資料２－１
科学技術・学術審議会　学術分科会
人文学・社会科学特別委員会（第７回）
令和３年６月２８日

デジタル・ヒューマニティーズ（DH）の概要と人文学DX

一般財団法人人文情報学研究所　主席研究員

永崎研宣

本件に関わる自己紹介

- 日本学術振興会　人文学・社会科学ﾃﾞｰﾀｲﾝﾌﾗｽﾄﾗｸﾁｬｰ構築推進ｾﾝﾀｰ 研究員（2021年度より）
- 東京大学大学院人文社会系研究科次世代人文学開発ｾﾝﾀｰ人文情報学部門　客員研究員・非常勤講師（2012年より）
  - その他、DHの授業担当（関西大・同志社大・立教大・筑波大・大阪大・広島大）
- 京都大学人文科学研究所共同研究班「人文学にとってのWebを再探する」班長
- 国立国会図書館研究員（委嘱）（2014年より）
- 日本デジタル・ヒューマニティーズ学会議長（2019年より）
- Alliance of Digital Humanities Organizations運営委員（国際DH学会連合）（2019年より）
- 情報処理学会人文科学とコンピュータ研究会運営委員
- Text Encoding Initiative Consortium 理事 （2017-2018）
- 情報規格調査会SC2委員会委員（2013年より）
- ISO/IEC JTC1/SC2 リエゾンメンバー（SATからの代表として）（2017年より）
- SAT大蔵経テキストデータベース研究会技術担当（2005年より）

DHの概要について

研究全体の概要

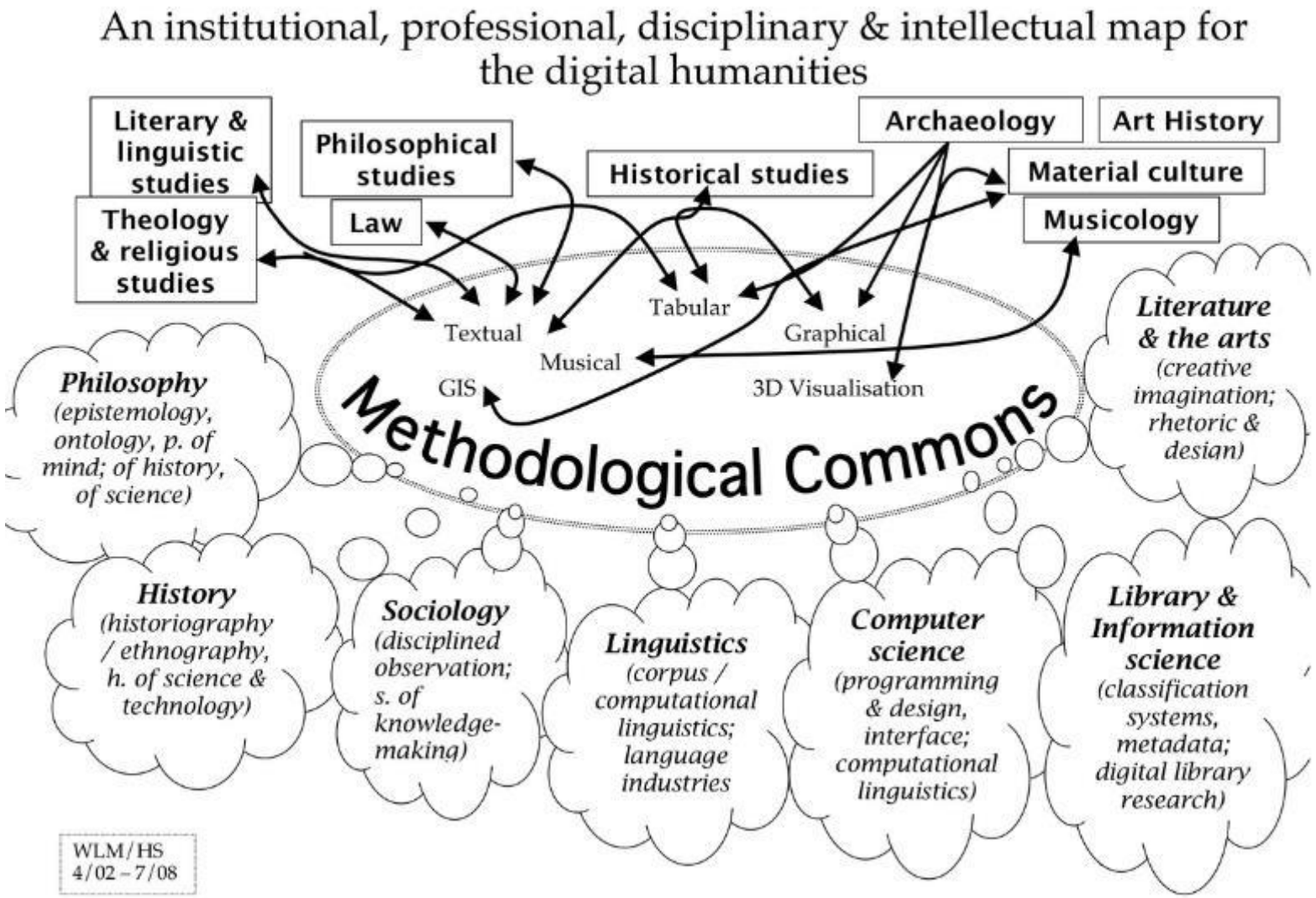
学会の動向

デジタル・ヒューマニティーズ（DH）とは

- 人文学の何らかの側面にデジタル技術を適用・応用する研究。
  - 1940年代に開始。
  - 1980年頃の隆盛（パソコンの登場による）
  - 2005年の国際DH学会連合設立。
    - 2006年、パリ・ソルボンヌ大学を皮切りに「デジタル・ヒューマニティーズ」を冠する国際学会が開催され、以後、毎年開催されてきた。
- 2006年、欧州ESFRIロードマップにより欧州DHインフラDARIAH開始
  - 2014年、 DARIAHは欧州15ヶ国で正式に設立。現在は26ヶ国が参加・協力。
- 2008年、米国人文学基金（NEH）がOffice of DHを設置
  - 2006年にはNEHが関係者を集めたフォーラムを開催
  - DH専門の研究助成を開始。AHRC（英）、DFG（独）との連携ファンド。
  - 8ヶ国連携のDigging into Data Challengeファンドにも参加。

第一回DH国際会議（2006年、パリ・ソルボンヌ大学）

DHの場を形成する理念的背景
＝タコツボ化を越える建設的な再構築の場



人文学の様々な分野・様々な手法をデジタル技術の応用を介して横断的に議論し共有するための場の形成

横断的な議論を通じて相互の方法論を自省し深化させる場にもなり得る

成果自体を横断的に産み出す場にもなり得る

参照：
https://digitalnagasaki.hatenablog.com/entry/2020/12/20/182659

DHの場を支える技術的背景

- 人文学のための国際規格の策定・改訂と運用
  - それぞれの分野の専門家コミュニティが取組み
    - テキスト資料
      - TEI （Text Encoding Initiative）協会による1987年からのTEIガイドライン策定
      - 欧米圏で進む、TEIガイドラインに準拠した人文学向けテキストデータの構築と共有
      - データ駆動型研究においては必須の構造化された応用データの基盤
      - 科研費基盤(S)事業による東アジア・日本語分科会の設立と2分科会提案によるルビの導入（2021年）
    - デジタル画像
      - IIIF (International Image Interoperability Framework)協会による2011年頃からの仕様策定
      - 欧米の文化機関に所属するWebエンジニアを中心としたコミュニティが推進
      - 国内外の文化機関で普及が進みつつある
    - 博物館・美術館資料の目録データ
      - 国際博物館会議（ICOM）が目録標準モデル CIDOC-CRMを策定
      - 2020年、バージョン7.0が公開
    - 記録史料の目録データ
      - 国際公文書館会議（ICA）が国際標準記録資料記述一般原則 ISAD(G)を策定
    - 文字コード
      - Unicode、ISO/IEC 10646における多様な文字への対応

