(HAIYAN)による被害は近年のフィリピン災害史上最大級のものとなっています。6-7mの高さにつながる風速が多くの人命を奪い、死者は6000人以上、行方不明者は1700名以上の人的被害となりました。また避難者は400万人程度、被災者は1200万人にも達するようですよ。これだけたくさんの方が避難を余儀なくされているのは、強風に多くての豪雨での被害を受けたことが影響しており、これは規模は巨大ですがフィリピンでこれまでより多岐にわたるタイプの被害です。一方、人的被害はその多くが高潮によるものであり、これは今までのフィリピンでは少なかったタイプの、想定外の被害と言えるかもしれません。（参考：[短絡的な見方から「高潮」の警報＝フィリピンの台風被害が拡大した理由](#)）。また、台風の問題によって南門が開かれたことで、避難先から居住区に戻って産業施設にあったケースもあるようです。この大災害を教訓として、今後高潮に対する被害をどうやって防いでいくか、また高潮に関する情報伝達と教育普及をどう改善していくかが、フィリピンにとっての大きな課題となります。

土木学会海岸工学委員会がまとめた「2013年台風(Haiyan)によるフィリピン災害には、「土木学会・フィリピン土木学会合同台風3.0号津波早期警戒グループ」がまとめたものや、他の観測データやシミュレーション結果があります。また、2013年フィリピン高潮合同調査会によも、現地の観測データ（高潮観測局）があります。ただし、シミュレーション結果は観測データをうまく再現できていない点があるため、シミュレーションでは考慮されていないメカニズムによる順位上昇があったのではないかと考えられています。高度な高度の予測による津波のような問題の発生は、専門家にとっても想定外の事象だったと言えるかもしれません。今後どうやってメカニズムを理解し、それを各地の防災対策に活かしていくかが、専門家にとっても大きな課題となります。

ではフィリピンでは、どのような防災、減災対策を進めれば良いのでしょうか。もちろん、堤防を築くとか、海浜に住まないとか、堤防式決壊対策などありますが、フィリピンにおいてはコストがかかりすぎる対策は現実的でない点もあります。最初期シェルターを作るのも一案ですが、これも国際的な援助が必要そうです。そこでもう少し現実的な案として、天然の堤防としてのマングローブの活用があります。今回の高潮被害でも、マングローブの緑林を進んでいた場所の方で被害がなかったために、フィリピン当局は今後はマングローブの植林を進めていくとのことです。もちろんフィリピン沿岸にはマングローブが広がっており、それが育ち盛んな状態になっていたのですが、それが枯死してしまったかと書かれ、エビの養殖池を作るための伐採が主な原因とのこと。そして、そこを養殖されたエビが日本にも大量に輸出されていることを考えると、この災害は決して日本人に無関心なこととは言えません。マングローブは生態系保全や二酸化炭素削減にも効果があります。自然のメカニズムをうまく利用して被害を軽減する方法について、知恵を出しながら考えていくことが必要です。



データに対するバイアス

フレーミング：論理的には等価の記述が異なる反応（行動）につながる。

- 術後一か月の生存率は90%です。
- 術後一か月の死亡率は10%です。

前者の表現の方が、手術を選んだ人が圧倒的に多かった。



客観的事実は同じでも、表現から連想する内容は異なる。

参考：ダニエル・カーネマン「ファスト&スロー」、早川書房、2012

データを解釈する文脈

順位	降水量	開始日時	終了日時	グラフ	気象災害
1	800.0	2004年09月28日18時 (JST)	2004年09月29日17時 (JST)	描画	検索
2	759.0	2008年09月18日19時 (JST)	2008年09月19日18時 (JST)	描画	検索
3	679.0	2011年09月03日18時 (JST)	2011年09月04日17時 (JST)	描画	検索
4	584.0	2001年10月09日16時 (JST)	2001年10月10日11時 (JST)	描画	検索
5	557.0	2001年08月21日02時 (JST)	2001年08月22日01時 (JST)	描画	検索
6	531.5	2011年07月19日03時 (JST)	2011年07月20日02時 (JST)	描画	検索
7	529.0	1991年09月18日10時 (JST)	1991年09月19日09時 (JST)	描画	検索
8	508.0	1	データはできるだけ手 ための文脈と共に表示 脈化)。文脈とは、 現、マルチメディア表 る尺度などがある。		
9	507.0	1			
10	497.0	1			
10	497.0	2			
12	496.0	2			
13	472.0	2			
14	430.0	2			
15	423.0	2			
16	421.0	2			
17	408.0	1			
18	400.0	2			
19	389.0	2003年08月08日09時 (JST)	2003年08月09日08時 (JST)	描画	検索
20	386.0	2012年05月01日23時 (JST)	2012年05月02日22時 (JST)	描画	検索

データはできるだけ孤立させず、解釈のための文脈と共に表示する（データの文脈化）。文脈とは、ランキングや確率表現、マルチメディア表現、人が実感できる尺度などがある。

