人文ＤＸを指向する情報基盤の構築

Construction of an Information Infrastructure for the Humanities DX

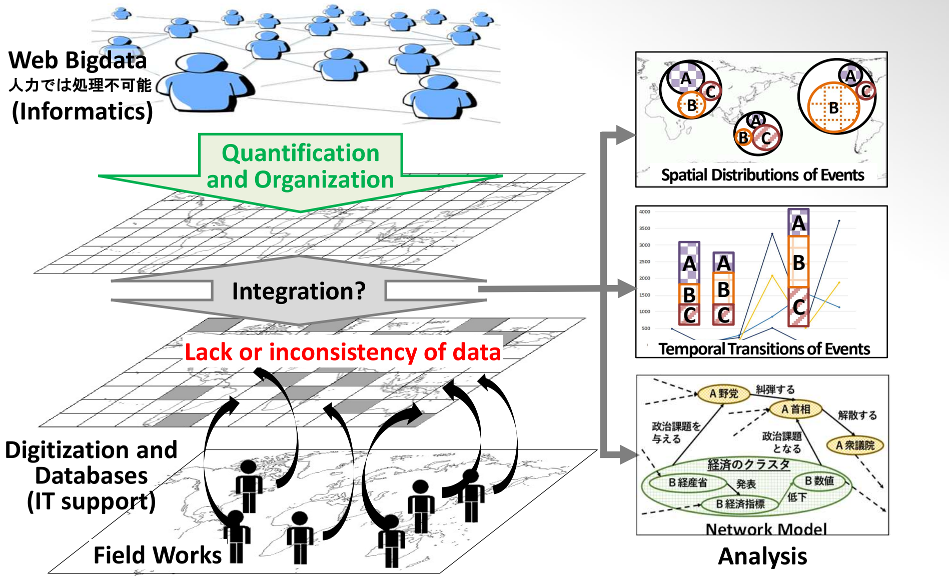
原 正一郎[*1]、馬場 弘樹[*2]、森 信介[*1]、村上 勇介[*1]、関野 樹[*3]、山田 太造[*4]、後藤 真[*5]

Shoichiro Hara*1, Hiroki Baba*2, Shinsuke Mori*1, Yusuke Murakami*1, Tatsuki Sekino*3, Taizo Yamada*4, Makoto Goto*5

1.京都大学(Kyoto University) 2.一橋大学(Hitotsubashi University)
3.国際日本文化研究センター(International Research Center for Japanese Studies)
4.東京大学(University of Tokyo) 5.国立歴史民俗博物館(National Museum of Japanese History)

AI等の活用を推進する研究データエコシステム構築事業シンポジウム
2023/09/29@学術総合センター

モチベーション:情報学で地域研究を



目 的

DX：情報学をドメインとした計量的な地域研究の試みと、データ駆動型地域研究の基盤整備。そこで本研究では、大量のテキストデータの計量的処理をユースケースとして以下の研究を行う。

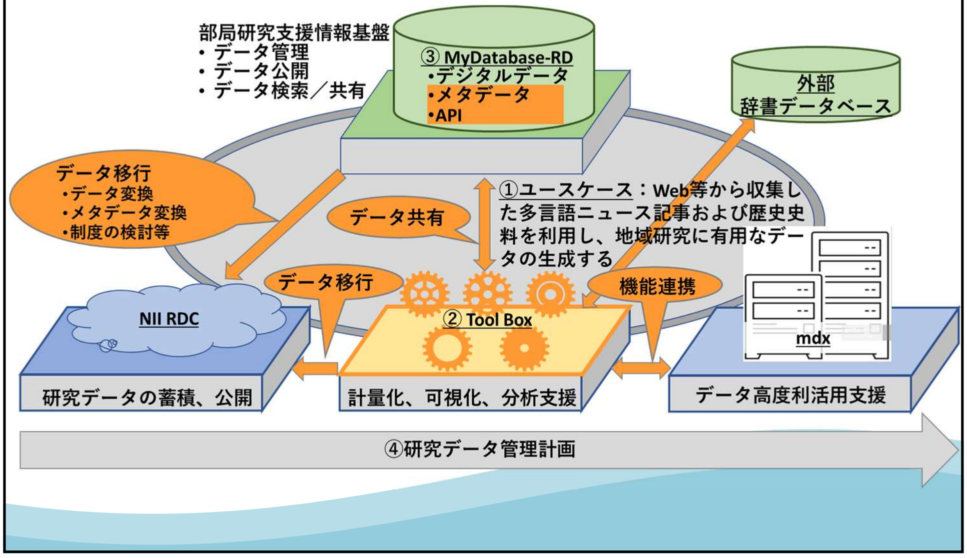
① ユースケース：Webニュースデータから地域変動を検出する計量的手法を開発する（mdxの利用）

② ツール化と公開：データ分析を支援する使いやすい計算環境を、ツールボックスとして構築・公開する（mdxとGakuNin RDMの利用）

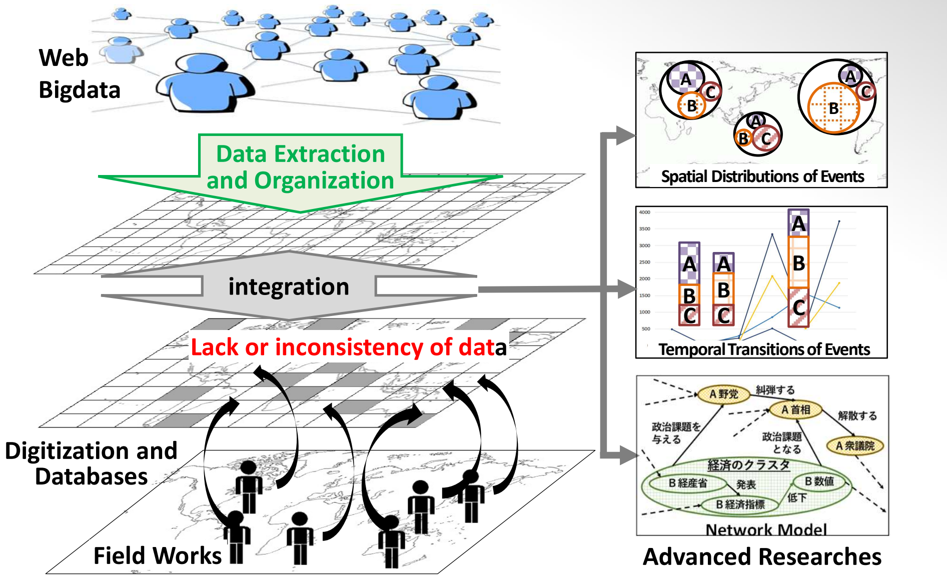
③ データ保存と共有：大量・多様な研究データの長期保存・共有・利活をどのように実現するか？　*まずはローカルなデータシステム構築し*、次に公的（？）な外部サーバとの連携を考える（GakuNin RDM の利用）