研究インフラ側からの支援として

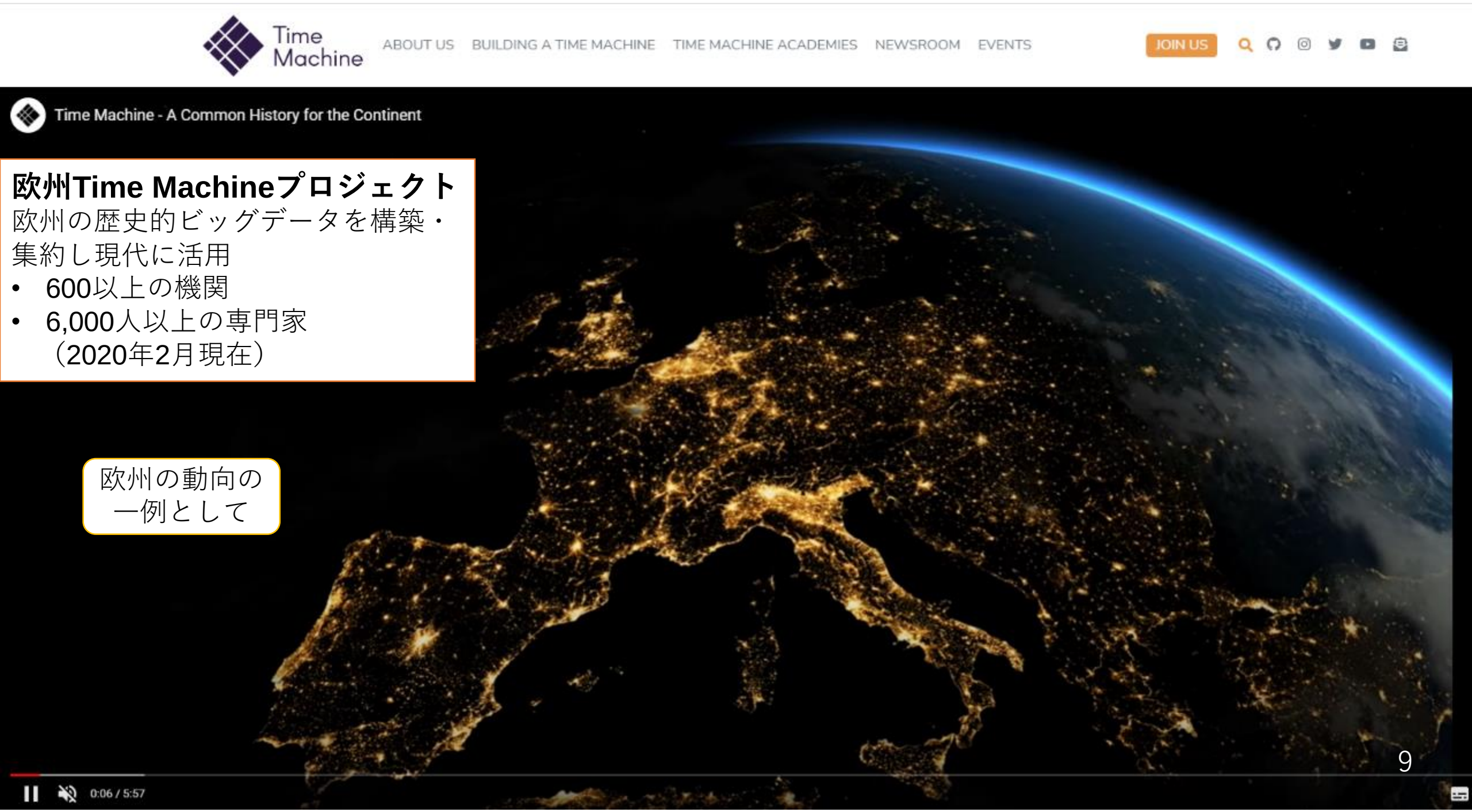
- 欧州における Social Sciences & Humanities Open Marketplace