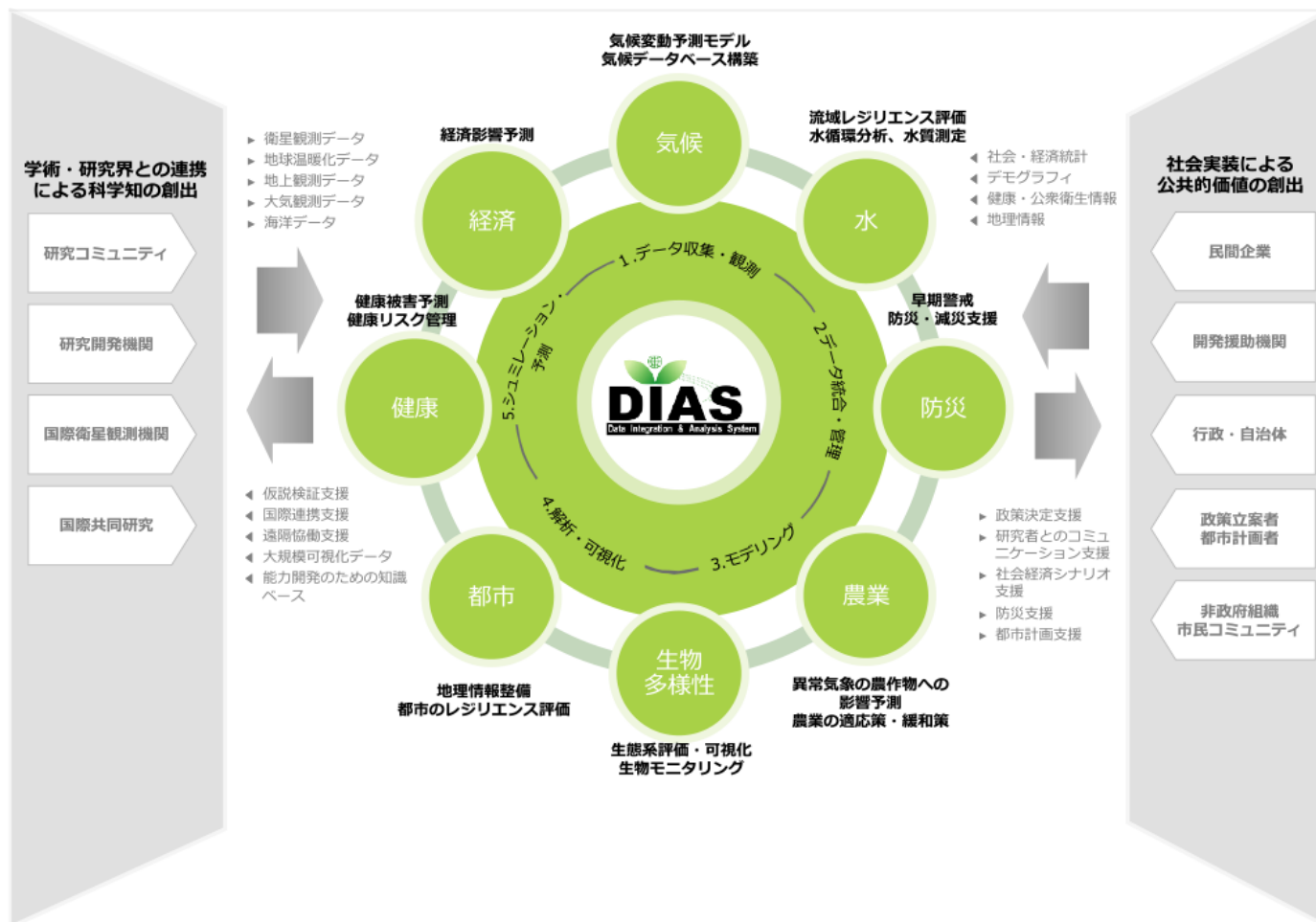
順位	降水量	開始日時	終了日時	グラフ	気象災害
1	119.0	2007年08月01日09時 (JST)	2007年08月02日08時 (JST)	描画	検索
2	118.0	2003年08月07日20時 (JST)	2003年08月08日13時 (JST)	描画	検索
3	102.0	1995年07月24日19時 (JST)	1995年07月25日09時 (JST)	描画	検索
4	97.0	1996年08月17日22時 (JST)	1996年08月18日05時 (JST)	描画	検索
5	93.0	2001年08月22日13時 (JST)	2001年08月23日08時 (JST)	描画	検索
6	88.0	2000年10月01日06時 (JST)	2000年10月01日14時 (JST)	描画	検索
7	87.0	1981年08月04日17時 (JST)	1981年08月05日16時 (JST)	描画	検索
孤立させず、解釈の 示す（データの文 ランキングや確率表 表現、人が実感でき				描画	検索
				描画	検索
				描画	検索
				描画	検索
				描画	検索
				描画	検索
				描画	検索
				描画	検索
				描画	検索
19	69.0	2006年08月21日18時 (JST)	2006年08月22日17時 (JST)	描画	検索
21	67.0	1982年10月24日23時 (JST)	1982年10月25日22時 (JST)	描画	検索

