

CODHシンポジウム

日本文化とAIシンポジウム2019

AIがくずし字を読む時代がやってきた

木簡情報のオープンデータ化と文字画像DB
連携の強化

独立行政法人 国立文化財機構
奈良文化財研究所 都城発掘調査部
史料研究室長

馬場基



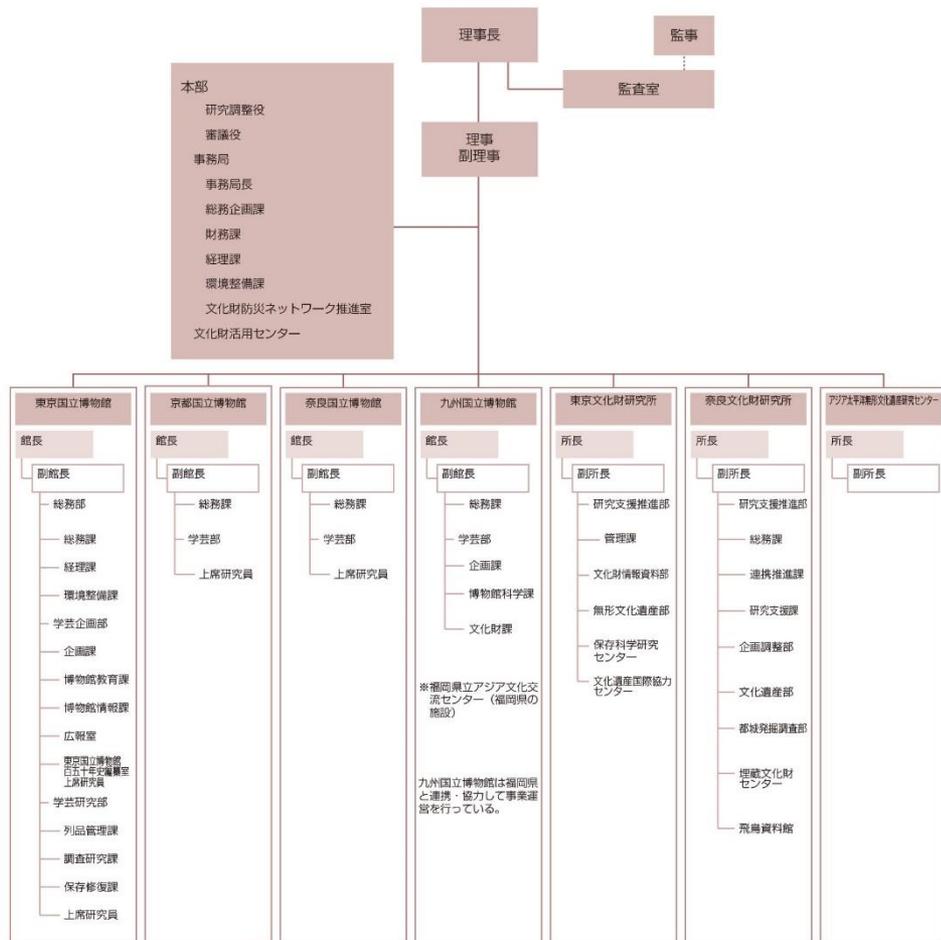
所管の法人等

独立行政法人国立文化財機構 法人番号

3010505001183

独立行政法人は国が提供していた行政サービスをより柔軟に実施するために国から独立した組織です。「独立行政法人国立文化財機構」は、東京国立博物館、京都国立博物館、奈良国立博物館、九州国立博物館の4博物館を設置し、有形文化財を収集し、保管して国民の皆様の観覧に供するとともに、4博物館と東京文化財研究所、奈良文化財研究所、アジア太平洋無形文化遺産研究センターの計7施設にて文化財に関する調査及び研究等を行うことにより、貴重な国民的財産である文化財の保存と活用を図ることを目的としています。これにより文化財の保存と活用をより一層効率的かつ効果的に推進し、文化財保護行政を支えてまいります。

組織図





インデックス付き
システムファイル
(題籤軸)



キーホルダー



膨大な消し残り
(削屑)



古代式生体認証IDカード
(画指)



日本全国の木簡出土点数: 40万点以上
奈文研所蔵木簡点数: **25万点**以上

古代式嚴封封筒
(封緘木簡)

「平城宮跡出土木簡」国宝指定 (2017/9)

常食菜甚惡



赤外線画像

此
所
不
得
小
便

ナマの素材としての可能性

直接様々な情報を伝える
多様な視点からの分析が可能

学会・社会からの関心

関心が高く、積極的公開の必要

資料としての「弱点」

文字数が少なく断片的
膨大な手間と時間が必要

重要・魅力的⇔近づきにくい

変化する情報集合

様々な遺跡・遺構から次々と出土
情報が分散しがち

脆弱遺物という限界

現物の調査を開放できない
資料の真実性をいかに担保するか



情報学との連携による解決を模索

データベース開発・提供を軸とした究を展開

1 出土文字史料に関する国内最大の「現業部門」

喫緊の課題

- ・ 史料体の管理・保全を廻る問題
 - 管理台帳・管理状況情報・作業効率化 等
- ・ 史料情報の公開・公表を廻る課題
 - 画像データ作成・文字情報検討・公開 等