人文・社会科学のデジタル研究・教育に関する総合ポータル

- データセット
- ツール＆サービス
- 教材
- 刊行物
- ワークフロー

JDCatの拡大版と言える
⇒連携の可能性も

cessda CLARIN DARIAH-EU European Social Survey SHARE



欧州Time Machineプロジェクト

欧州の歴史的ビッグデータを構築・集約し現代に活用

- 600以上の機関
- 6,000人以上の専門家
  （2020年2月現在）

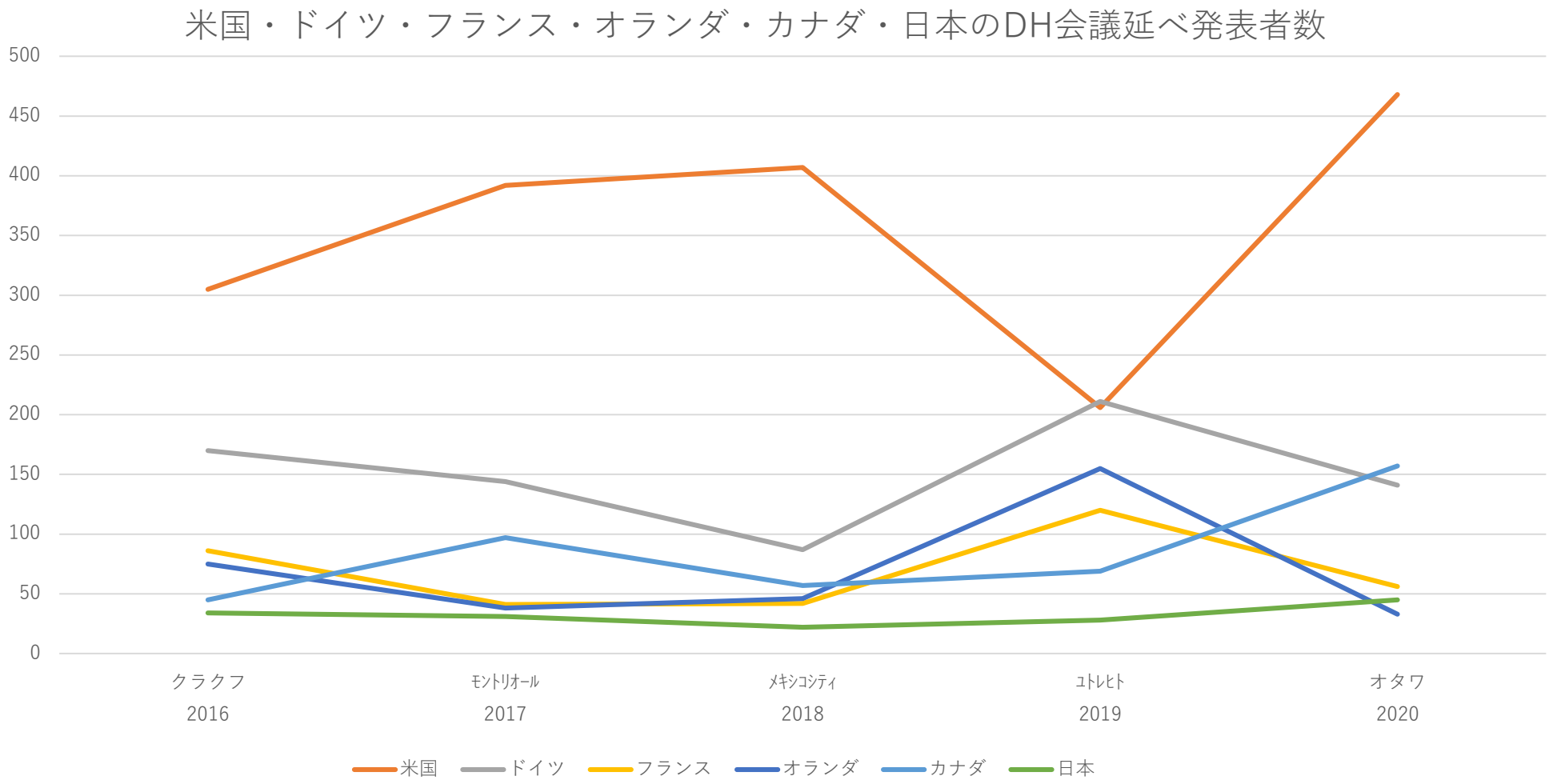
欧州の動向の一例として

国内外のDHコミュニティ

発表者数の推移から

国際DH連合学術大会における延べ発表者数

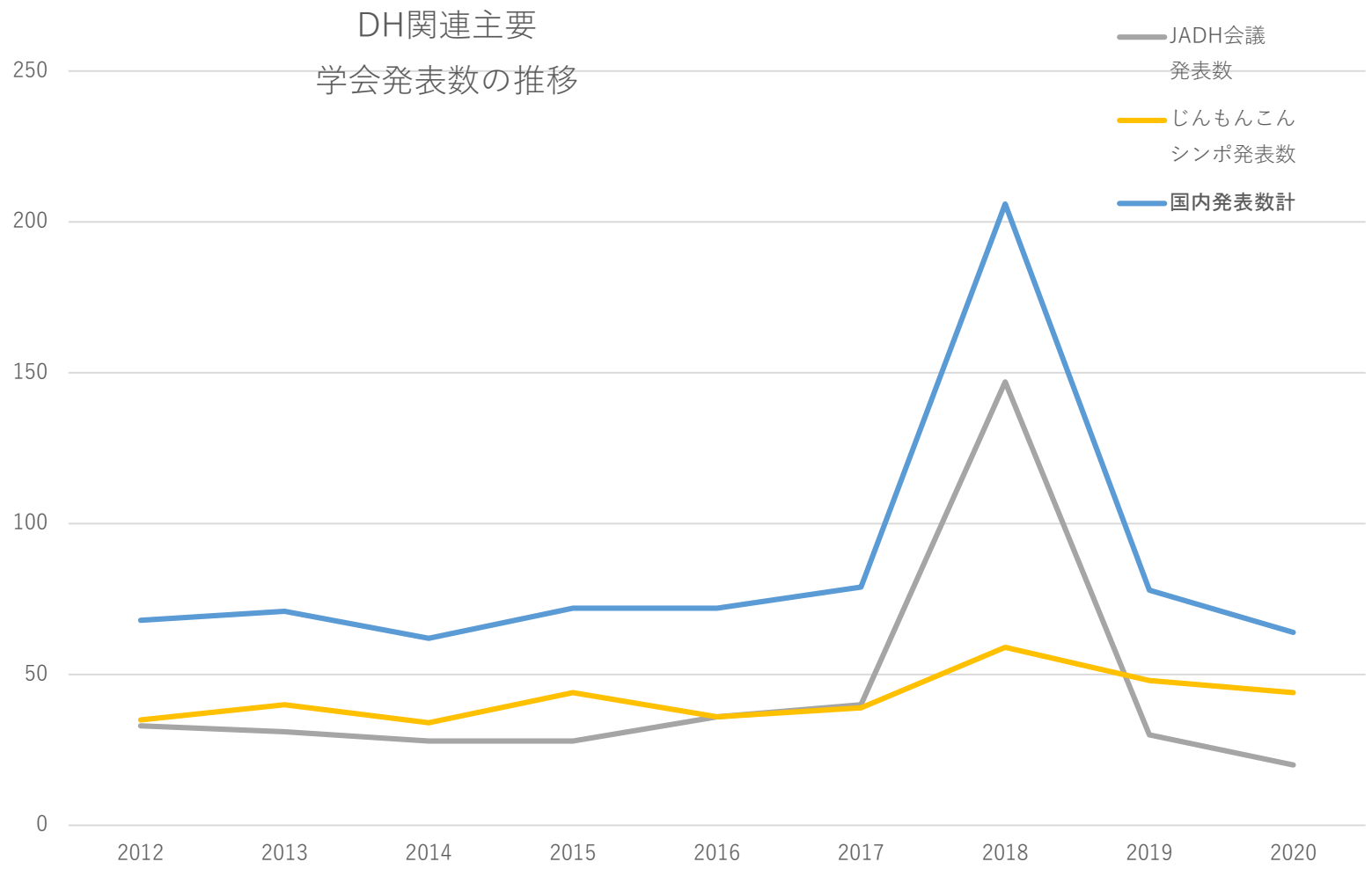
2016-2020における国際的な研究発表の動向として



国際DH連合学術大会における共同発表者の国際ネットワーク

2016-2020における国際的な研究発表の動向として

国内のDH関連学会の発表件数



※2018年JADH会議は、TEI (Text Encoding Initaitive) 年次大会と共催）

※2018年のTEI年次大会が欧米以外の場（東京）で開催されたのは初めてのこと

東アジアの動向

- 台湾（数位人文）
  - 中央研究院・国立台湾大学を中心とした長い取組み
    - DH基盤データの整備・公開／DH研究プラットフォームの構築運用
    - DH教育カリキュラムへの取組み（政府の**教育部數位人文創新人才培育計畫**）
    - 国際会議の継続的な開催（PNC（Pacific Neighborhood Consortium, 1997年より）, DADH(Digital Archives and Digital Humanities Conference), 2009年より）
  - 台湾DH学会の設立とジャーナルの刊行
- 中国（数字人文、數碼人文）
  - DHセンターの設立
    - 武漢大学、上海師範大学、中国人民大学、北京大学
  - DH関連の主な活動
    - DHフォーラムの開催（北京大学（2015年より）
    - DHジャーナルの刊行（清華大学（2020年））
    - 資料デジタル化の推進
      - 浙江大学　CADAL
      - 山東大学　全球漢籍合璧工程調査目録編纂複製作業
      - 中国国家図書館など中国の図書館10館が古典籍のデジタル公開 (https://current.ndl.go.jp/node/43982)
- 中国学DHの国際ネットワーク
  - 中国圏外ではハーバード大学、ライデン大学、ダラム大学、京都大学が注目される
- 韓国
  - 政府レベルでのデジタルデータ構築の取組み
  - Humanities Research Institute による取組み
    - 2018年よりAI人文学国際会議を開催
  - 韓国DH学会設立に向けて準備中

データ駆動型研究の事例紹介

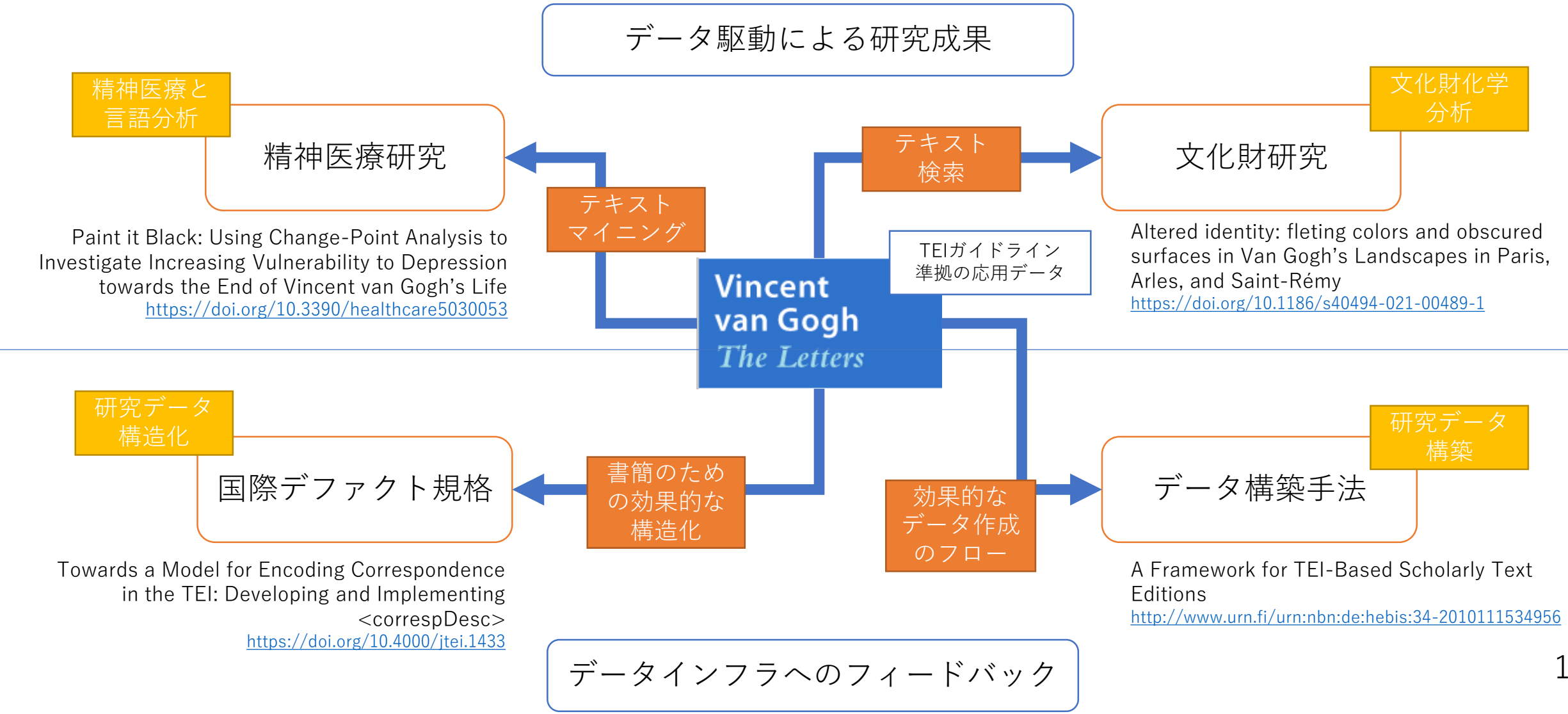
一つのデータセットから様々な研究が産み出される事例として

- ゴッホの手紙
- 和歌テキストデータベース

データ駆動型研究のプロセス自体が社会的課題の解決につながる事例として

- 「みんなで翻刻」

データ駆動型研究とデータインフラ活動へのフィードバック：Vincent van Gogh The Lettersにおける事例の一部



Paint it Black: Using Change-Point Analysis to Investigate Increasing Vulnerability to Depression towards the End of Vincent van Gogh's Life
https://doi.org/10.3390/healthcare5030053

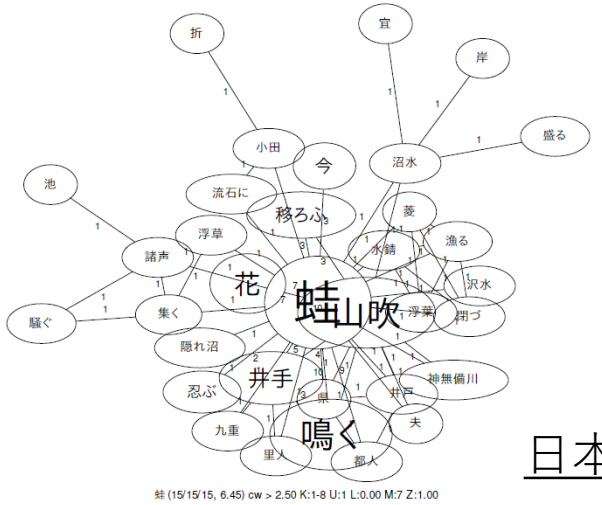
Altered identity: fleting colors and obscured surfaces in Van Gogh's Landscapes in Paris, Arles, and Saint-Rémy
https://doi.org/10.1186/s40494-021-00489-1

Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc>
https://doi.org/10.4000/jtei.1433

A Framework for TEI-Based Scholarly Text Editions
http://www.urn.fi/urn:nbn:de:hebis:34-2010111534956

和歌のテキストデータ

日本の伝統文化の再発見



和歌の精選された語彙における
看過されてきた意味の発見

山元 啓史「山吹」をめぐる和歌語彙の空間
http://id.nii.ac.jp/1001/00079399/

ジェンダー研究

日本の伝統文化における
女性の位置づけの定量的な分析

近藤みゆき『王朝和歌研究の方法』笠間書院

日本語資料のための構造化研究

永崎研宣, 乾 善彦他「万葉集伝本研究のためのデジタル基盤構築」
http://id.nii.ac.jp/1001/00209265/

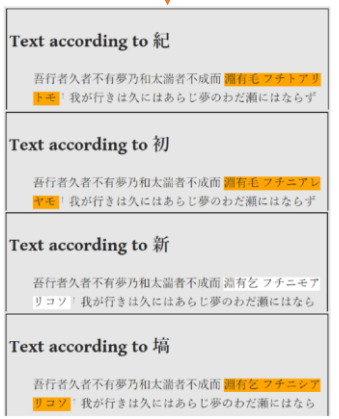
日本の古典を国際的な規格に準拠させる
ために必要な事項を探究

```
<text>
  <body style="writing-mode:vertical-rl">
    <ab/>
    <div n="廣瀬本万葉集第二巻">
      <pb corresp="#page2805960"
        facs="https://www.iiif.ku-orcas.kansai-u.ac.jp/iiif/2/210464810%2F0044.tif/full/full/0/default.jpg"
        n="四六丁表" source="#page2805960"/>
      <div corresp="#盤姫皇后">
        <div xml:id="旧国歌大観番号八五">
          <p xml:id="p00076">相聞</p>
          <p xml:id="p00077">難波高津宮御宇天皇代<note resp="#万葉集" type="割書" xml:id="note00001"
              >大鷦鷯天皇　諡曰仁徳天皇</note></p>
          <p style="text-indent:3em" xml:id="p00078">盤姫皇后思　天皇御作歌四首</p>
          <lg type="短歌" xml:id="manyo0085">
            <l xml:id="manyo0085_01"><lb/>君之行氣長成奴山多都祢迎加将行待尓可将待</l>
            <l corresp="#manyo0085_01" xml:id="manyokun0085_01"><lb/>キミカユキケナカクナリヌヤマタツネ
                <note place="right" resp="#定家" type="注" xml:id="note00002"
                >或本ムカヘニ</note>ムカヘカ<anchor type="noteEnd"/>ユカムマチニカマタ <note place="right"
                resp="#定家" subtype="異訓" type="合点" xml:id="note00003">ツムノ</note>モ<anchor
                type="noteEnd"/></l>
          </lg>
```