尾鷲の24時間降水量

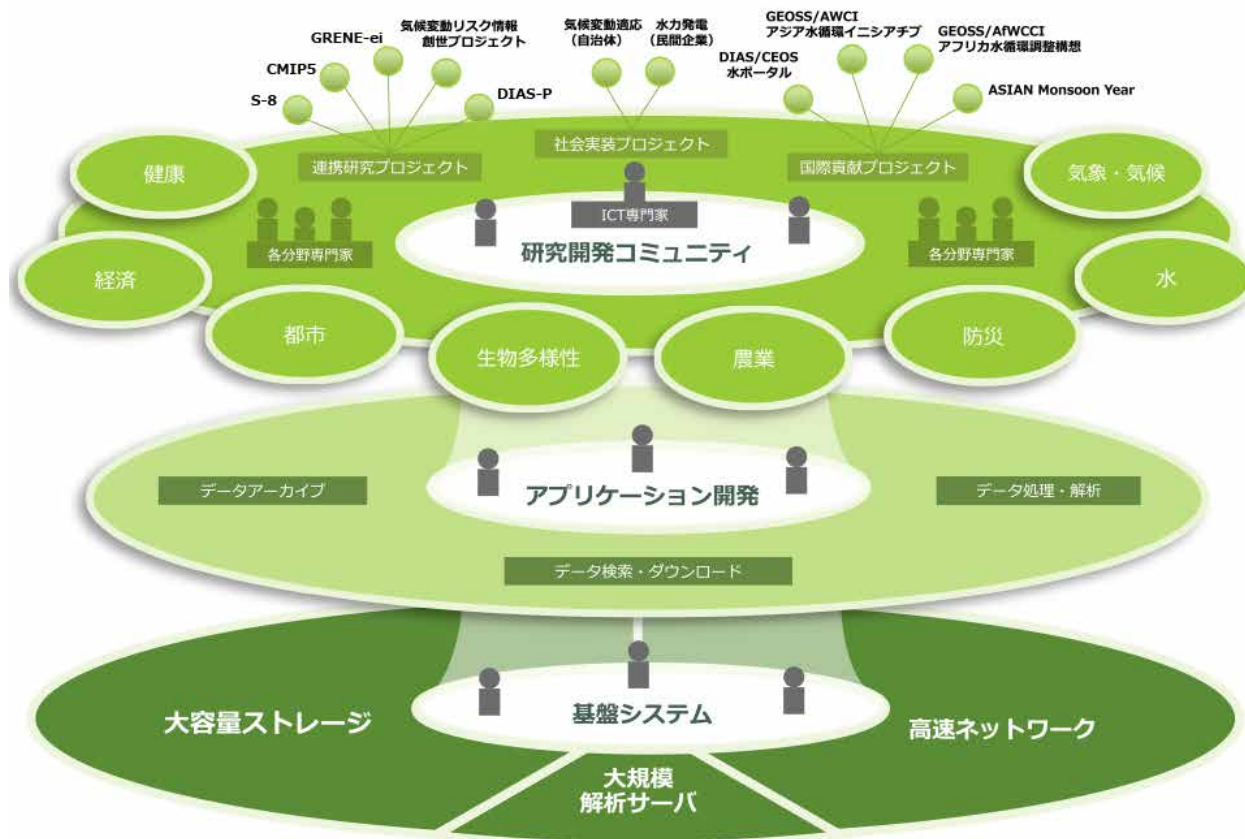
稚内の24時間降水量

DIAS : データ統合解析システム

<http://www.diasjp.net/>



DIAS = データ + コミュニティ



コミュニティ基盤

データ基盤

<http://www.diasjp.net/about/system/>

DIASのデータマネジメント

1. **ドメインの専門家にヒアリング**しながら、メタデータのフォーマットを決める。
2. **メタデータの国際標準**を参照しながら、相互運用性を確保する。
3. **メタデータシステム（管理・検索）**を構築する（または仕様書を書いて外注する）。
4. **メタデータの内容をチェック**し、システム全体として一貫性を保つように改善する。
5. **メタデータ入力を推進**するため、普及活動を開催しながらコミュニティと付き合う。

メタデータ形式の検討

- ⊕ データ提供者が作成しているデータセット紹介のドキュメントを分析
- ⊕ ドキュメントに含まれているメタデータの収集
- ⊕ ISO19115（地理空間メタデータ標準）の項目との対応づけ
- ⊕ 検索に役立つキーワードについての検討

試行錯誤

メタデータ ワーキンググループ

当初検討
サンプル
データ
A1 - A6



データ提供者作成

データセットドキュメンテーション

データ提供者



情報システム
研究者

メタデータ
入力ツール



アプリケーション



入力項目確認

メタデータ標準

キーワード

インタフェース検討

データ分類マトリクス

検索と表示

ドキュメンテーション中の章立て、内容进行分析し、メタデータ項目との対応付けを行う。データセットごとに内容が異なるため当初のサンプルデータに対して詳細な検討を行う。