2 出土文字史料研究のナショナルセンター

喫緊の課題

- ・ 国際共同研究を含めた出土文字史料研究の促進
 - 研究情報の共有・共同研究
 - 文字研究・歴史研究等機関との連携
- ・ 史料の保全や公開の促進
 - 国内埋蔵文化財調査期間との協力

1999年
木簡DB公開

東京大史
奈良文化財研究所と奈良文化財研究所との連携による、木簡データベースの公開。これは、これまで、それぞれが持っている古い文字を公開するシステムを開発した。



システムの高次化
⇒木簡研究・公開の強力なツールに発展

2005年
木簡字典公開

「し、今秋か、発した。」から18世紀より、7世にわたる木簡の選が一目で分かる。歴史中、古文書など、歴史的な文書や、その歴史をみる歴史学、東大編纂所は奈良時代から明治維新までの古文書や公文



- ・木簡の個々の文字が一覧可能
- ・類例を元に釈文を検討
- ・木簡の文字の研究の展開
- ・豊富な画像公開

2009年
東京大学史料編纂所との連携DB公開

書など紙に書かれた、奈良文化財研究所、研究する、唯一の研究機関。木簡のデータベース上で公開している文字は、2機関合わせて約2万種、画像数では18万件近くに上り、データベースへのアクセス数は東大編纂所の



- ・研究機関を超えての連携
- ・前近代の1000年以上の文字画像を一度に探せる
- ・代表字形のみを一覧可能
- ・一般利用者の利用もより容易に



寛書を交わす東大史料編纂所の加藤友康所長(左)と奈良文化財研究所の田辺征夫所長

2016年

MOJIZO<画像から検索>公開
(東京大学史料編纂所との連携)

2016年

MOJIZO<画像から検索>公開
(東京大学史料編纂所との連携)

- ・画像から文字画像を検索する新システムの実現
- ・研究者にとって頼もしい「相談相手」の登場
- ・一般利用者にとって木簡やくずし字がさらに身近に

東大で開かれた記者会見には、東大側から加藤友康所長、奈良側から田辺征夫所長が出席。奈良文化財研究所は過去40年以上、世界文化遺産にも登録されている平城宮跡の発掘調査を続けており、田辺所長は「今後新たに見つかる文字を解明する際にも、今回の提携は大きな意義がある」と語った。【井上樹】

