電子くずし字字典データベースにおける現状と展望

山田太造（東京大学史料編纂所）

アウトライン

- 電子くずし字字典DB概要
- データ
- 他機関とのシステム連携
- 展望

東京大学史料編纂所

- 日本史に特化した唯一の附置研究所
- 明治の修史事業（明治2年（1869））
- 明治21年（1888）東京大学（当時帝国大学）へ
- 事業
  - 日本史史料の調査・蒐集（採訪）
  - 史料集編纂
  - 出版（冊子・データベース）

電子くずし字字典DB（1）

- 経歴
  - 2000年開発着手
  - 2001年データ追加
  - 2004年所内公開
  - 2006年所外公開

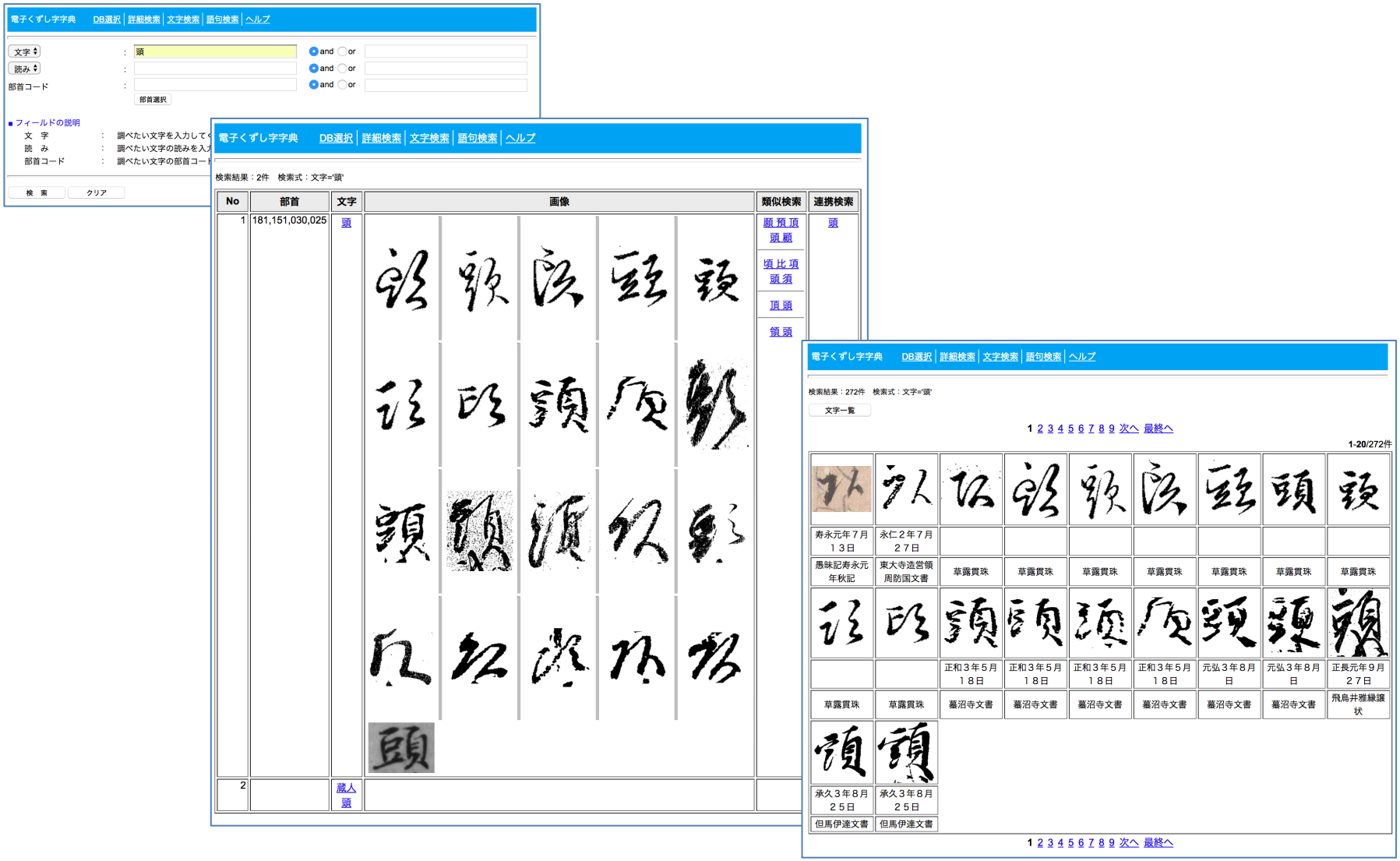
- データ
  - 史料編纂所が対象とする日本史史料から
    - 奈良時代から江戸時代初期（17世紀前半）
    - 102種類の史料群
  - 2013年度末
    - 単文字：198,852件（字種：5,963件）
    - 語彙：9,882件（語彙種：2,492件）

電子くずし字字典DB（2）

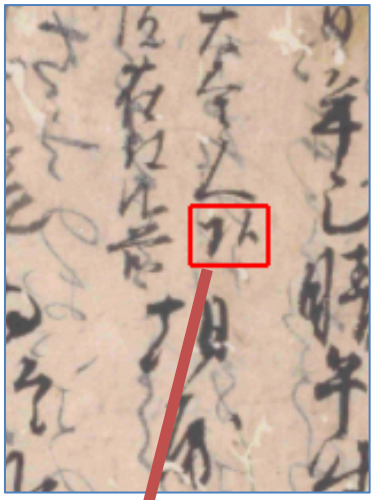
- ”各種史料のデジタル画像化を前提として，出典データを持った字形画像データの蓄積を意図したところである．”