2017/01/18

メタデータ入力ツールは必須メタデータ項目を含めドキュメントの章立てに対応

オープンデータとデジタルヒューマニ
ティーズ

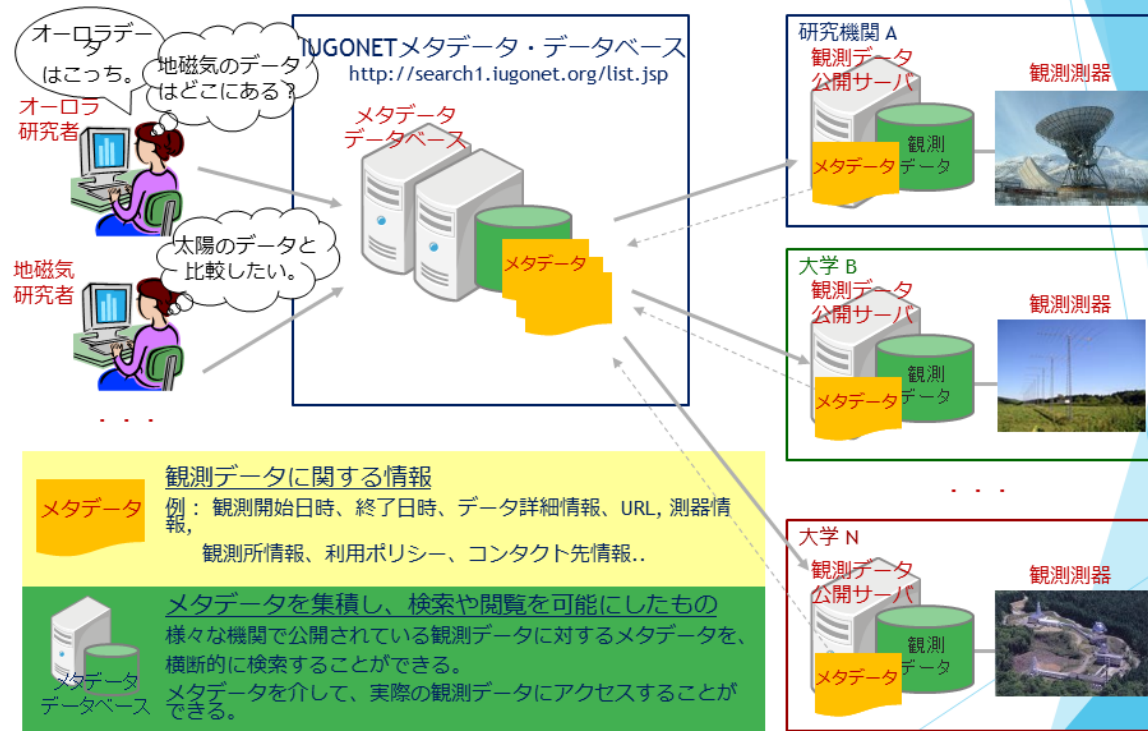
メタデータ入力キャンプ



- メタデータの登録が滞っているデータ提供者に対し、入力方法・入力内容などに関する疑問に答え、入力を促進する場を設定。
- 2014年7月、11月、2015年2月に実施。

IUGONETの事例

IUGONET紹介



サイエンス成果の創出に向けたデータアーカイブと解析環境の融合. 梅村直生ほか
第3回オープンサイエンスデータ推進ワークショップ@京大を一部改変

11

南山 泰之, 超高層大気観測データのメタデータ作成実験経過報告, 第2回 SPARC Japan セミナー2016

オープンデータとデジタルヒューマニ
ティーズ

図書館員がメタデータに挑む

概要

▶ 図書館職員によるIUGONETメタデータ作成を通じて、

1. 研究者へのインセンティブ

- ▶ 研究者、特にメタデータ作成業務従事者の負担軽減
- ▶ 図書館が持つ流通面での知見の提供

2. 図書館による新規サービス展開

- ▶ 研究者との協働による、より直接的な研究支援

の可能性を探る。

南山 泰之, 超高層大気観測データのメタデータ作成実験経過報告, 第2回 SPARC Japan セミナー2016

オープンデータとデジタルヒューマニ
ティーズ

段階的な作業

パイロット開始 (6月～)

図書職員とのコラボレーション

No.	難易度	内容	所感 (図書)	所感 (IUG)
1	低	既存のメタデータ(XML)を機械的に修正する	◎ 容易	◎ OK
2	中-	既存のメタデータ(XML)を新しいスキーマバージョンに則したものに変更する	○ 出来た	○ OK
3	中	PIにインタビューしながら新しくサイエンスメタデータを作成する	○ 出来た	○ OK
4	中	外部サーバを使って単純なデータ公開・更新を行う	(予定)	
5	中+	目前サーバを使って単純なデータ公開・更新を行う	(予定)	
6	高	IUGONETのようなサービスのデータ公開部分を運用する	(予定)	
7	高+	IUGONETのようなサービスのデータ登録部分を運用する	(予定)	