④ 研究データ管理手法：これまでのデータ管理は自己流。研究データを適切に管理する知識・スキル・評価の標準的な手法の確立が必要（ Basic Rubric とGakuNin RDM の利用）

研究の概要



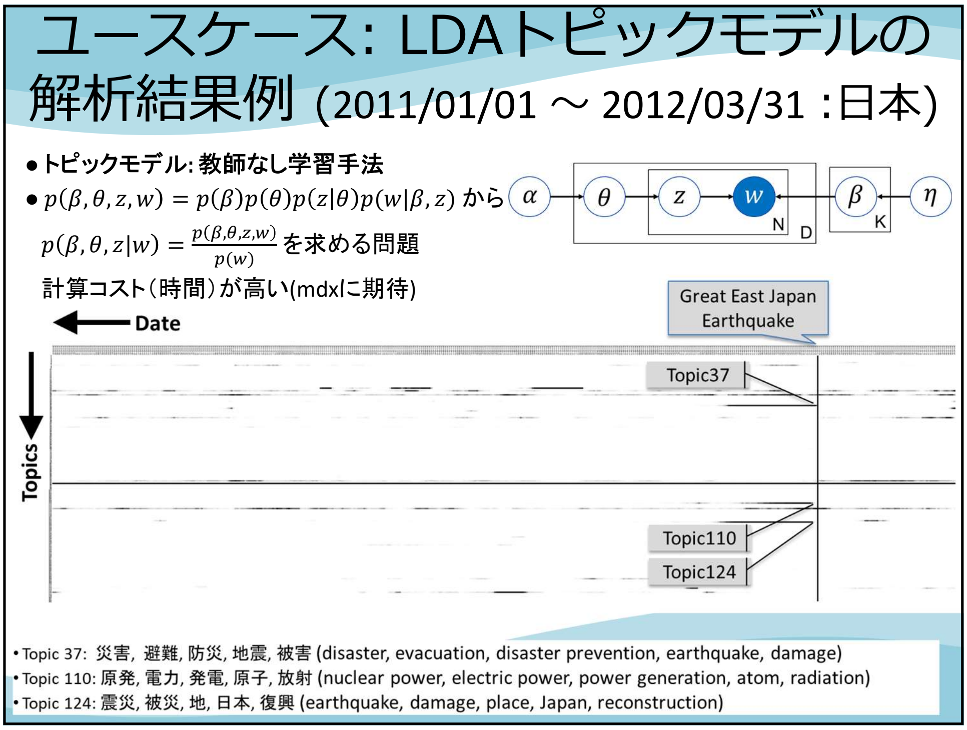
ユースケース:Webニュースデータから地域変動を検出する計量的手法の開発



ユースケース:データサンプル

- 収集しているデータ
  - 毎日新聞: Web 収集および購入CD-ROM)
  - 朝日新聞: Web 収集
  - 読売新聞:Web 収集
  - AFP (English and Spanish news articles via AFP Forum)
- この発表で利用したデータ
  - 毎日新聞CD-ROM版
  - 全国版と46地方版
  - 2010-01-01 ～ 2019-12-31

	Articles	Sentences	Words	Characters
National	607,671	7,791,338	219,677,468	321,075,103
Local	1,338,053	14,032,449	401,913,172	587,034,329
Total	1,945,724	21,823,787	621,590,640	908,109,432



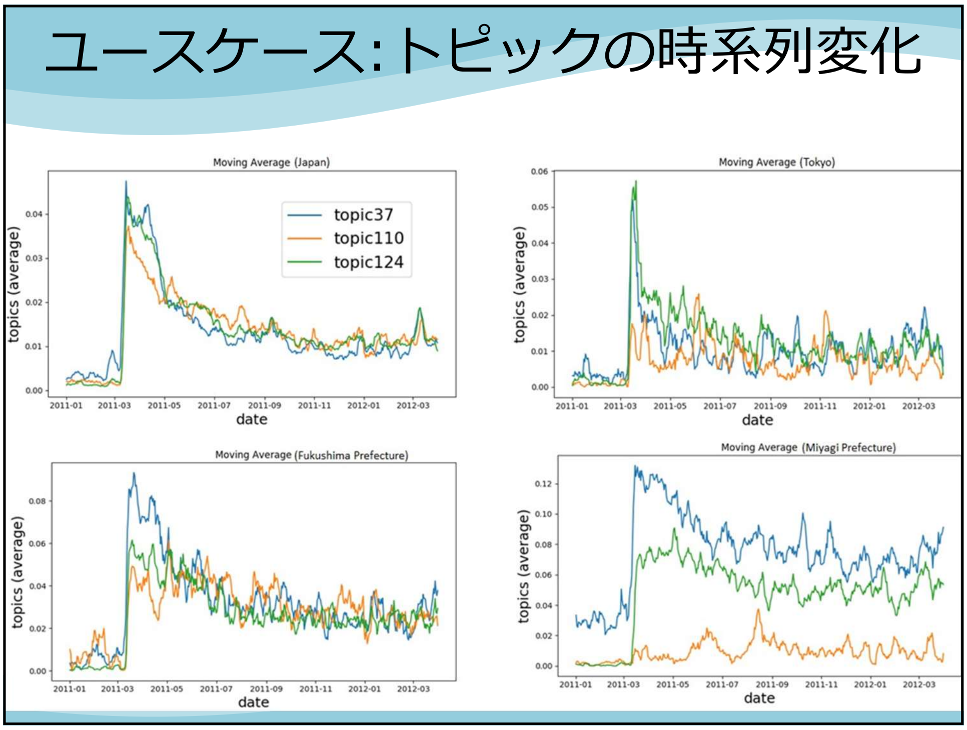
ユースケース: LDAトピックモデルの解析結果例 (2011/01/01 ～ 2012/03/31 :日本)