日本語文法の探究

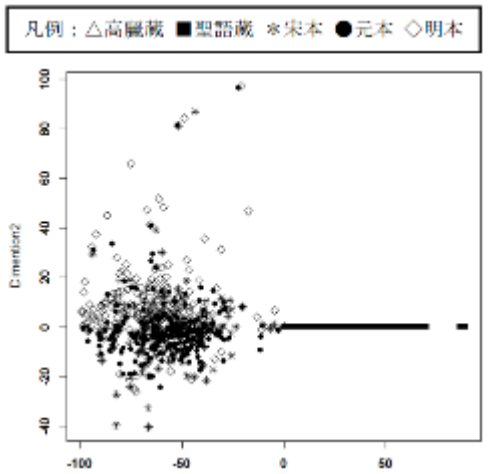
現代も用いられる
助詞の用法の歴史を明らかに

（大学院生による研究事例）



小池俊希『日本語歴史コーパス』へのTEI適用に基づく諸本比較――『万葉集』における「読添えのモ」を事例として――
http://id.nii.ac.jp/1001/00204772/

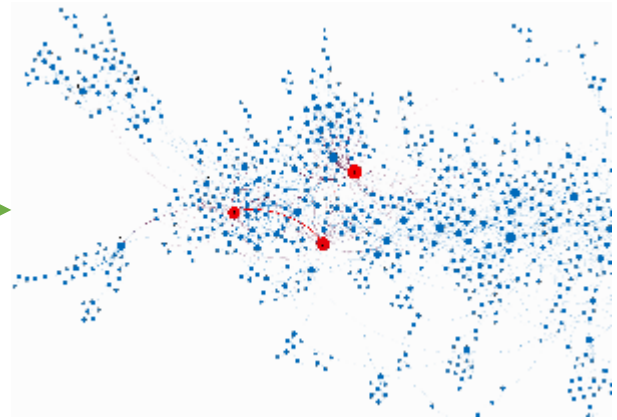
永崎研宣他「人文学資料としてのテクスト構造化の意義を再考する」
http://id.nii.ac.jp/1001/00096423/



8世紀～16世紀のテキスト伝承の分析

Marcus Bingenheimer, “On the Use of Historical Social Network Analysis in the Study of Chinese Buddhism: The Case of Dao’an, Huiyuan, and Kumārajīva”
https://doi.org/10.17928/jjadh.5.2_84

4～5世紀中国僧の社会ネットワーク分析



仏典テキストデータ

仏教対話AI「ブッダボット」－伝統知と人工知能の融合－
（京大こころの未来研究センター）
https://www.kyoto-u.ac.jp/ja/research-news/2021-03-26-3