史料の崩し字 スマホで解読

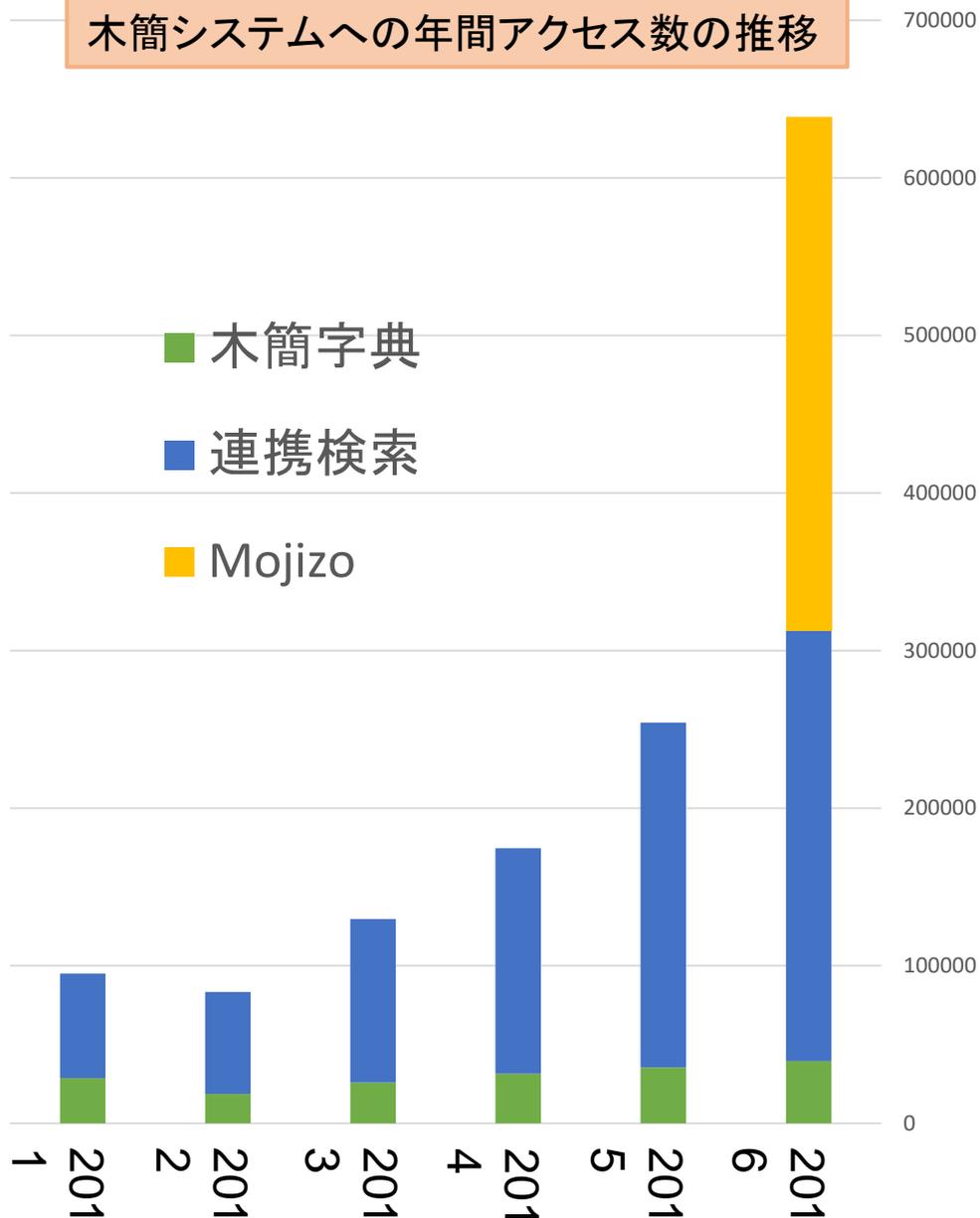


奈文研など 画像検索システム公開

専門家でも判読が難しい木簡や古文書などの崩し字の解読を手助けする画像検索システムを、奈良文化財研究所(奈研)と東大史料編纂所が共同で開発し、25日、奈良文化財研究所で公開した。スマートフォンでも利用でき、木簡の文字を調べると、その文字の類似性の高い木簡の文字画像を一度に探せる。代表字形のみを一覧可能に。一般利用者の利用もより容易に。研究機関を超えての連携。前近代の1000年以上の文字画像を一度に探せる。代表字形のみを一覧可能に。一般利用者の利用もより容易に。



木簡システムへの年間アクセス数の推移



大きな成果を達成

新たな3つの課題・期待

- 1 データの「量」
- 2 データの「質」
- 3 連携力の強化

*連携検索は2013年度分のログ未取得のため推計値

1 利用者増加＝網羅性の必要 ⇒データの「量」

木簡字典

収録データ内容：飛鳥・藤原京跡、
平城宮・京跡、
各地出土木簡を収録

収録データ数：文字種 約 1,900種
木簡点数（表裏別） 約 15,000点
文字画像数 約100,000件

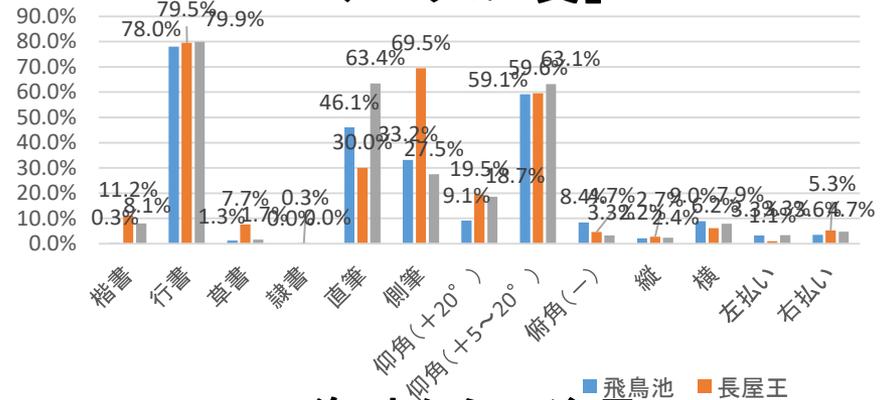
それでも全体に対する比率は決して高くない

- ・ 出土木簡点数40万点以上
木簡DB収録点数約17万点（約43%）
- ・ 木簡DB収録文字数約30万字
木簡字典収録文字数約10万字（約33%）
・・・全出土「文字」の25%程
（出土木簡中には全く読めないものも含まれるため）

拠点機関が「提供」するだけでは限界も・・・
国際連携には従来の方法では難しい……

より「連携」しやすい・より「共有」しやすい
多くの研究者・市民がつながる「参加誘発型」のスキームが実現できないか？

2 多分野での研究が展開 ⇒データの「質」



3 海外からの注目 ⇒連携力の強化

朝日新聞 2016/02/19 (金) 31面



協力を交わす松村恵司所長(左)と台湾の中央研究院歴史語言研究所の黃進興所長。奈文研提供

奈文研 台湾の研究所と協約 木簡読み解きでタッグ

奈文研の奈良文化財研究所は17日、台湾の中央研究院歴史語言研究所（史語所）と、力を合わせて木簡の読み解きや漢字文化の研究にあたる協約を結んだ。

史語所は、漢代の簡牘（木簡）の研究拠点。持っている簡牘の多くはモンゴルに近い砂漠地帯の出土品。ちぎれちぎれになったものがほとんど。日本の木簡に比べ、状態がよいという。紙が広まる前の時代、版簿や手紙、荷札など幅広く使われた。

奈文研所は今後、分析の手法を伝え合ったり、書かれた内容と比較したりする「対照」や漢字文化の考察を深めることを目指す。奈文研（栗田優美）

研究からはまず、木簡に赤外線などを当てて撮影し、表面の劣化などで見分けにくくなった文字を読み取る方法を伝授。将来は、双方のデータベースをつなげることも視野に入れる。

台湾での締結式に臨んだ松村恵司・奈文研所長は「双方の経験、知識、画像データなどの資源を提供し合うことは、文化財保全と研究発展に役立つと期待している」とコメントを出した。