- トピックモデル: 教師なし学習手法
- $p(\beta, \theta, z, w) = p(\beta)p(\theta)p(z|\theta)p(w|\beta, z)$ から

$p(\beta, \theta, z|w) = \frac{p(\beta,\theta,z,w)}{p(w)}$ を求める問題

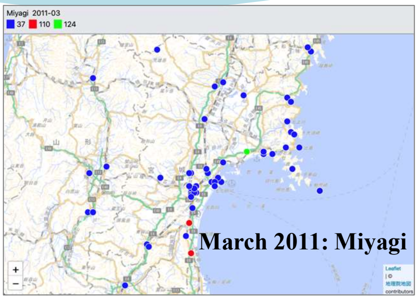
計算コスト（時間）が高い(mdxに期待)

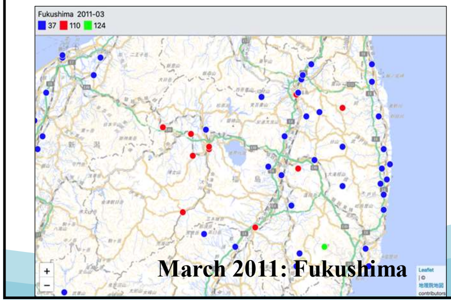
- Topic 37: 災害, 避難, 防災, 地震, 被害 (disaster, evacuation, disaster prevention, earthquake, damage)
- Topic 110: 原発, 電力, 発電, 原子, 放射 (nuclear power, electric power, power generation, atom, radiation)
- Topic 124: 震災, 被災, 地, 日本, 復興 (earthquake, damage, place, Japan, reconstruction)

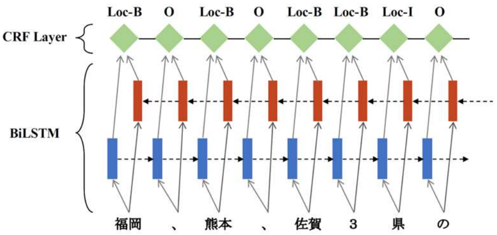


ユースケース:トピックの時系列変化

ユースケース:
トピックの空間分布の変遷







① BiLSTM-CRF:固有表現抽出。学習データから地名抽出法を学習する

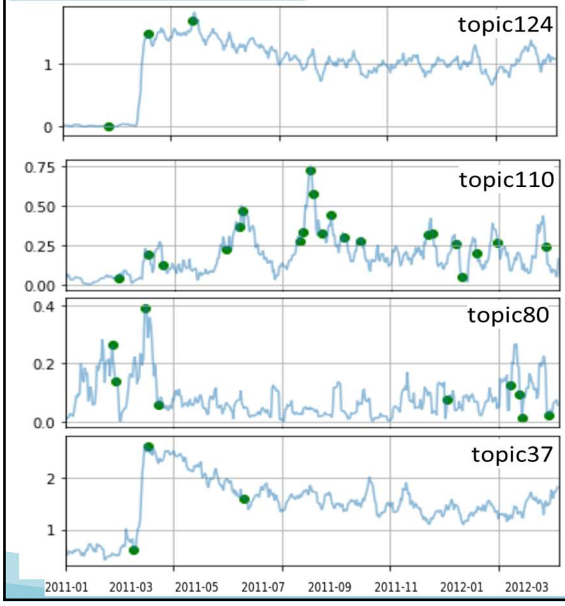
② 地名辞書を検索し、地名に対応する緯度・経度の候補集合を取得する

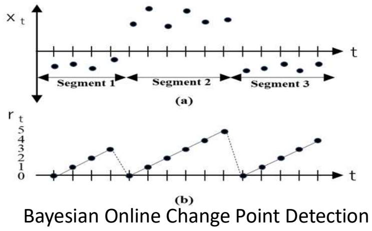
③ 単純な戦略を用いて緯度・経度を一つに絞り込む

⇒ このアノテーションデータを共有する（何処にどのように蓄積する？）

🟦 Topic 37
🟥 Topic 110
🟩 Topic 124

ユースケース:
トピック変化点抽出の試み (宮城県)





Bayesian Online Change Point Detection

- Topic 124: 震災, 被災, 地, 日本, 復興 (earthquake, damage, place, Japan, reconstruction)
- Topic 110: 原発, 電力, 発電, 原子, 放射 (nuclear power, electric power, power generation, atom, radiation)
- Topic 80: 病院/医療/患者/看護/救急 (hospital, medical, patients, emergency)
- Topic 37: 災害, 避難, 防災, 地震, 被害 (disaster, evacuation, disaster prevention, earthquake, damage)

ユースケース:トピック内容の変遷

Topic 80: 病院/医療/患者/看護/救急 (hospital, medical, patients, emergency)