東アジア・日本の人文学資料の国際標準化
http://www.l.u-tokyo.ac.jp/news/2021/13300.html

情報学との連携による社会的課題の解決の可能性へ

データ駆動型研究のプロセスがもたらす意義

雉尾弥—日本書紀にみる赤気に関する一考察
http://id.nii.ac.jp/1013/00005767/

古気象研究
（地球温暖化）

歴史のなかの地震・噴火: 過去がしめす未来
https://ci.nii.ac.jp/ncid/BC05967143

古地震研究
（防災・減災）

地域史
（地方創生）

ジェンダー問題
（男女共同参画）

日本文化の
再発見

データ駆動型研究へ

古文書
古記録
テキスト
データ

古典籍
テキスト
データ

機械可読性の高いデータ構築

データ駆動型研究の
プロセスを通じた社会貢献

＝データ作成期間中の
社会貢献

協働プラットフォーム

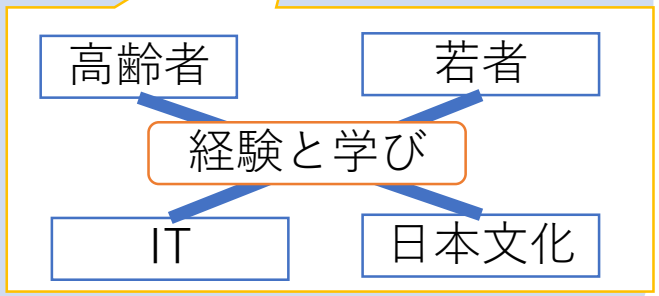
みんなで翻刻
https://honkoku.org/

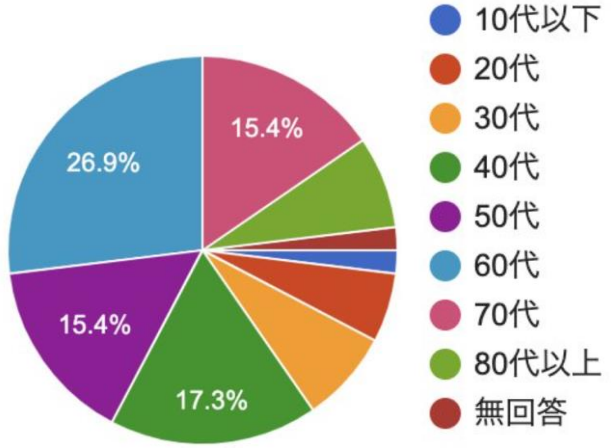
貢献

学び

少子高齢化社会に
おける共創の場

人力＋AI協働
の学びの場

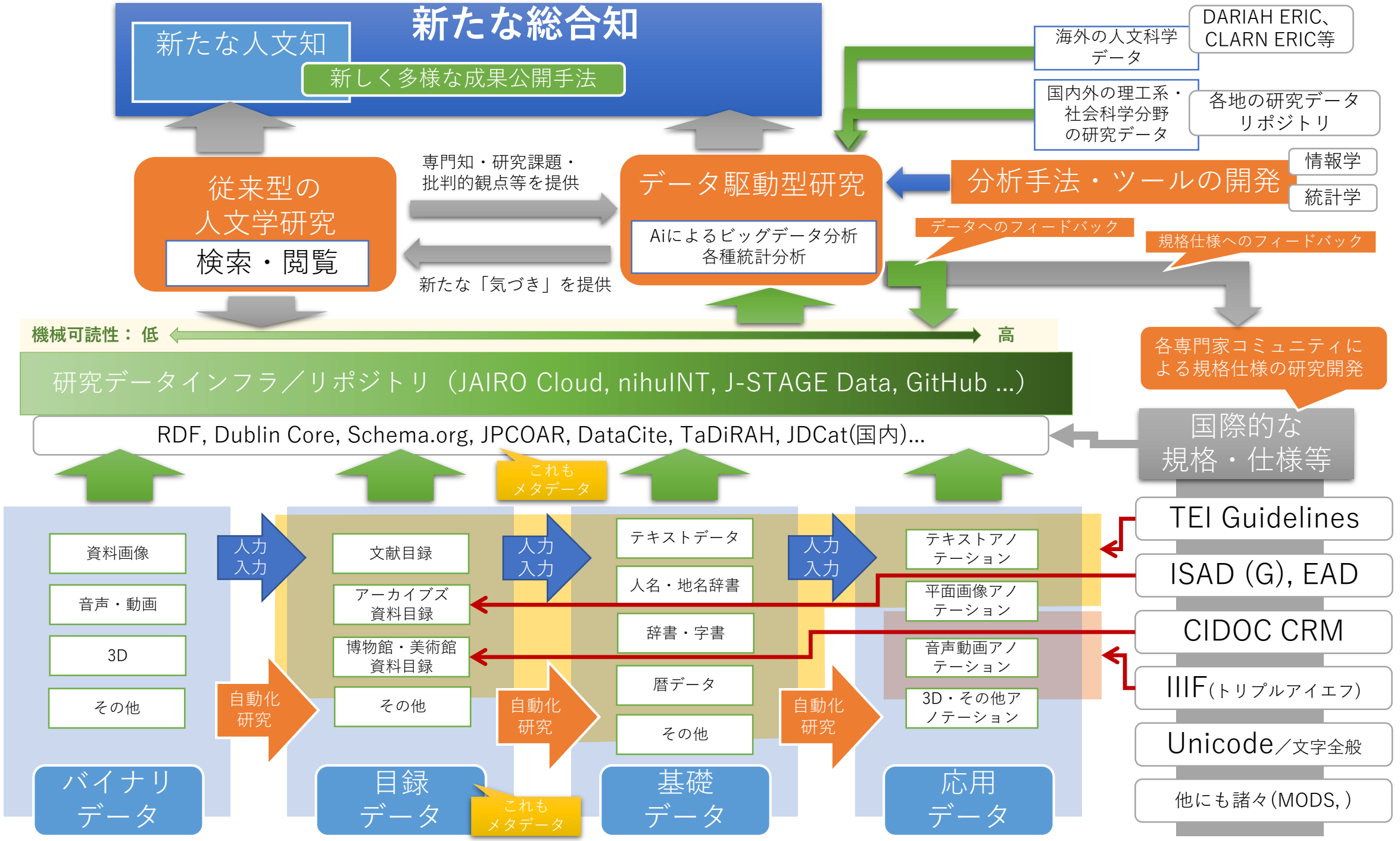




『みんなで翻刻』の運用成果と参加動向の報告
http://id.nii.ac.jp/1001/00208573/

必要となるインフラ・環境

既存の様々な取組みを踏まえた見取り図として


新たな総合知
新たな人文知
新しく多様な成果公開手法
海外の人文科学データ
DARIAH ERIC、CLARN ERIC等
国内外の理工系・社会科学分野の研究データ
各地の研究データリポジトリ
従来型の人文学研究
検索・閲覧
専門知・研究課題・批判的観点等を提供
新たな「気づき」を提供
データ駆動型研究
Aiによるビッグデータ分析
各種統計分析
分析手法・ツールの開発
情報学
統計学
データへのフィードバック
規格仕様へのフィードバック
機械可読性：低
高
各専門家コミュニティによる規格仕様の研究開発
研究データインフラ／リポジトリ（JAIRO Cloud, nihuINT, J-STAGE Data, GitHub ...）
RDF, Dublin Core, Schema.org, JPCOAR, DataCite, TaDiRAH, JDCat(国内)...
国際的な規格・仕様等
これもメタデータ
資料画像
音声・動画
3D
その他
人力入力
文献目録
アーカイブズ資料目録
博物館・美術館資料目録
その他
人力入力
テキストデータ
人名・地名辞書
辞書・字書
暦データ
その他
人力入力
テキストアノテーション
平面画像アノテーション
音声動画アノテーション
3D・その他アノテーション
自動化研究
自動化研究
自動化研究
TEI Guidelines
ISAD (G), EAD
CIDOC CRM
IIIF（トリプルアイエフ）
Unicode／文字全般
他にも諸々(MODS,)
バイナリデータ
目録データ
これもメタデータ
基礎データ
応用データ

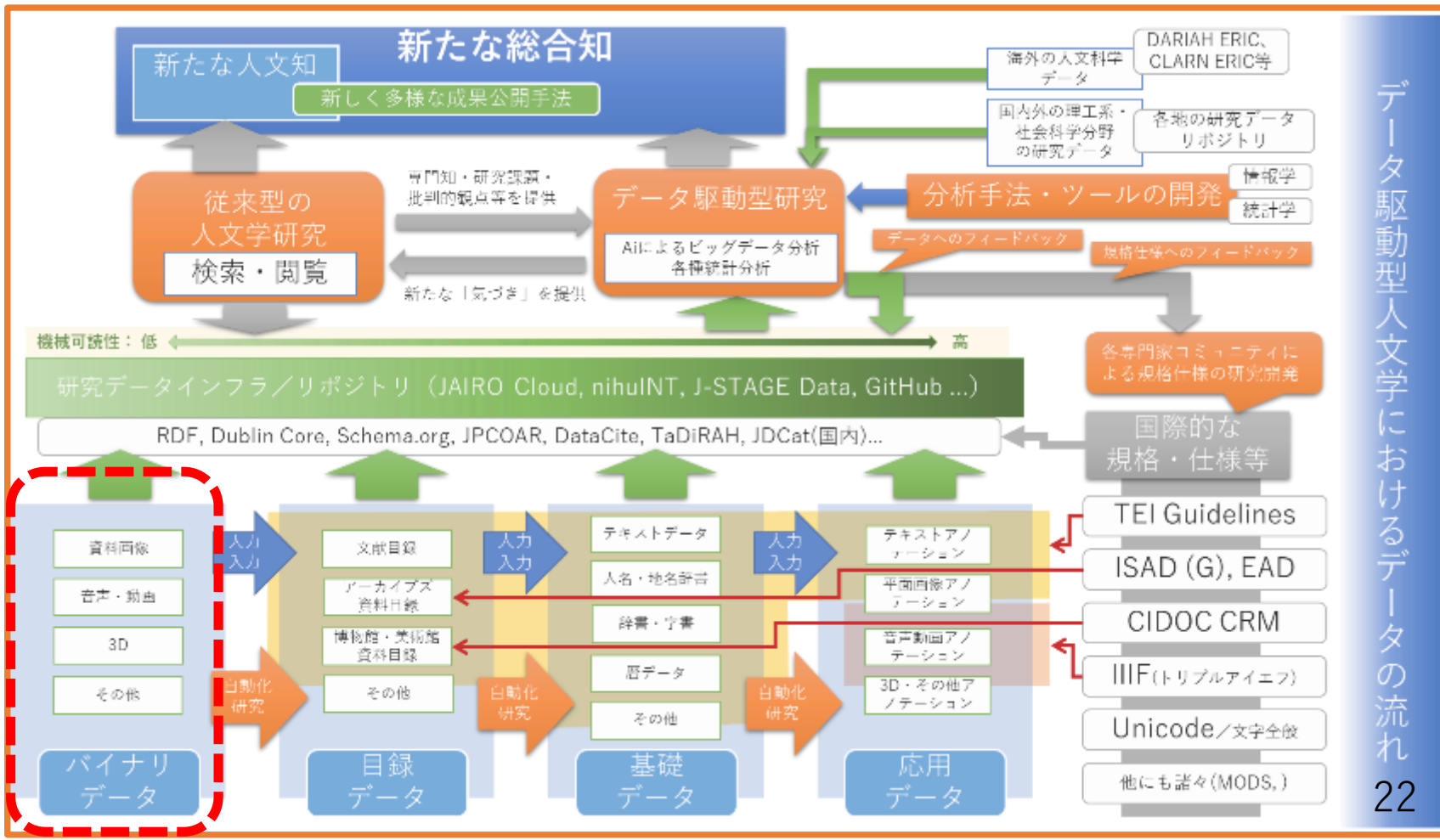
データ駆動型人文学のデータの流れを踏まえたフローの事例

- SAT大蔵経データベース https://21dzk.l.u-tokyo.ac.jp/SAT/

バイナリデータ

撮影・公開画像
- 大正新脩大蔵経 頁画像
- 大正新脩大蔵経 図像編画像
- 万暦版大蔵経 頁画像
- その他仏典頁画像

計約30万コマ・約50TB



データ駆動型人文学におけるデータの流れ

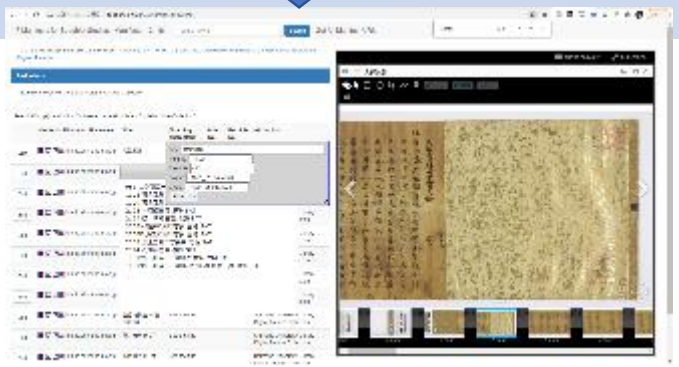
データ駆動型人文学のデータの流れを踏まえたフローの事例

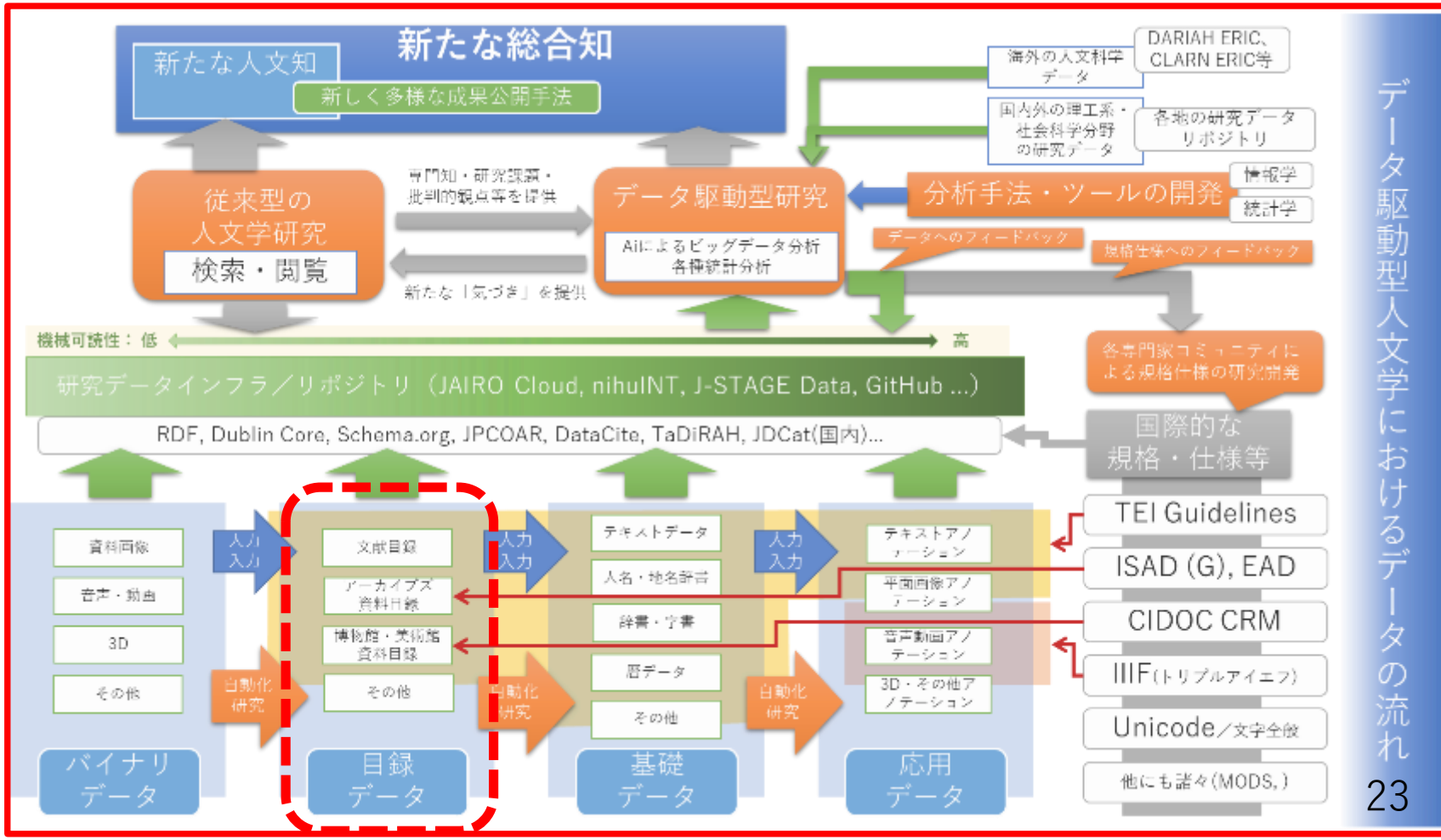
- SAT大蔵経データベース https://21dzk.l.u-tokyo.ac.jp/SAT/