出土文字史料情報のオープンデータ化
出土文字史料の連携強化
出土文字史料情報の多様化・多元化

科学研究費補助金の交付を受けての研究・開発
国立文化財機構の事業の中への位置づけ

科研プロジェクトその1 木簡テキスト情報のオープンデータ化

https://mokkanko.nabunken.go.jp/ja/

Wooden Tablet Database

木簡をさがす

文字画像をさがす

キーワード: 武蔵国 × 032 ×

検索する

項目検索: 内容, 国名, 木簡の形, 出典, 年紀・人名, 大きさ・樹種等, 遺跡

検索条件: 《すべて検索》:【検索文字】武蔵国 032

ダウンロード

表示件数: 16 / 21

1 2 次

すべての検索結果: 25件

本文	寸法(mm)			型式番号	出典	遺跡名
	縦	横	厚さ			
□□(国古カ)五十戸	(119)	28	3	032	飛鳥藤原京1-724(荷札集成-273・飛13-13...	飛鳥池遺跡北地区
无耶志国薬鳥	162	17	4	032	藤原宮4-1726(飛20-31下・荷札集成-74・木...	藤原宮跡西南南門地区
白朮四□	90	16	3	032	藤原宮4-1731(木研11-34頁-2(19)・飛9-10...	藤原宮跡西南南門地区

絞込

- ◆型式番号
032 (24件)
081 (1件)
- ◆内容分類
付札 (3件)
荷札 (21件)

Wooden Tablet Database

木簡庫とは

使い方

ヘルプ

いいね!

シェア

B! 2

木簡をさがす

すべて

本文

カテゴリー
(意味検索)

文字画像をさがす

テキストから

画像から

キーワード

武蔵国 x 032 x

検索する

<<TOPへ戻る

項目検索

■検索条件: 《すべて

すべての検索結果: 25件

検索結果一覧データのダウンロード

検索結果データの「本文・寸法・型式番号・出典・遺跡名・調査主体・発掘回数・遺構番号・地区名・R番号」の一覧表をダウンロードできます。地区名とR番号を組み合わせると、木簡のIDになります。

データのダウンロードおよび利用に際しては、以下の内容に同意してください。

- 1 データの利用にあたっては、出典(奈良文化財研究所・木簡庫データベース)を明記してください。
- 2 データの利用は、利用者の責任において行うものとします。内容の正確さには万全を期してはいますが、データを利用したことにより、また利用できなかったことによりいかなる損害が生じても、当研究所は一切の責任を負いません。
- 3 木簡庫データベースのデータは、予告なく修正する場合があります。

同意する

検索結果一覧データをダウンロード

機構プロジェクト 木簡画像のオープンデータ化

Browser address bar: <https://colbase.nich.go.jp/>

Page title: ColBase

Navigation: 国立文化財機構 | National Insti... ColBase

ColBase 国立博物館所蔵品統合検索システム

博物館を選ぶ Language

フリーワード (作品名・作者など) 検索

画像あり

詳細検索

Pick Up



閲覧ランキング

- 
- 
- 
- 
- 

各機関より

九州国立博物館

釜を中心に344件の画像を追加し...