2011-03-11	医療	病気	健康	薬	厚生	手術	市立	治療	診察	入院
2011-03-12	市立	精神	病院	専門	体制	医療	医師	医学	患者	厚生
2011-03-13	経過	機関	床	態勢	市立	病院	家族	精神	外来	患者
2011-03-14	態勢	病院	家族	市立	精神	外来	患者	診療	厚生	軽症
2011-03-15	病院	家族	市立	精神	外来	患者	診療	厚生	軽症	床
2011-03-16	家族	市立	精神	外来	患者	病院	診療	厚生	軽症	床
2011-03-17	市立	精神	外来	患者	病院	診療	厚生	軽症	床	ベッド
2011-03-18	市立	医療	診察	処方	外来	患者	病院	薬	診療	厚生
2011-03-19	体制	市立	患者	病院	機関	医療	外来	救急	診療	受付
2011-03-20	機関	医療	外来	患者	救急	病院	診療	受付	異常	家族
2011-03-21	搬送	看護	人工	家族	市立	健康	入院	受診	患者	病院
2011-03-22	医療	体調	医師	処方	命	患者	救急	薬	看護	保健
2011-03-23	ケア	床	異常	寝	医療	治療	命	患者	点滴	病棟
2011-03-24	市立	命	医療	患者	救急	病院	診療	受付	張り出	心身
2011-03-25	医療	患者	救急	病院	診療	受付	市立	張り出	心身	家族
2011-03-26	家族	市立	入浴	浴槽	機関	医療	病院	精神	健康	受付
2011-03-27	痛み	市立	家族	支え	搬送	機関	受付	菅原	入浴	命
2011-03-28	菅 原	入浴	支え	家族	命	市立	態勢	過ごせ	病院	体制
2011-03-29	家族	命	市立	態勢	過ごせ	病院	体制	ケア	健康	精神
2011-03-30	体制	ケア	健康	家族	精神	受付	態勢	市立	病院	診療
2011-03-31	市立	病院	診療	家族	命	精神	体制	医療	クリニック	公的
2011-04-01	家族	命	市立	精神	体制	医療	クリニック	公的	健康	菅原
2011-04-02	公的	家族	健康	精神	命	菅原	医師	医学	医者	入浴
2011-04-03	命	家族	菅原	医 師	医学	医者	入浴	寝	医療	病気
2011-04-04	医師	医学	医者	命	入浴	寝	医療	病気	救急	病院
2011-04-05	機関	健康	病院	診療	命	市立	家族	入浴	栄養	寝
2011-04-06	家族	入浴	栄養	寝	薬	班	命	過ごせ	助 成	医療
2011-04-07	家族	入浴	栄養	寝	薬	班	命	過ごせ	助成	医療
2011-04-08	命	過ごせ	助成	医療	治療	精神	受付	家族	市立	経過
2011-04-09	受付	家族	市立	経過	命	体制	栄養	医師	健康	看護
2011-04-10	家族	命	体制	栄養	医師	健康	看護	医療	厚生	床
2011-04-11	家族	入院	命	栄養	班	ケア	健康	心身	病院	痛み
2011-04-12	家族	心身	病院	痛み	栄養	在宅	乏し	命	受付	入院
2011-04-13	家族	入院	病院	菅原	外来	診療	専門	命	市立	健康
2011-04-14	家族	市立	命	健康	入浴	機関	精神	経過	専門	支え
2011-04-15	家族	市立	命	健康	入浴	機関	精神	経過	専門	支え
2011-04-16	機関	寝	菅原	家族	ケア	手術	市 立	支え	治療	入院
2011-04-17	市立	家族	床	小児	病院	乏し	態勢	機関	専門	院長
2011-04-18	病院	院長	搬送	手術	健康	看護	水分	専門	家族	ケア

Medical, Emergency, Cure

Medical Phase

～　Mid March

Medical, Emergency, Cure

Health, Care, Life, Support

Medical → Health Phase

Topic 50:temporary house, toilet etc.

Mid March ～ late March

Family

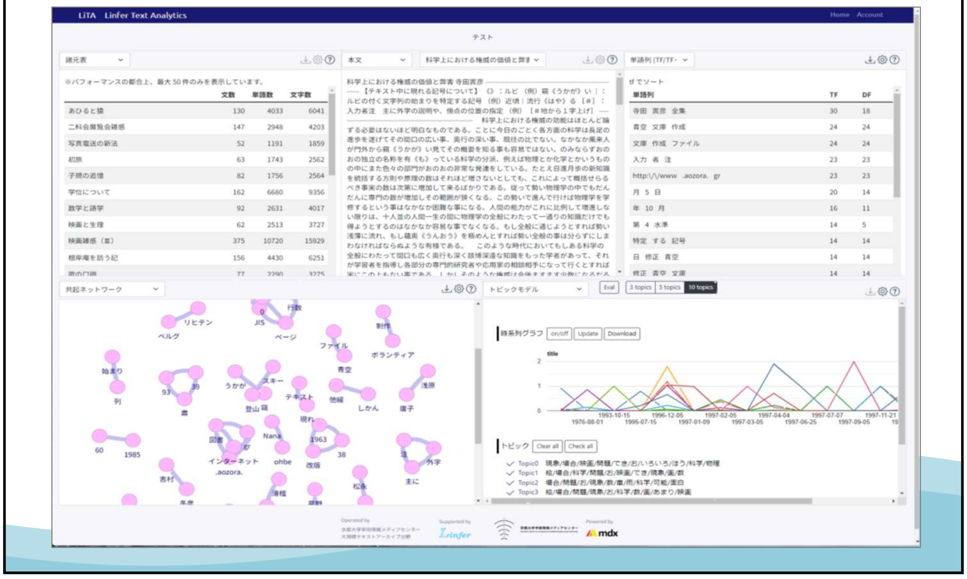
Health, Care, Life, Support

Health Phase

Topic 50:temporary house, recovery etc.