サイエンス成果の創出に向けたデータアーカイブと解析環境の融合. 梅村直生ほか
第3回オープンサイエンスデータ推進ワークショップ@京大を一部改変

南山 泰之, 超高層大気観測データのメタデータ作成実験経過報告, 第2回 SPARC Japan セミナー2016

オープンデータとデジタルヒューマニ
ティーズ

持続可能な体制は作れる

ここまでの検証

- ▶ 既存の情報源をメタデータ化する作業は、図書館の目録作成業務の延長
 - ▶ 標準化されたスキーマや入力ルールが定められていれば、**分野を問わず**対応可能
 - ▶ メタデータ作成業務を集約でき、機関としてのコストが下がる
 - ▶ もっとも、データ取得者による確認作業は省略できない
- ▶ 人的な持続可能性の問題
 - ▶ 作業は慣れや労力を要するが、**マニュアル作成によって標準化・引き継ぎが可能**
 - ▶ メタデータ作成に要する時間は、**図書や地図とあまり変わらない**
 - ▶ 図書館のネットワークで共有することで、全体の知見になる

コード	一般資料種別	コード	特定資料種別
なし	下記のいずれでもないもの	なし	下記のいずれでもないもの
a	地球資料 (cartographic material)	a	天体図 (celestial globe)
		b	惑星儀 / 月球儀 (planetary or lunar globe)
		c	地球儀 (terrestrial globe)
		d	ダイアグラム (diagram)
		e	地図 (map)
		f	断面図 (profile)
		g	模型 (model)
		h	リモートセンシング画像 (remote-sensing image)
		i	断面図 (section)
		j	景観図 (view)
		なし	上記のいずれでもないもの
b	文字資料 (点字) (braille)	なし	
		a	スコア (full score)
		b	ミニチュアスコア (miniature or study size)
		c	鍵盤楽器伴奏譜 (accompaniment reduced for keyboard)
		d	ヴォイススコア (voice score)
		e	コンダクトスコア / ピアノコンダクタースコア (condensed score or piano-conductor score)
		f	クローススコア (close score)
		g	邦楽譜
		h	複合形式の楽譜 (multiple score format)
		i	不明 (unknown)
		なし	上記のいずれでもないもの (other than score format)
d	文字資料 (書写資料) (manuscript text)	GM1	「なし」に対応するSMDのうち適切なコードを用いる
e	地図 (書写資料) (manuscript map)	GM2	「a」に対応するSMDのうち適切なコードを用いる
f	楽譜 (書写資料) (manuscript music)	GM3	「c」に対応するSMDのうち適切なコードを用いる
g	静止画像 (投影) (projected image)	a	フィルムストリップ (カートリッジ) (filmstrip cartridge)
		b	フィルムストリップ (filmstrip)
		c	フィルムストリップ (その他) (other filmstrip type)