ColBaseより

Colbase

国立文化財機構の4つの国立博物館の所蔵品を検索できる ColBase を公開しました。

2017年03月27日

100%

科研プロジェクトその2

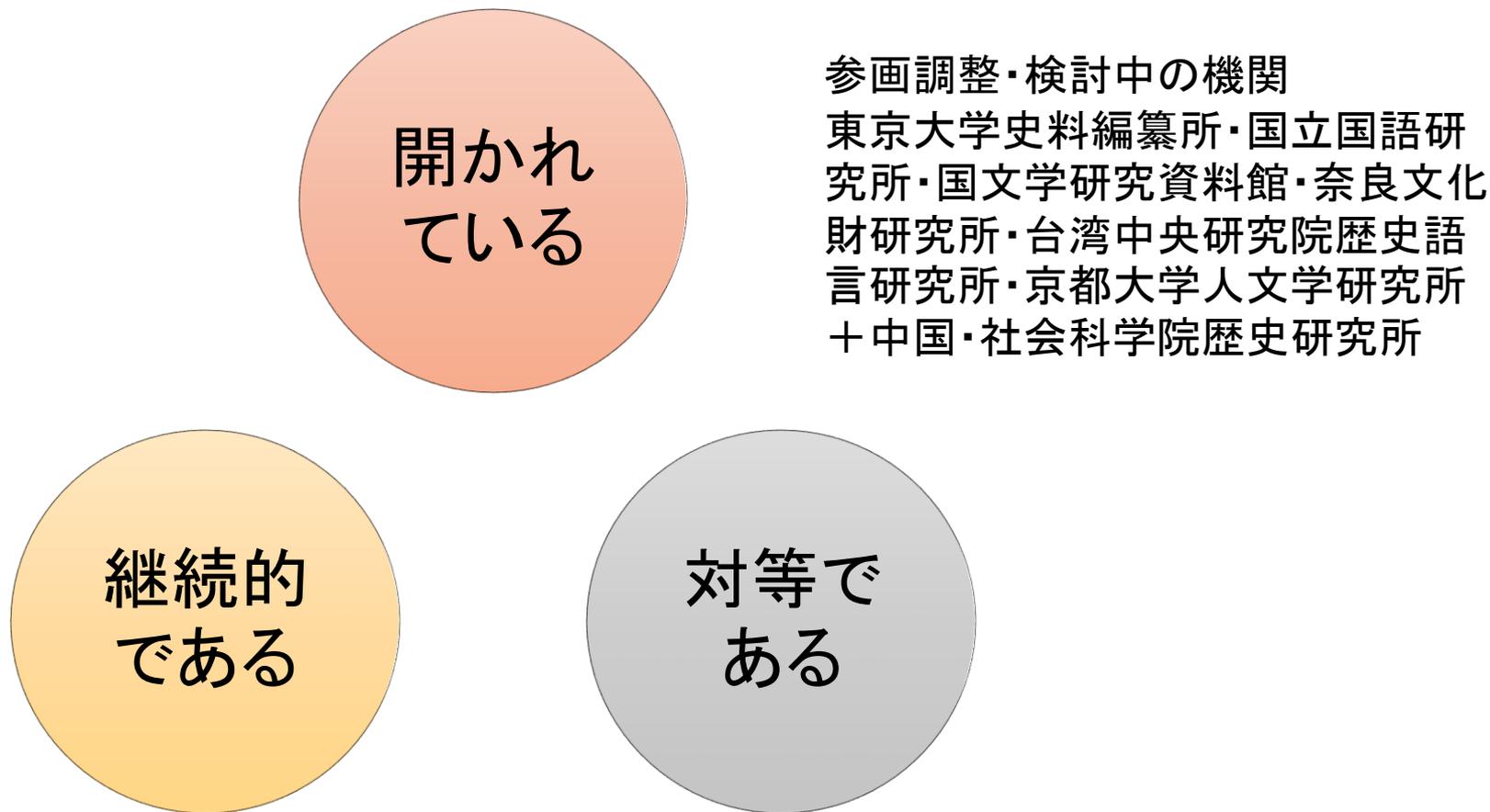
木簡文字画像の連携検索強化

今回の他機関連携の特徴
3段階の同意と共同の積み重ね

- 1 連携・協力の「意志」を共有
- 2 「意志」実現の方法を共有
 - ・・・宣言を年内完成を目指して最終検討
- 3 方法実現のためのツールを開発
 - ・・・β版ほぼ開発終了
 - 年度内を目処に公開準備