April ～

ツールボックスの開発（mdx）



LiTA　Linfer Text Analytics

※パフォーマンスの都合上、最大 50 件のみを表示しています。

	文数	単語数	文字数
あひると猿	130	4033	6041
二科会展覧会雑感	147	2948	4203
写真電送の新法	52	1191	1859
初旅	63	1743	2562
子規の追憶	82	1756	2564
学位について	162	6680	9356
数学と語学	92	2631	4017
映画と生理	62	2513	3727
映画雑感（Ⅲ）	375	10720	15929
根岸庵を訪う記	156	4430	6251
[illegible]	77	2290	3275

科学上における権威の価値と弊害 寺田寅彦 ―――――――――――――――――――――――― 【テキスト中に現れる記号について】 《》：ルビ （例）窺《うかが》い ｜：ルビの付く文字列の始まりを特定する記号 （例）近頃｜流行《はや》る ［＃］：入力者注　主に外字の説明や、傍点の位置の指定 （例）［＃地から１字上げ］ ―――――――――――――――――――――――― 科学上における権威の効能はほとんど論ずる必要はないほど明白なものである。ことに今日のごとく各方面の科学は長足の進歩を遂げてその間口の広い事、奥行の深い事、既往の比でない。なかなか風来人が門外から窺《うかが》い見てその概要を知る事も容易ではない。のみならずおのおの独立の名称を有《も》っている科学の分派、例えば物理とか化学とかいうものの中にまた色々の部門がおのおの非常な発達をしている。たとえ日進月歩の新知識を統括する方則や原理の数はそれほど増さないとしても、これによって概括せらるべき事実の数は次第に増加して来るばかりである。従って勢い物理学の中でもだんだんに専門の数が増加しその範囲が狭くなる。この勢いで進んで行けば物理学を学修するという事はなかなか困難な事になる。人間の能力がこれに比例して増進しない限りは、十人並の人間一生の間に物理学の全般にわたって一通りの知識だけでも得ようとするのはなかなか容易な事でなくなる。もし全般に通じようとすれば勢い浅薄に流れ、もし蘊奥《うんおう》を極めんとすれば勢い全般の事は分らずにしまわなければならぬような有様である。　このような時代においてもしある科学の全般にわたって間口も広く奥行も深く該博深遠な知識をもった学者があって、それが学習者を指導し各部分の専門的研究者や応用家の相談相手になって行くとすれば実にこの上もない事である。しかしそのような権威は

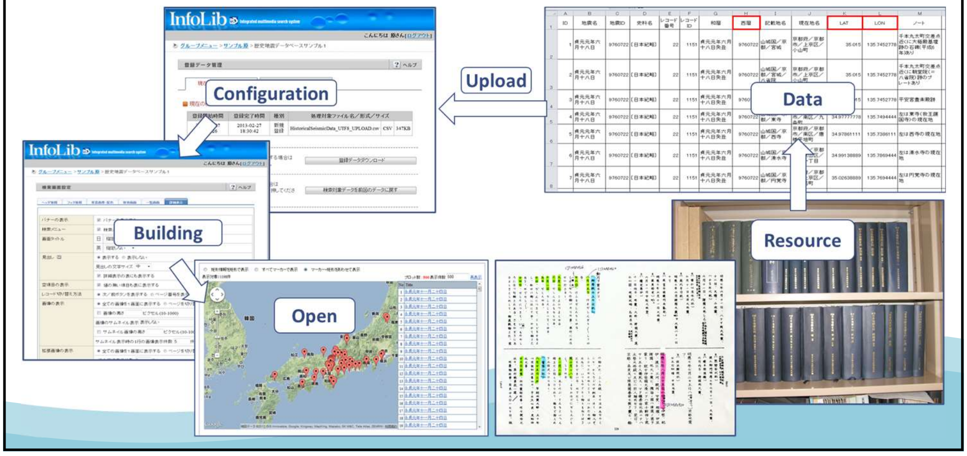
tf でソート

単語列	TF	DF
寺田　寅彦　全集	30	18
青空　文庫　作成	24	24
文庫　作成　ファイル	24	24
入力　者　注	23	23
http:\/\/www　.aozora.　gr	23	23
月　5　日	20	14
年　10　月	16	11
第　4　水準	14	5
特定　する　記号	14	14
日　修正　青空	14	14
修正　青空　文庫	14	14

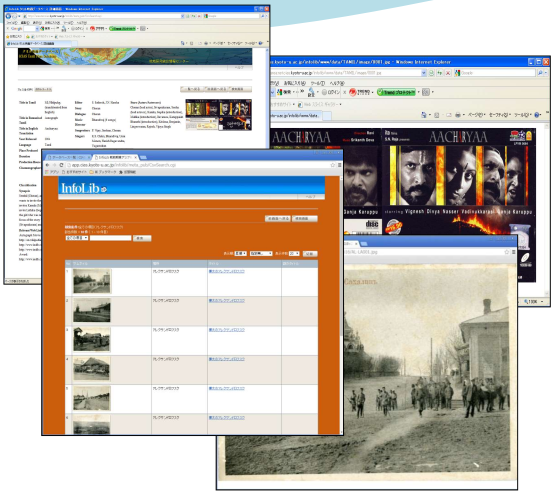
✓ Topic0　現象/場合/映画/問題/でき/お/いろいろ/ほう/科学/物理
✓ Topic1　絵/場合/科学/問題/お/映画/でき/現象/画/数
✓ Topic2　場合/問題/お/現象/数/車/雨/科学/可能/面白
✓ Topic3　絵/場合/問題/現象/お/科学/数/画/あまり/映画

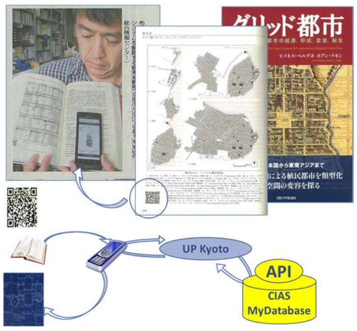
データ保存と共有:MyDatabase

- 多くは書誌的データベース（例：画像データと検索用メタデータ）だが、メタデータは研究者・資料毎に異なるので図書館システムには馴染まない
- 利用者のデータリテラシーはそれほど高くない
- 簡単に作れて変更も容易なデータベースシステムが必要