目録データ

文献目録

- 仏教学独自のモデル
- 記述方法はTEIガイドラインに準拠中
- 外部サイトの仏典画像も対象
- Web協働編集システムで構築中





データ駆動型人文学のデータの流れを踏まえたフローの事例

- SAT大蔵経データベース https://21dzk.l.u-tokyo.ac.jp/SAT/

基礎データ

TEIガイドラインに移行中

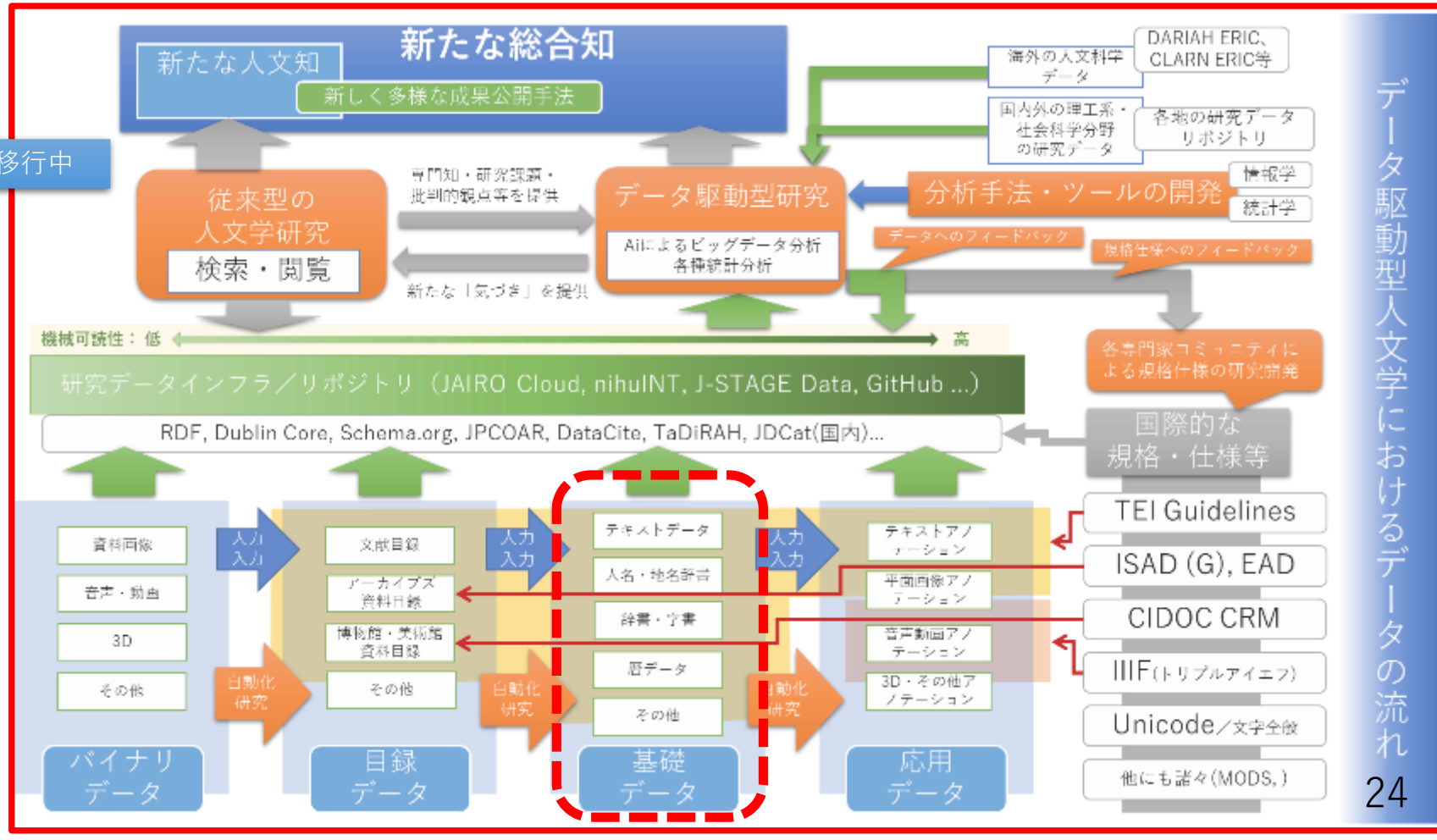
テキストデータ
- 1億字強の全文テキストDB
  - 人の目で全体を2回校正
- 現代日本語訳データ

辞書・字書

TEIガイドラインに準拠
- 外部DBと動的連携
  ≒データ再利用

一部TEIガイドラインに準拠

辞書DBとの連携



データ駆動型人文学のデータの流れを踏まえたフローの事例

- SAT大蔵経データベース https://21dzk.l.u-tokyo.ac.jp/SAT/

応用データ

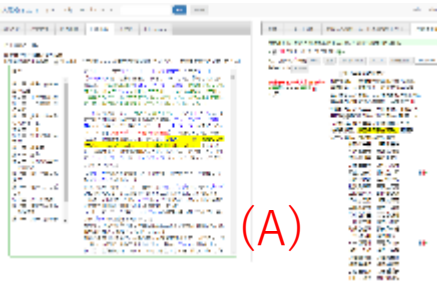
TEIガイドラインに準拠

テキストアノテーション

- 現代語訳と原文との文章単位でのリンクデータ…(A)
- 文書間の文章・フレーズ単位での引用構造を記述

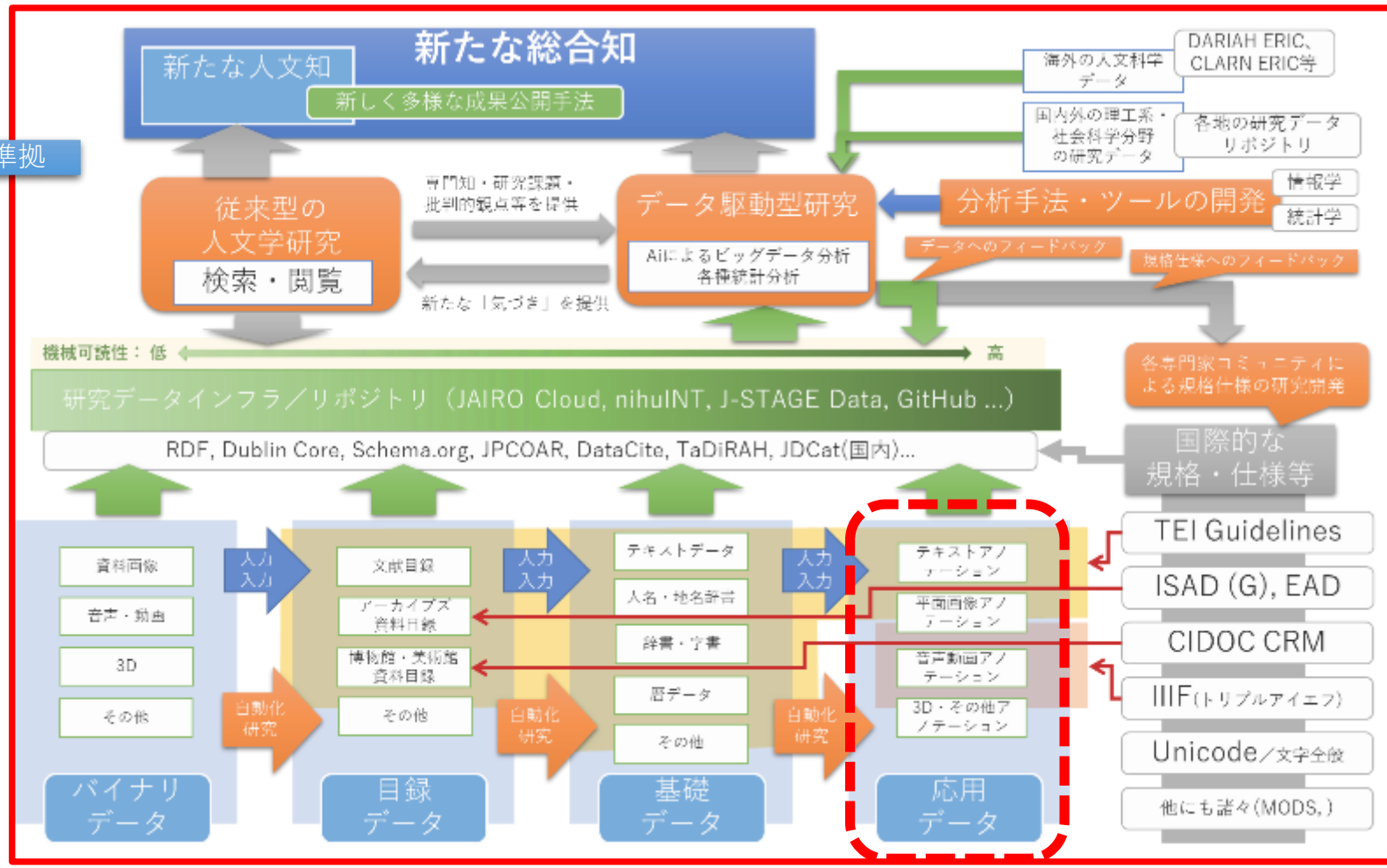
画像アノテーション

- 異体字の字形をIIIF準拠で記述・表示
- 異文を画像でIIIF準拠の記述・表示…(B)
- 仏画の各種属性をIIIF準拠で記述・表示



(A)