目録システムコーディングマニュアル

http://catdoc.nii.ac.jp/MAN2/CM/furoku1_1.html

南山 泰之, 超高層大気観測データのメタデータ作成実験経過報告, 第2回 SPARC Japan セミナー2016

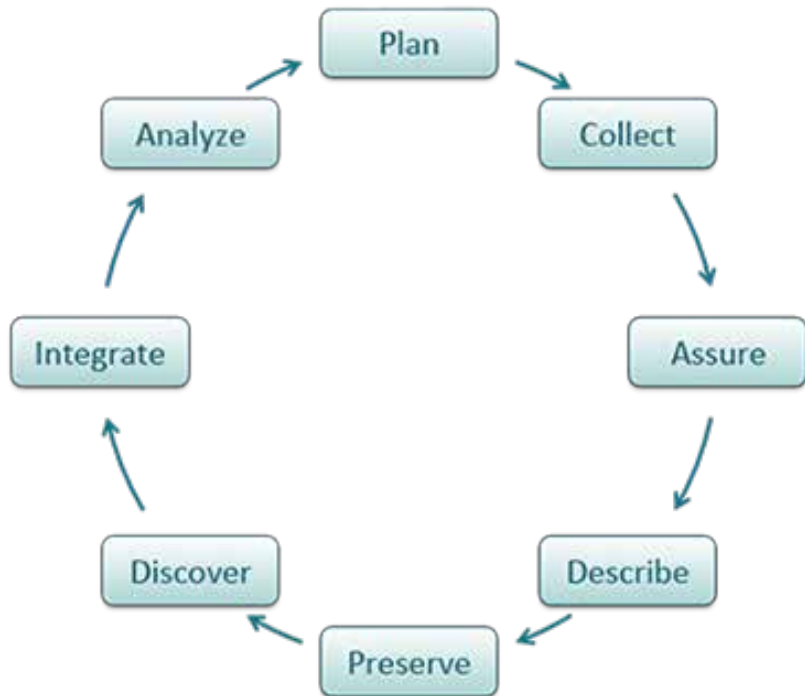
オープンデータとデジタルヒューマニ
ティーズ

研究データとキュレーション

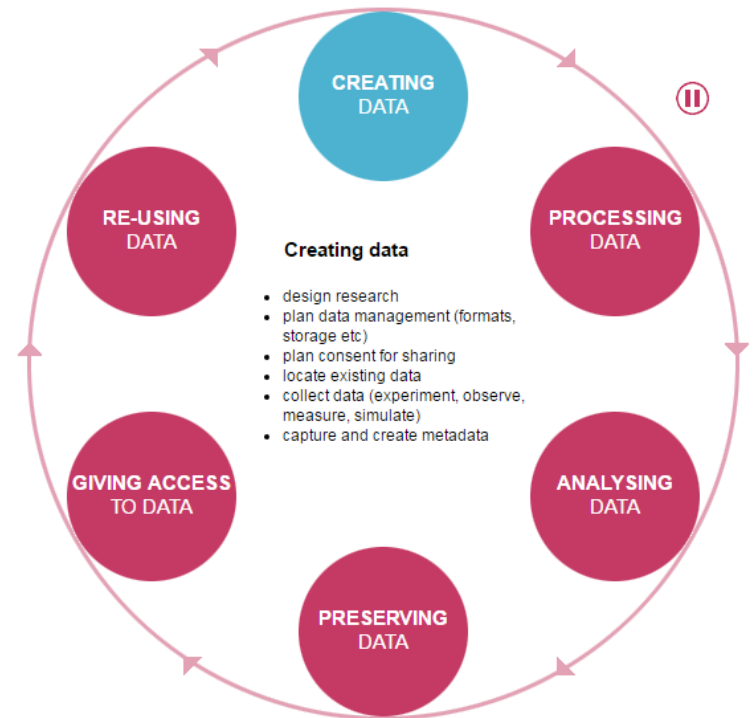
データライフサイクル

- データが誕生してからたどるプロセスを、段階の遷移として表すモデル。
- 唯一の正解はない。分野ごとにライフサイクルの図式は異なる。
- ライブラリアン／キュレーターは、プロセスの一部に関わる。
- プロジェクトに適したライフサイクルを、データの特성에応じて考える。

循環するもの

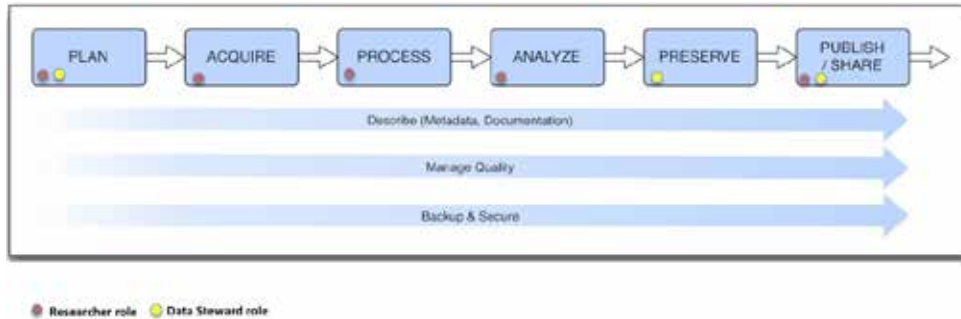


<https://www.dataone.org/data-life-cycle>



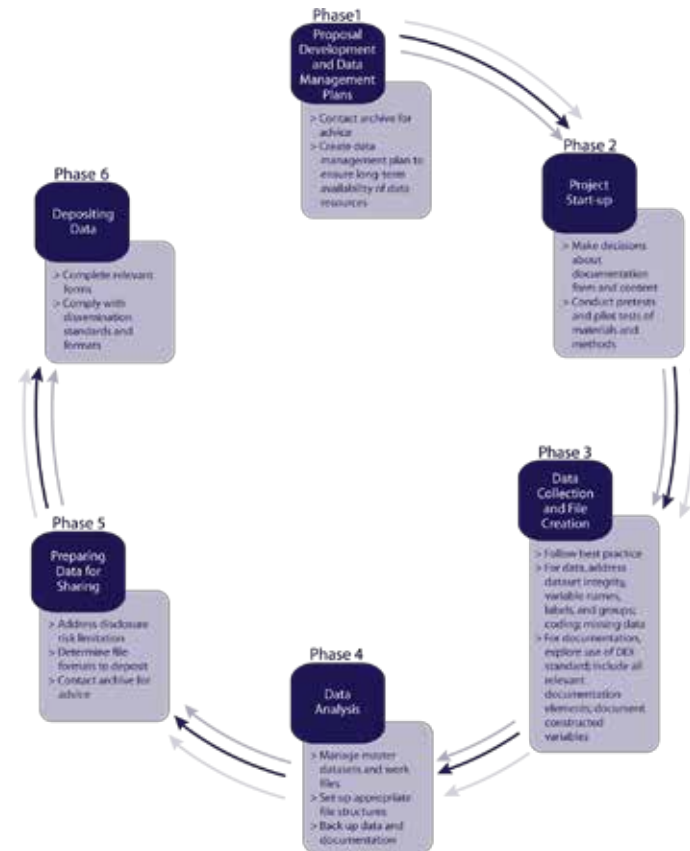
<http://www.data-archive.ac.uk/create-manage/life-cycle>