データ保存と共有:MyDatabaseの例





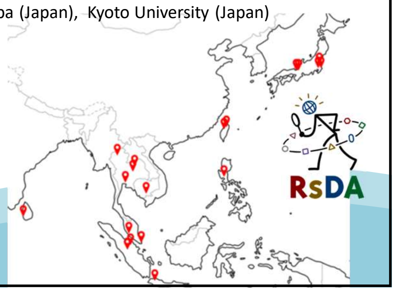
- データベースの作成には時間と手間がかかるので数が増えない
- データ公開なら比較的容易であり、最近ではデータ公開が要求される
- 論文との紐付けが無いので図書館のデータレポジトリには馴染まない
- とりあえずMyDatabaseの機能拡張で対応する

データ保存と共有:国際共同研究

Research small Data Alliance in east and southeast asia (RsDA)

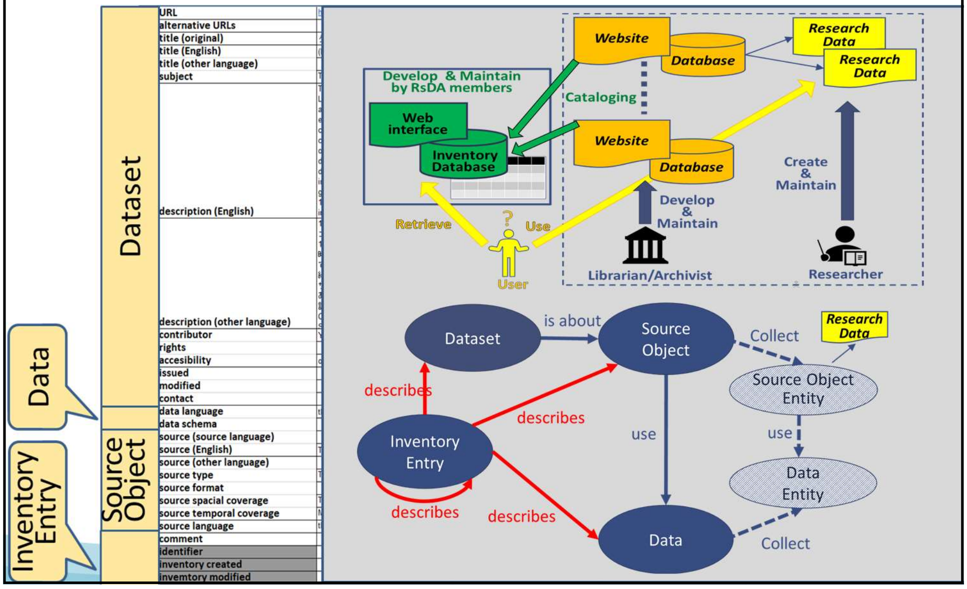
1. ASEANの主要学術記憶機関における**学術デジタル資産**に関する現状の理解と共有
2. 研究データの共有と保存にどのように対処すべきか、利用可能な技術は何か、誰がこの問題に貢献すべきか等について、参加者が意見やアイデアを交換し共有する機会を提供。
3. コラボレーションのためのネットワークを構築するための実践的なアプローチについて議論

Member Institutes: Academia Sinica (Taiwan), Chiang Mai University (Thailand), Chulalongkorn University (Thailand), International Research Center for Japanese Studies (Japan), Khon Kaen University (Thailand), National Museum of Japanese History (Japan), National Taiwan Normal University (Taiwan), National University of Singapore (Singapore), Royal University of Phnom Penh (Cambodia), Suranaree University of Technology (Thailand), Universitas Indonesia (Indonesia), Universitas Islam Riau (Indonesia), Universitas Lancang Kuning (Indonesia), Universiti Teknologi MARA (Malaysia), University of Colombo (Sri Lanka), University of the Philippines Diliman (Philippines), University of Tsukuba (Japan), Kyoto University (Japan)



データ保存と共有:メタデータ開発

（MyDatabaseの機能拡張⇒ GakuNin RDMへ）



URL
alternative URLs
title (original)
title (English)
title (other language)
subject
description (English)
description (other language)
contributor
rights
accesibility
issued
modified
contact
data language
data schema
source (source language)
source (English)
source (other language)
source type
source format
source spacial coverage
source temporal coverage
source language
comment
identifier
inventory created
invemtory modified

研究データ管理

1. どこにデータを蓄積したら良いのか（システムの問題）
   - 小規模の人文学系研究所（センター等）では恒久的な人員や機材の維持が困難
   - MyDatabaseは十数年に渡ってデータサービスを継続してきたが既に耐用年数を超えている
   - MyDatabaseでは収容できない大規模データの構築が進みつつある
   - 大学の研究データ管理システムあるいはGakuNin RDMへの期待
2. そもそもデータ管理はどのように行うのか（運用の問題）
   - 実はよく分かっていない（自己流：どのデータをどのように残す？）
   - GakuNin RDM利用するとしても、前提となる知識やスキルが必要ではないか（データとメタデータだけでデータの再利用は可能なのか？）

   ⇒ 必要なスキルや評価の標準化が必要（例えば、京都大学学際融合教育研究推進センターのアカデミックデータ・イノベーションユニットが開発したBasic Rubric）
   - ユースケースを通じてRubricの利用と評価を試みる

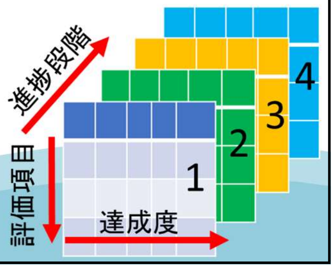
研究データ管理:Rubricの役割と作成

Rubricの役割

- 研究のための準備：Data Management Plans 提出等
- 研究進捗状況のチェック（データ取得・保管状況、処理・解析、文書化等）
- オープンサイエンス(データ)時代への対応