1 連携・協力の「意志」を共有＝コンセプトの作成

当初の開発参加機関以外の機関も容易に連携可能



長期的運用を目指す

データ公開機関が相互に対等

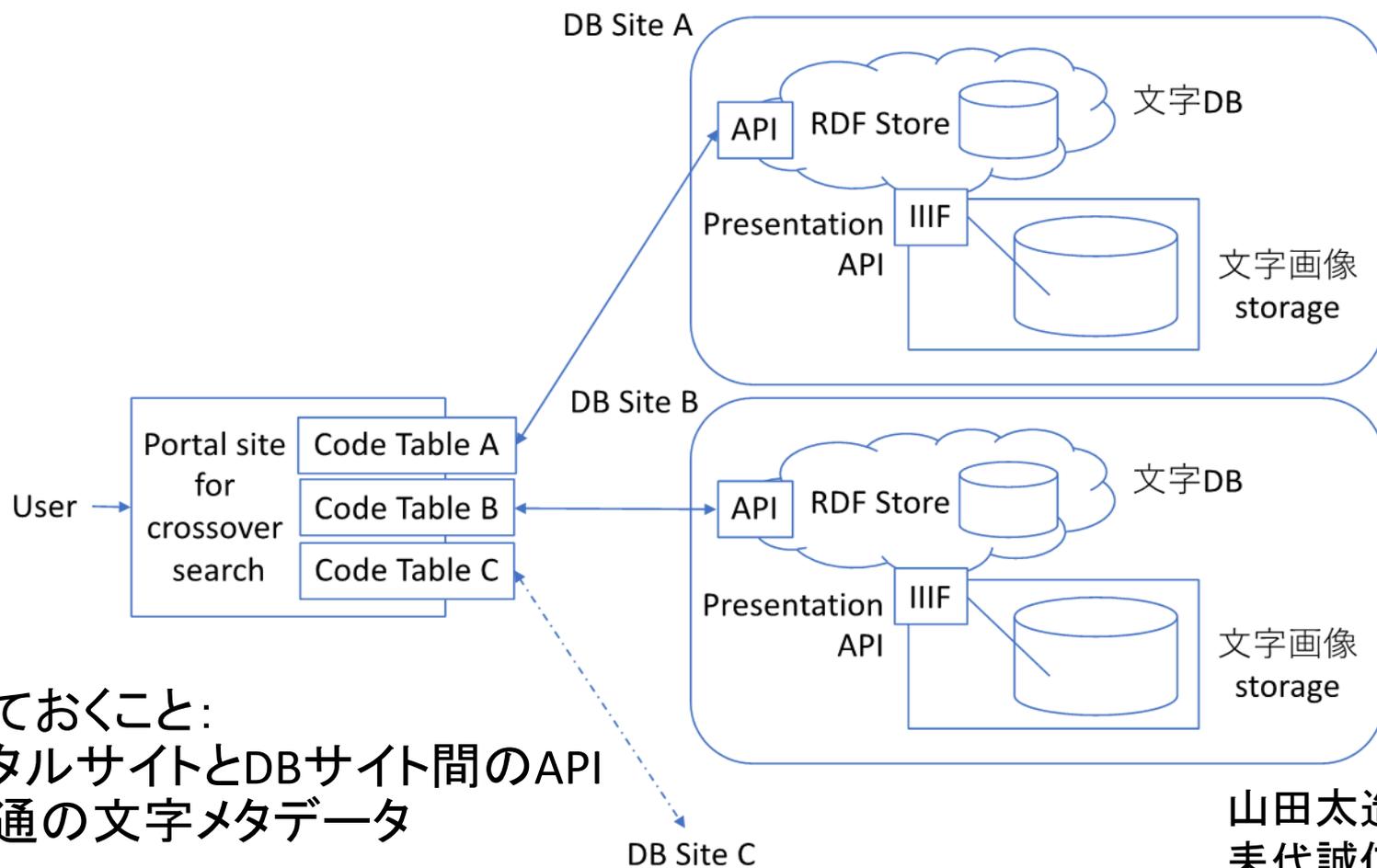
※みなさまのご参加を歓迎しお待ち申し上げます

2 「意志」 実現の方法を共有

汎用性の高い方法の組み合わせが妥当

「連携検索システム開発」ではなく

「共通検索を可能にするフレームワークの構築」



- 決めておくこと:
ポータルサイトとDBサイト間のAPI
= 共通の文字メタデータ

山田太造氏発案
末代誠仁氏作図

以下・山田太造氏による提言に基づき
科研研究グループで同意した内容です

共通メタデータ

メタデータは基本はDB/機関の事情による
共通するデータ項目は設定

文字：Unicode

符号化方式UTF-8で表現した文字

出典：URI？

時間：わかる場合のみ Hutimeの活用も

空間：地域程度でもよい

DB名：データベース名

機関名：機関名

文字画像：URIでよいか。

IIIF Presentation API対応・IIIF Image API？

切り取った文字画像そのもの？

各DBサイトの構築

各機関での事情に応じて構築

前述のAPIには対応

※IIIF Content Search APIでの検索

基本は単一のmanifestファイルを検索。

実装面も考慮すると、このAPIに従って実施する必要性はまだあまりないか

IIIF Presentation APIを利用

manifestのAnnotations・Web Annotationもあり？

表示

史料画像ありの場合：manifestファイル内で部分指定
ビューアでそれをオーバーレイ

ない場合：文字画像そのものを提示

manifestでの記述

文字に関するメタデータ・ライセンス

以上を私なりに簡単に要約すると

- 各文字画像はiiifに準拠したデータとする
- 各文字画像に関するマニフェストを公開する
- 各文字画像には以下の情報を付与する
 - 画像提供機関
 - 各画像の固有ID（機関ごとのルールによる）
 - 文字コード（unicode）
- 各文字画像には可能であれば以下の情報を付与する
 - 時空間情報
- 各文字画像は以下の条件を満たすものとする
 - 解像度300dpi程度（目標）
 - creative commons cc-BYSA 相当 以上
- 有効な連携検索のために補足的な情報を整備する
 - 文字コードマッピングデータ
 - 連携検索用API

3 方法実現のためのツールを開発

国際的・機関間の分業・連携により開発

文字コードマッピング

国語研・京大人文研・台湾中研院史語所

ポータルサイト開発

奈文研・東大編纂所

ポータルサイトネーミング

国語研

奈文研と編纂所のデータを検索できるβ版開発

→ ほぼ終了

実験ののち他機関対応作業へ着手

史的文字データベース連携検索ポータルサイト



The screenshot shows the search results for the character '国' (kuni). The search criteria are '検索文字: 国'. The results are displayed in two sections: '奈良文化財研究所' (Nara Cultural Heritage Research Institute) with 16 items, and '東京大学史料編纂所' (The University of Tokyo Institute of Cultural Materials Studies) with 4 items. Each item includes a thumbnail image of the character on a document and a '詳細' (Details) link. Below the results, the text 'III Fビューア (Mirador) 各機関のDB' (III F Viewer (Mirador) Databases of each institution) is shown, with arrows pointing to the Mirador viewer and the database details.

Coming soon

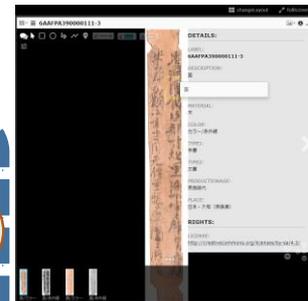
This section lists databases that are 'Coming soon' and are enclosed in a blue dashed border:

- HNG 漢字字体規範史DB** (HNG Kanji Font Style Normative History DB)
- 台湾中研院 居延漢簡** (Taiwan Academia Sinica Juyan Han Slips)
- 国文研 日本古典籍くずし字データセット** (Kokugon Ken Japanese Classical Texts Kuzushiji Data Set)

検討中

This section lists databases that are 'Under consideration' and are enclosed in a green dashed border:

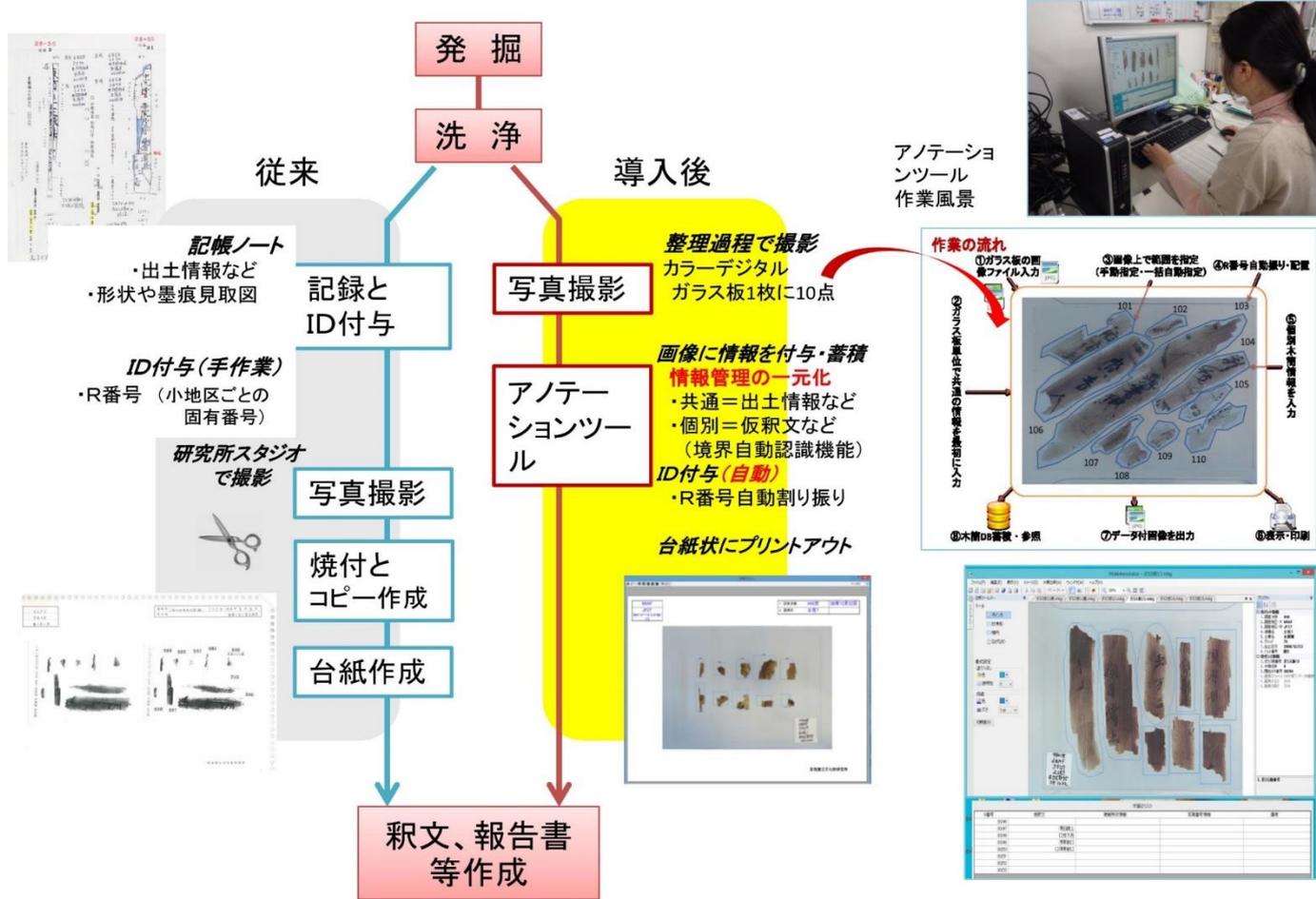
- 中国 中国社会科学院 歴史研究所** (China Chinese Academy of Social Sciences Institute of History)



○今回のポータルサイトは一事例

調査・研究への展開

削屑の整理作業とアノテーションツールの導入

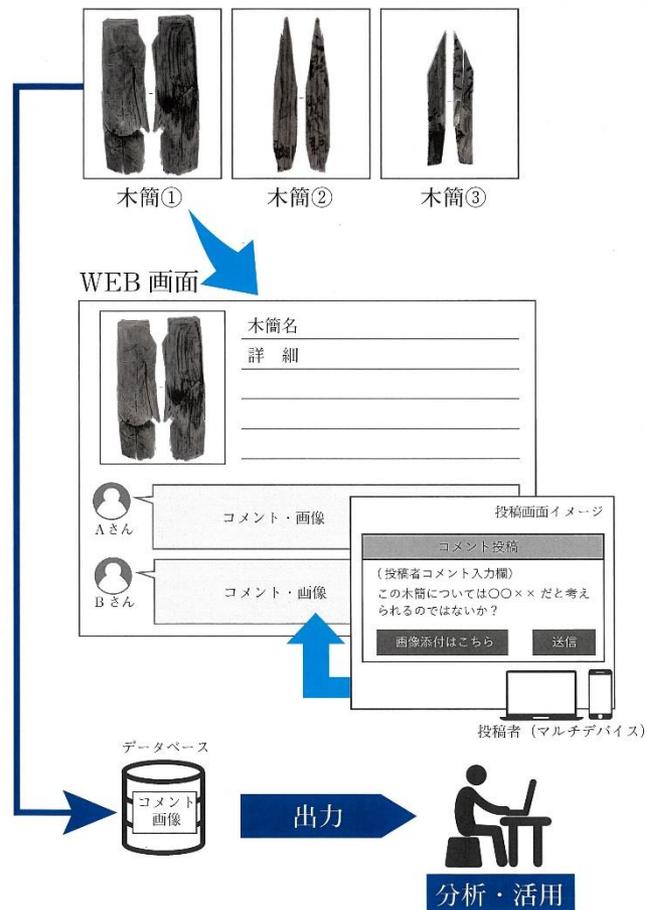


作業の効率化・情報の一元化 (情報間の関係性の構築)
 + 経験知の収集・蓄積

多様なプレイヤー参加による「知」の爆発的展開への期待



多機関連携ポータルサイト



様々な知識・バックグラウンドを持つ人々の気付き・暗黙知を集める

↓
分析・新たな発見

集めるところまでは検討中
分析手法は未着手

「しょうてん」の問題

EsT character = 升

Edit New Account RDF (Turtle)