線形に進むもの



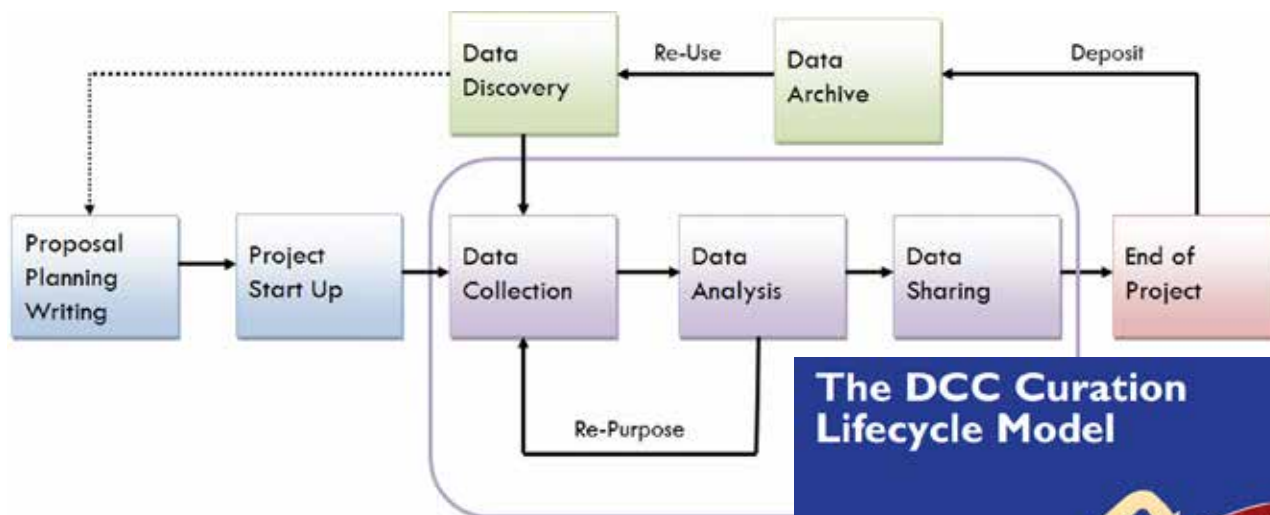
USGS

データスチュワードは
PLAN、PRESERVEを担
い、PUBLISH/SHAREは
研究者との共同作業。

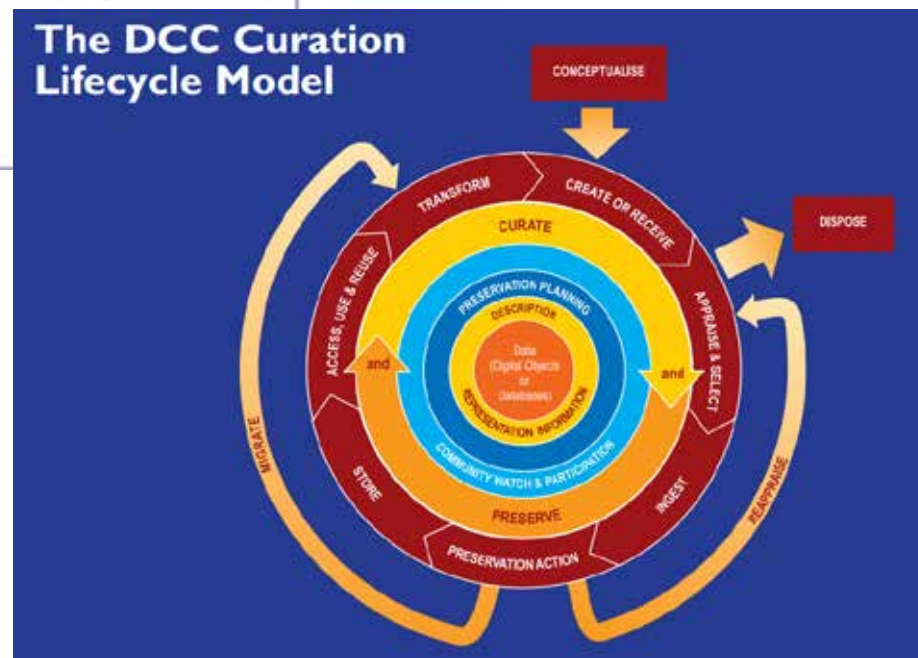


<https://www.icpsr.umich.edu/files/ICPSR/access/deposit/datalifecycle.png>

複雑なもの



<http://data.library.virginia.edu/data-management/lifecycle/>



データライブラリアン

1. データの性質に応じた**データライフサイクル**を定義し、一部を実行する能力。
2. 組織レベルやプロジェクトレベルの**データ管理計画（DMP）**立案。
3. データの利活用と長期保存を見据えた**メタデータ**形式の策定・入力支援。
4. データのデポジットやデータ論文など、**共有と公開**に向けた作業の支援。