Rubricの役割

- 目標: 研究のライフサイクルの各段階でRDMに関する知識およびスキルの学習あるいは自己評価ができる仕組みの作成
- 分野: 様々な分野向けのルーブリックを作成。それらを基に、できる限り一般化する⇒Basic rubric
- 対象: 学生・若手研究者
- 特徴: 進捗段階に対応した4つのルーブリックに分割
  1. データに関する計画
  2. データの処理と整理
  3. データの解析
  4. データの共有または公開



ルーブリックの各項目毎に4段階の達成度を設定

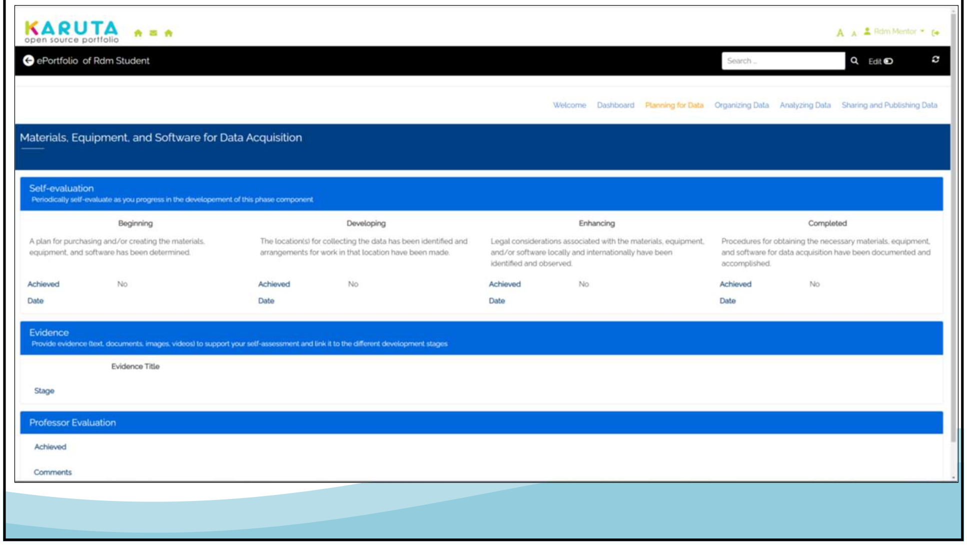
研究データ管理:Basic Rubricの例
(Planning for Data)

達成度

評価項目

PLANNING FOR DATA	Beginning	Developing	Enhancing	Completed
Data to Be Obtained	The data to be produced and the materials required to produce it have been determined.	The categories of data to be acquired and the relationship of the categories to specific research purposes have been identified.	The resolution, quality, and minimum amount of data, as well as necessary subjects for acquiring the data, have been identified.	The researcher is confident of the categories, resolution, quality, and minimum amount of data, and that needed subjects can be reliably acquired.
Methodology for Obtaining Data	The methodology required to obtain the data has been determined.	The instruments, data transfer plan, and storage device required to obtain and store the data have been identified.	Preparation or implementation of the instruments, data transfer arrangements, and storage device has begun.	Preparation or implementation of the instruments, data transfer arrangements, and storage device is complete.
Materials, Equipment, and Software for Data Acquisition	A plan for purchasing and/or creating the materials, equipment, and software has been determined.	The location(s) for collecting the data has been identified and arrangements for work in that location have been made.	Legal considerations associated with the materials, equipment, and/or software locally and internationally have been identified and observed.	Procedures for obtaining the necessary materials, equipment, and software for data acquisition have been documented and accomplished.
Team Members	Team members responsible for data acquisition and dissemination have been identified.	Appropriate roles have been assigned according to the skills and interests of individual team members.	A suitable research environment has been prepared to encourage information exchange, consultation, and cooperation among team members.	Team members fully understand the division of roles, as well as their own personal role, and are well prepared to conduct and disseminate the research.
Funding for Data Acquisition	Funding required for data acquisition has been estimated and available sources of funding identified.	Funding applications include team member expenses, purchase of equipment and materials, software for data acquisition, and legal expenses.	Requests for funding have been submitted to relevant funding sources.	Sufficient funding for all expenses related to data acquisition has been acquired.
Data Re-Use	Changes that may be required in ways that the research data is managed for re-use have been considered.	The availability of the data for re-use by other researchers has been determined.	Decisions have been made regarding the exclusive right of team members to have use of the data during an institutionally determined embargo period.	Plans for potential re-use of the data, including necessary changes to its management, are complete and accepted by all team members.
Open Versus Closed Data	Academic journals for potential publication have been identified and their conditions for data disclosure are understood. Sharing data in public repositories has been considered.	Decisions have been made as to which data (and materials) can be published and shared and which cannot. Relevant international and local legalities and proprietary issues have been considered.	The choice of licenses and repositories for publishing and sharing the data, as well as potential users of the data (and materials), have been considered.	The choices for publishing and sharing the data (and materials) have been documented and agreed to by all team members.
Planning for Publishing and Sharing Data	The requirements of funding agencies and research institutes being considered to manage the disclosure of the data (and materials) are understood.	Funding for disclosure of the data (and materials) has been estimated. It has been decided if the researcher's institution will be responsible for data disclosure or if the responsibility will be outsourced.	If needed, suitable repositories for outsourcing the publishing and sharing process have been identified and responsibilities for ongoing expenses have been determined.	A Research Data Management (RDM) plan for publishing and sharing the data (and materials) has been created, documented, and agreed upon by all team members and by the researcher's institution.