(B)

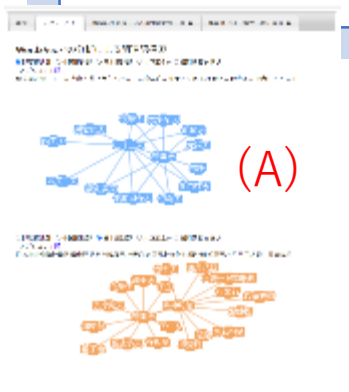


データ駆動型人文学のデータの流れを踏まえたフローの事例

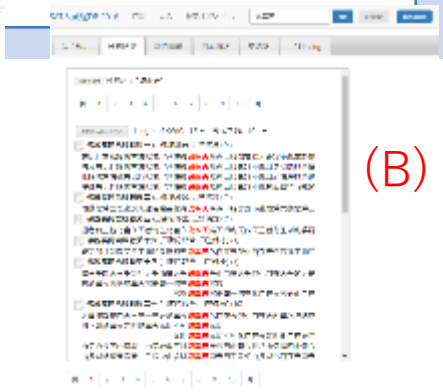
- SAT大蔵経データベース https://21dzk.l.u-tokyo.ac.jp/SAT/

データ駆動型研究

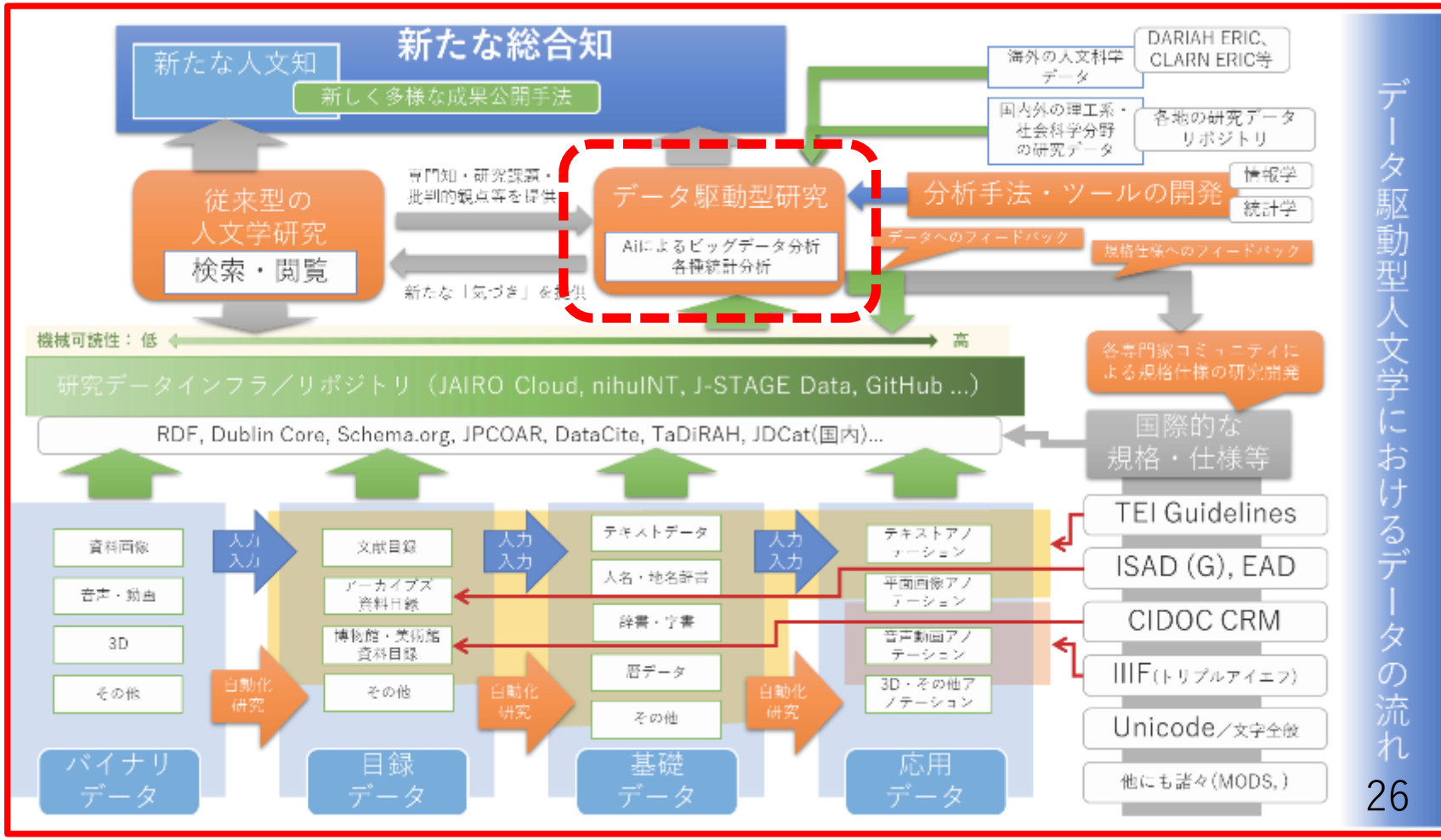
- AI関連技術による仏典分析機能…(A)
  - Word2Vecを用いた任意のカテゴリにおける単語の文脈分析と比較機能
- 単語の登場頻度によるテキスト分析機能…(B)
  - 大規模テキスト向け全文検索ソフトウェアによる高速かつ簡便な機能
- 脚注の統計分析による伝承系統の研究



(A)



(B)



データ駆動型人文学におけるデータの流れ

データ駆動型人文学のデータの流れを踏まえたフローの事例

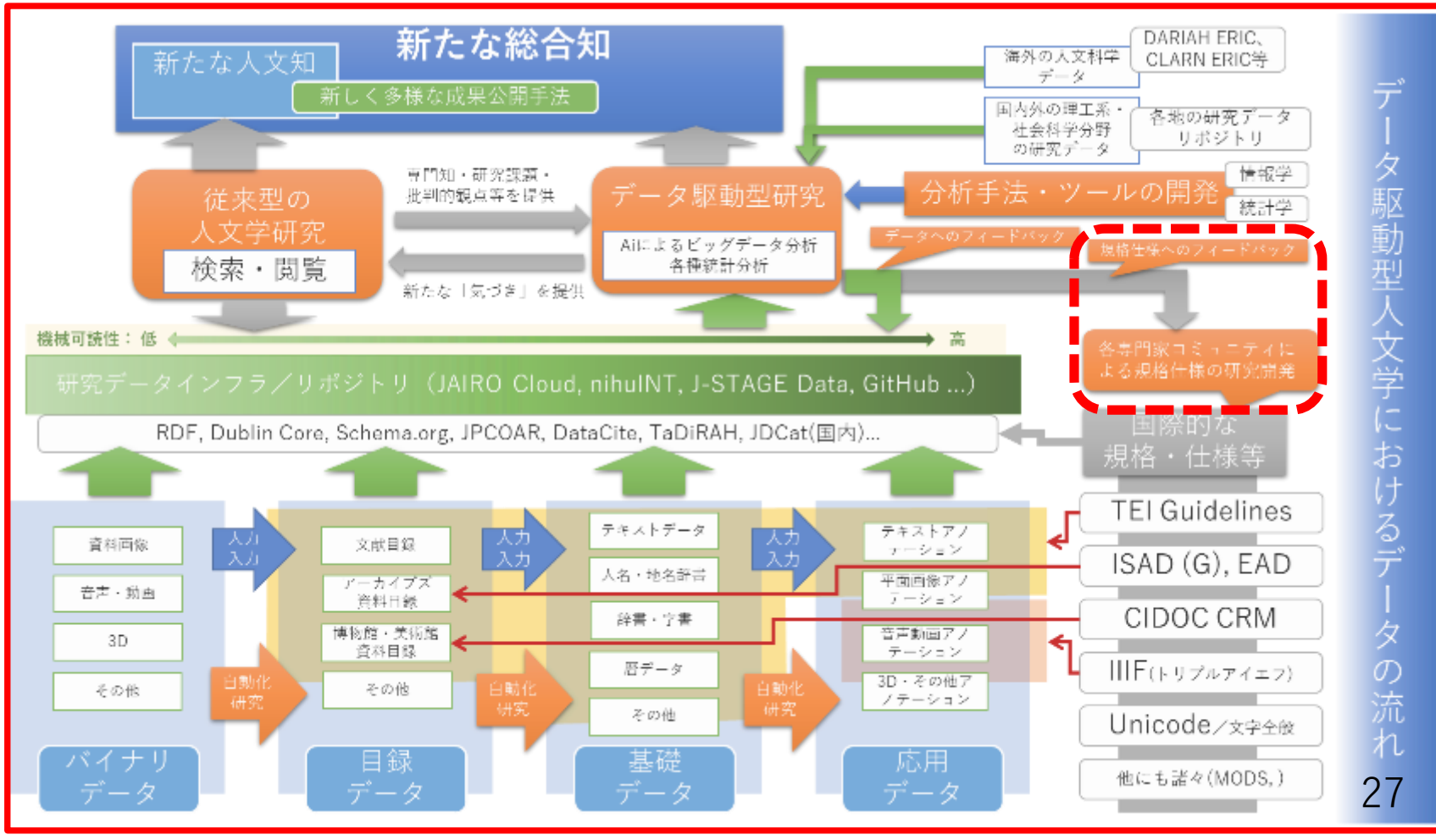
- SAT大蔵経データベース https://21dzk.l.u-tokyo.ac.jp/SAT/