求められるスキル@2016年

1. **データライフサイクル**に関する知識とDOIなどによるデータ管理の基礎。
2. **メタデータ**規格や語彙に関する国際標準との相互運用性を踏まえた知識。
3. **データ共有と利活用**を促進する方法に関する基本的な動向の把握。
4. 【上級】研究コミュニティとの議論やエスノグラフィ的観察に基づき、**データ管理システム**を設計し構築する能力。

求められるスキル@2026年？

1. **人工知能がすべてを代替する？** 代替可能なスキルを学ぶのは、人生戦略上まずい。。
2. **データ整理はアクセス手段が目的？** ならば、現状の方法にこだわる必要はない。
3. **人工知能を使いこなせる** スキルは欲しい（すべてのライブラリアンに関係）。
4. **人工知能の下で働く？** 人工知能をより良くするために、より良いデータを入力する。
5. 長期的に維持できる価値は何か？

担い手は若手かシニアか？

- **若手によるデータマネジメント**：キャリアパスを用意して取り組んでもらおう。
- **シニアによるデータマネジメント**：最後のご奉公（？）として貢献してもらおう。
- **若手案の利点**：新しい可能性を開拓し、今後大きく伸びる可能性がある。
- **シニア案の利点**：豊富な知識と経験があるし、キャリアも考えなくていい。。

Digital Curation @ King's College London

Digital Curation MA

Overview

A two-year international programme providing a combination of theoretical knowledge and practical training, equipping you to manage and curate the digital information and digital assets of organisations across public and private sectors. Students study for one year at King's and one year at Humboldt University Berlin, and choose from an exceptional range of options covering different aspects of digital curation.

Reasons you should consider the MA in Digital Curation are:

- You will receive an outstanding education that will equip you to become a leader in the field of digital curation and information management.
- Employers are looking for skilled professionals with knowledge and expertise in managing and curating their valuable digital information and assets.
- You will study in two of Europe's most exciting cities and have the opportunity to...

Key information

Application status: **Open**

Duration: **Two years FT, September to September**. Students spend four semesters over the course of two years on the programme: two consecutive semesters at Humboldt and two semesters at King's, beginning the programme at Humboldt in all cases.

Study mode: **Full-time**

Course location: **Dr Mark Hedges (King's): mark.hedges@kcl.ac.uk; Dr Tobias Blanke (King's): +44 (0) 20 7848 1973; Professor Michael Sauter (Humboldt)**

Course contact for further information: **Postgraduate Admissions Team, Admissions Office tel: +44 (0) 20 7848 7000 fax: +44 (0) 20 7848 7200**

Course contact form: [Postgraduate admissions enquiry form](#)

Awarding institution: **King's College London**

Faculty: **Faculty of Arts & Humanities**

Department: **Department of Digital Humanities**

<http://www.kcl.ac.uk/study/postgraduate/taught-courses/digital-curation-ma.aspx>

- Libraries, museums, galleries, and archives.
- Digitally literate professionals to work in education and heritage institutions.
- Media organisations, publishing houses, government and industry; research institutions and healthcare and law firms.

データキュレーター

- **(元々の) キュレーター**：美術品の「価値」を見だし、展示を企画・実行する。
- **データベースキュレーター**：データの統合やアノテーションを通して、データベースの価値を高める作業を行う。
- **データキュレーター**：編集を通してデータの「価値」を高める仕事。
- キュレーターのキーワードは「**価値**」？

様々なデータ専門職

- **データライブラリアン**：データを組織化する（本来の価値を引き出す）。
- **データキュレーター**：データを編集する（プラスの価値を引き出す）。
- **データアナリスト（データサイエンティスト）**：データを（アルゴリズム的に）分析する。
- **データエンジニア**：システムを構築する。

おわりに

書籍（雑誌）とデータの違い

- **書籍や雑誌**：整理・編集済みの最終生産物だけを扱っていただければよかった。
- **データ**：ライフサイクルの全過程を扱わなければならない。
- **関係者（ステークホルダー）やプロセスが多い**ため、メタデータも複雑化する。
- その分、未開拓の部分も大きく、これから発展するテーマです！！

人材育成

- **データと倫理**：データの品質や利用に対する倫理、（将来の）社会に対する責任。
- データであれ書籍であれ、**人類の知的成果を組織化し継承する**という役割は同じ。
- **外に出てデータの現場を知り**、観察やコミュニケーションをして欲しい。
- 文字と違い**人間はデータが読めない**ので、分析や可視化の「基本」スキルも欲しい。

まとめ

- オープンサイエンスは様々な動きの収束点、大同団結とも同床異夢とも言える。
- 研究データのキュレーションとは、データを編集して新たな価値を見いだすこと。
- 書籍（雑誌）とデータは違いもあるが、人類の知的生産物を扱う点では同じ。
- データ専門職へのニーズは今後高まるので、より多くの人々に参加して欲しい。

関連ウェブサイト

- オープンサイエンス

- <http://agora.ex.nii.ac.jp/~kitamoto/research/open-science/>

- 人文学オープンデータ共同利用センター

- <http://codh.rois.ac.jp/>

- デジタル台風

- <http://agora.ex.nii.ac.jp/digital-typhoon/>

- DIASプロジェクト

- <http://www.diasjp.net/>