&BUCS+S347; → 包摂

升 升

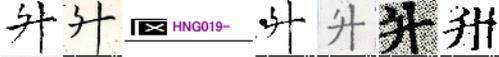
部首：十部 (R024)

漢字構造：  升

= UCS: U+5347 (21319)  

←denotational: &BUCS+S347;

→ [HNG] 中国写本: 

→ [HNG] 中国版本:  

→ [HNG] 日本写本: 

→ 説文小篆: 

→formed: 

<http://www.chise.org/est/view/character/升>

1/3 ページ
連携検索【代表文字一覧画面】

1/2 ページ

TOP 奈良文化財研究所 東大大学史料編纂所

『木簡画像データベース・木簡字典』『電子くずし字字典データベース』連携検索



検索文字 = ' 升 '

■検索文字: 升 ■代表文字検索結果 9件 ■全ての文字画像を表示



各画像下の「詳細」ボタンを押すと奈良文化財研究所「木簡字典」の詳細画面が表示されます。

電子くずし字データ ■代表文字検索結果 14件 ■全ての文字画像を表示 (丑)

					
年月日未詳 基沼寺文書	慶長4年3月3日 小早川秀秋知行方目録	(年月日未詳) 長谷場文書	(年月日未詳) 長谷場文書	(年月日未詳) 長谷場文書	(年月日未詳) 長谷場文書

<http://r-jiten.nabunken.go.jp/ichiran.php>

2019/01/14

木簡釈読者の「常識」と字体・字形研究者のコラボレーション

	書体					横角度			一画目					最終画					長い画			
	楷	行	草	隸		20° +	5~20°	5° -	ハネ	トメ	ハライ	側筆	直筆	ハネ	トメ	ハライ	側筆	直筆	タテ	ヨコ	左ハライ	右ハライ
合計	10.0	0.0	10.0	0.0	0.0	0.0	7.0	3.0	1.0	7.0	2.0	9.0	1.0	5.0	4.0	1.0	8.0	1.0	3.0	2.0	0.0	5.0
比率		0.0	100.0	0.0	0.0	0.0	70.0	30.0	10.0	70.0	20.0	90.0	10.0	50.0	40.0	10.0	80.0	10.0	30.0	20.0	0.0	50.0
合計	12.0	0.0	11.0	1.0	0.0	0.0	10.0	2.0	0.0	9.0	3.0	10.0	1.0	1.0	5.0	6.0	9.0	0.0	2.0	5.0	0.0	4.0
比率		0.0	91.7	8.3	0.0	0.0	83.3	16.7	0.0	75.0	25.0	83.3	8.3	8.3	41.7	50.0	75.0	0.0	16.7	41.7	0.0	33.3
合計	19.0	10.0	9.0	0.0	0.0	0.0	0.0	19.0	1.0	14.0	4.0	6.0	10.0	1.0	13.0	5.0	5.0	5.0	2.0	12.0	2.0	3.0
比率		52.6	47.4	0.0	0.0	0.0	0.0	100.0	5.3	73.7	21.1	31.6	52.6	5.3	68.4	26.3	26.3	26.3	10.5	63.2	10.5	15.8

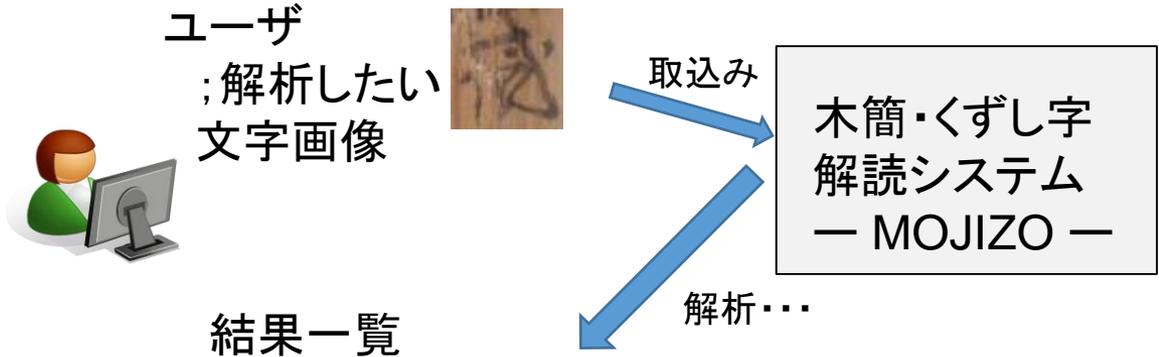
上から 晋 日 新羅

ある程度の方向性は確認できる可能性がありそう
 晋簡と日本木簡が近い・新羅木簡やや遠い という傾向か

個別の文字をみると、新羅と日本の類似も多く見られる
 例：「波」字を草書由来の書き方をする 等

個別事例（字形・用法）の蓄積＋量的事例（全体的傾向）

専門家でも解読できない2割に挑戦します



結果一覧
似ている画像を表示

上位8件画面→各機関100件まで表示する

奈文研
編纂所

リンク
それぞれのDBへ

- 奈文研
- 木簡字典DB
- 史料編纂所
- くずし字DB

文字画像の増加 → システムの強化・検索方法の強化が必要