研究データ管理:ePortfolio



ePortfolio of Rdm Student

Materials, Equipment, and Software for Data Acquisition

Self-evaluation
Periodically self-evaluate as you progress in the developement of this phase component

Beginning	Developing	Enhancing	Completed
A plan for purchasing and/or creating the materials, equipment, and software has been determined.	The location(s) for collecting the data has been identified and arrangements for work in that location have been made.	Legal considerations associated with the materials, equipment, and/or software locally and internationally have been identified and observed.	Procedures for obtaining the necessary materials, equipment, and software for data acquisition have been documented and accomplished.
Achieved No	Achieved No	Achieved No	Achieved No
Date	Date	Date	Date

Evidence
Provide evidence (text, documents, images, videos) to support your self-assessment and link it to the different development stages

Evidence Title

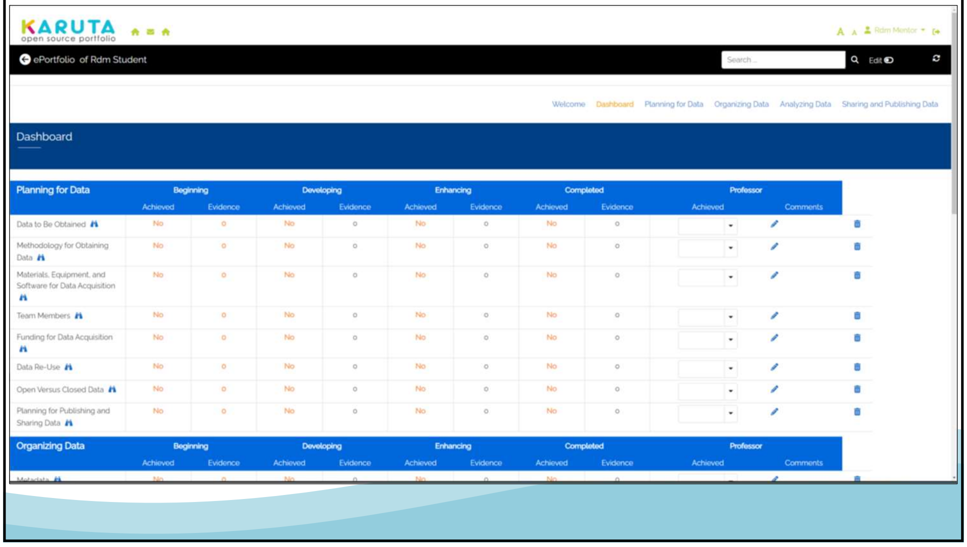
Stage

Professor Evaluation

Achieved

Comments

研究データ管理: ePortfolio



KARUTA open source portfolio

Rdm Mentor

ePortfolio of Rdm Student

Search ... Edit

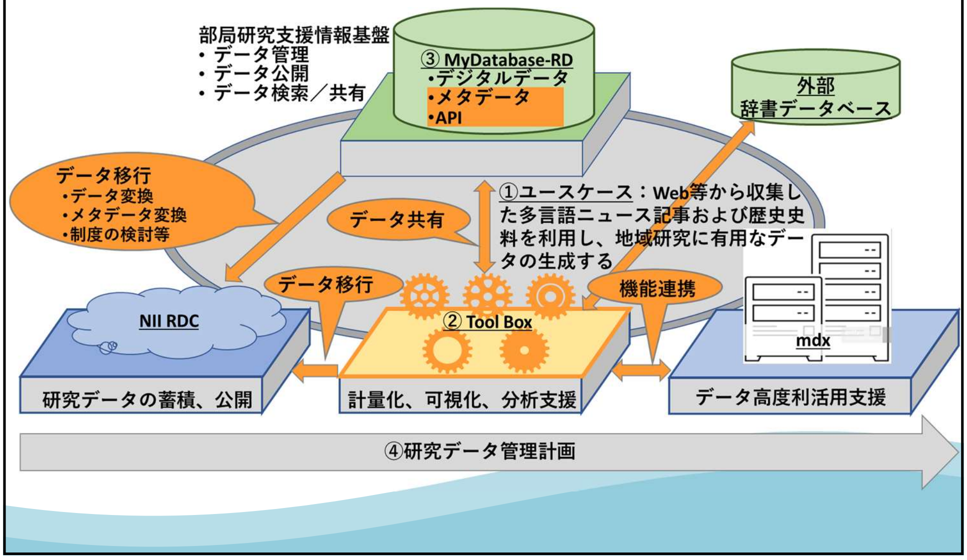
Welcome Dashboard Planning for Data Organizing Data Analyzing Data Sharing and Publishing Data

Dashboard

Planning for Data	Beginning		Developing		Enhancing		Completed		Professor		
	Achieved	Evidence	Achieved	Evidence	Achieved	Evidence	Achieved	Evidence	Achieved	Comments	
Data to Be Obtained	No	0	No	0	No	0	No	0	▾	✎	🗑
Methodology for Obtaining Data	No	0	No	0	No	0	No	0	▾	✎	🗑
Materials, Equipment, and Software for Data Acquisition	No	0	No	0	No	0	No	0	▾	✎	🗑
Team Members	No	0	No	0	No	0	No	0	▾	✎	🗑
Funding for Data Acquisition	No	0	No	0	No	0	No	0	▾	✎	🗑
Data Re-Use	No	0	No	0	No	0	No	0	▾	✎	🗑
Open Versus Closed Data	No	0	No	0	No	0	No	0	▾	✎	🗑
Planning for Publishing and Sharing Data	No	0	No	0	No	0	No	0	▾	✎	🗑

Organizing Data	Beginning		Developing		Enhancing		Completed		Professor		
	Achieved	Evidence	Achieved	Evidence	Achieved	Evidence	Achieved	Evidence	Achieved	Comments	
Metadata	No	0	No	0	No	0	No	0	▾	✎	🗑

統合はこれから・・・



部局研究支援情報基盤
- データ管理
- データ公開
- データ検索／共有

③ MyDatabase-RD
- デジタルデータ
- メタデータ
- API

外部
辞書データベース

データ移行
- データ変換
- メタデータ変換
- 制度の検討等

データ共有

①ユースケース：Web等から収集した多言語ニュース記事および歴史史料を利用し、地域研究に有用なデータの生成する

データ移行

機能連携

NII RDC

② Tool Box

mdx

研究データの蓄積、公開

計量化、可視化、分析支援

データ高度利活用支援

④研究データ管理計画