規格仕様への
フィードバック

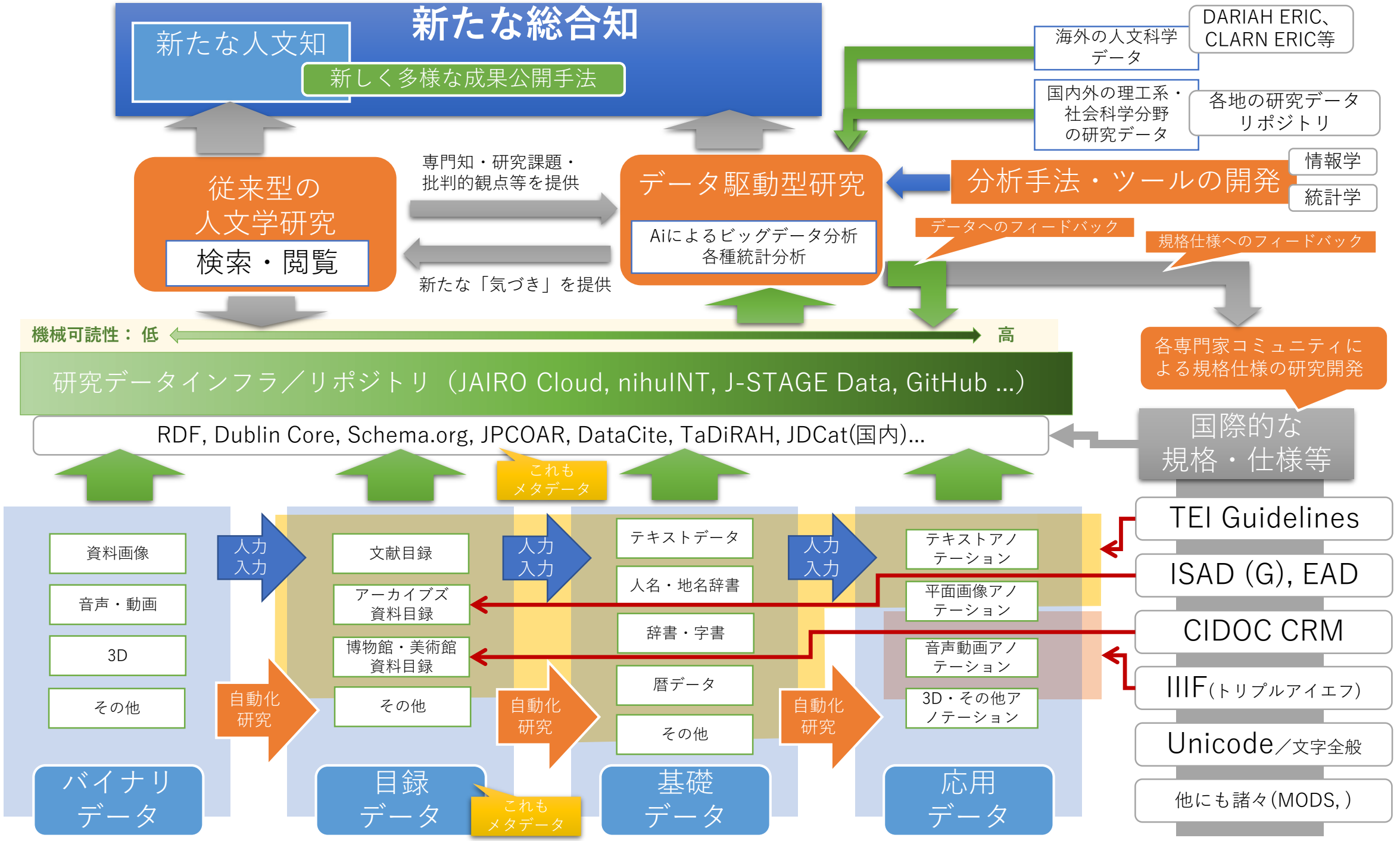
- ISO/IEC 10646への文字の符号化提案…(A)
  - 漢字3000字超、悉曇（梵字）の外字6文字及び悉曇への異体字処理機構導入の提案 （リンク1, リンク2）
  - 漢字に関しては学術団体として世界初の主体的参画
- TEIガイドラインにおける東アジア／日本語資料への対応強化の提案…(B)
  - 東アジア／日本語分科会の設立提案
  - ルビのセマンティクスの導入提案

(A)

(B)



データ駆動型人文学におけるデータの流れ


新たな総合知
新たな人文知
新しく多様な成果公開手法
海外の人文科学データ
DARIAH ERIC、CLARN ERIC等
国内外の理工系・社会科学分野の研究データ
各地の研究データリポジトリ
従来型の人文学研究
検索・閲覧
専門知・研究課題・批判的観点等を提供
新たな「気づき」を提供
データ駆動型研究
Aiによるビッグデータ分析
各種統計分析
分析手法・ツールの開発
情報学
統計学
データへのフィードバック
規格仕様へのフィードバック
機械可読性：低
高
各専門家コミュニティによる規格仕様の研究開発
研究データインフラ／リポジトリ（JAIRO Cloud, nihuINT, J-STAGE Data, GitHub ...）
RDF, Dublin Core, Schema.org, JPCOAR, DataCite, TaDiRAH, JDCat(国内)...
国際的な規格・仕様等
これもメタデータ
資料画像
音声・動画
3D
その他
人力入力
文献目録
アーカイブズ資料目録
博物館・美術館資料目録
その他
テキストデータ
人名・地名辞書
辞書・字書
暦データ
その他
テキストアノテーション
平面画像アノテーション
音声動画アノテーション
3D・その他アノテーション
自動化研究
TEI Guidelines
ISAD (G), EAD
CIDOC CRM
IIIF（トリプルアイエフ）
Unicode／文字全般
他にも諸々(MODS,)
バイナリデータ
目録データ
基礎データ
応用データ
これもメタデータ

国際的な動向への包括的な対応の必要性

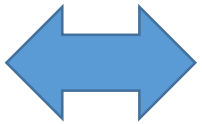
- 国際的なデータの規格・仕様の動向に対応するための方策の必要性

  - とりわけ、テキスト資料の構造をデータとして記述するTEIガイドラインへの対応は、その重要性にも関わらず日本では進んでいない。
  - 規格・仕様を推進する専門家コミュニティとの連携
  - 規格・仕様を日本で受容・検討・改訂するための環境整備
    - 日本語訳の共有・規格検討の場の設定・ツールの開発の共有…

人文学DXに向けた要件

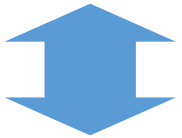
- データ構築のノウハウの集約と人材育成
  - 既存の学会、コミュニティ等との連携
    - JADH、IPSJ SIG-CH、RDUF…
- 国際的な規格に準拠したデータ構築のための継続的活動基盤
  - 規格の適用方法の確立
  - 規格改定のための検討と手続き
- データを蓄積するためのインフラの構築・維持
  - 国立国会図書館に納本された書籍の持続可能性と同等を目指す（JAILO CLOUDの活用）
- 構築されたデータを利活用可能な手法・ツールの開発
  - 人文学データに適した手法・ツールを開発する体制の整備・開発
  - 各実施機関への提供とフィードバックの収集
- 人文学側における開発者・運用者の育成（DH教育の一環）
  - すべて独力で開発する必要はないが、設計や運用の能力が人文学側に必要
- 上記の成果を広く周知・共有し総合知へと接続する枠組みを提供
  - JDCatの拡張

拠点機関



- 国文学研究資料館大型フロンティア事業による古典籍大規模デジタル画像
- 国立国語研究所による各種コーパス
- 東京大学史料編纂所による歴史データ
- CODHの各種人文学データセット
- 各種分析ツール（KHCoder、Voyant tools…）
- …

既存のリソース



実施機関

- データセットの構築・運用・分析
  - キュレイター、エンジニア、アナリストによる協働
  - 新たな独自のデータセットの構築・運用・分析
  - 既存のリソースの分析と連携活用⇒総合知への接続
- 成果の公表
  - 国内外の学会等での研究成果としての発表
  - 利用条件を整備した再利用可能なデータセットの公開
- 拠点機関の活動との連携
  - データ構築・運用・分析の経験知の提供と共有
  - 国際規格の適用方法や改訂についての情報提供と協働
  - 拠点開発ツールの適用とフィードバック
  - 独自開発ツールの拠点への提供
  - 人材育成に際しての積極的な協力
  - データセット公開にあたっての協働

データを活かした協働による成果発信