- DB設計時のポリシー
  - 難読字形や特殊な字形のみを採集するのではなく，可能な限り網羅主義をとること，
  - 単文字のみならず語彙も採録すること，
  - 字形画像の出典が明示できること，
  - 史料編纂所の研究者が随時登録することが可能なこと，
  - 似た字形を参照できる機能をもつこと，

検索(1)


電子くずし字字典
DB選択 | 詳細検索 | 文字検索 | 語句検索 | ヘルプ
文字
頭
読み
部首コード
部首選択
■ フィールドの説明
文 字 ： 調べたい文字を入力して
読 み ： 調べたい文字の読みを入力
部首コード ： 調べたい文字の部首コード
検 索
クリア
電子くずし字字典
DB選択 | 詳細検索 | 文字検索 | 語句検索 | ヘルプ
検索結果：2件 検索式：文字='頭'
No
部首
文字
画像
類似検索
連携検索
1
181,151,030,025
頭
願 預 頂
頭 顧
頃 比 項
頭 須
頂 頭
領 頭
頭
2
蔵人
頭
電子くずし字字典
DB選択 | 詳細検索 | 文字検索 | 語句検索 | ヘルプ
検索結果：272件 検索式：文字='頭'
文字一覧
1 2 3 4 5 6 7 8 9 次へ 最終へ
1-20/272件
寿永元年７月１３日
永仁２年７月２７日
愚昧記寿永元年秋記
東大寺造営領周防国文書
草露貫珠
正和３年５月１８日
元弘３年８月日
正長元年９月２７日
藁沼寺文書
飛鳥井雅縁譲状
承久３年８月２５日
但馬伊達文書
1 2 3 4 5 6 7 8 9 次へ 最終へ

検索（2）

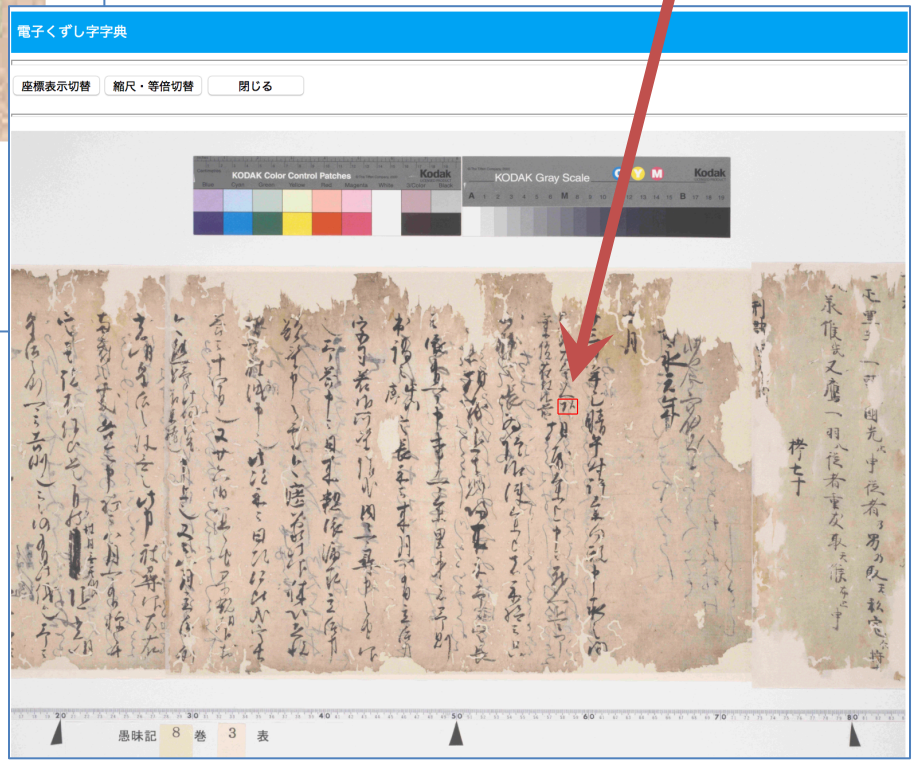


電子くずし字字典　DB選択 | 詳細検索 | 文字検索 | 語句検索 | ヘルプ

前　件　次　件　先　頭　最　終　一　覧　イメージ　類似検索1　類似検索2
類似検索3　類似検索4

1/272件

【文字】	頭
【部首】	181,151,030,025
【文字コード】	982D
【大漢和コード】	43490
【音読み】	トウ；ズ；ト
【訓読み】	あたま；かしら；こうべ
【史料区分】	原本
【架番号】	0073-10-8
【頁】	00000003
【和暦年月日】	寿永元年７月１３日
【文書名】	寿永元年七月十三日条
【史料群名】	愚昧記寿永元年秋記
【差出】	三条実房
【宛所】	
【備考】	
【属性】	公家



検索結果一覧

電子くずし字字典　DB選択 | 詳細検索 | 文字検索 | 語句検索 | ヘルプ

検索結果：2件　検索式：文字='頭'

No	部首	文字	画像	類似検索	連携検索
1	181,151,030,025	頭		願 預 頂 頭 顧 頃 比 項 頭 須 頂 頭 領 頭	頭

文字データ:単文字

管理番号	:	34020816
和暦年月日	:	正和３年５月１８日 西暦変換
西暦コード	:	1314 年 0 05 月 0 18 日 0

区分	:	単漢字
親字	:	頭 親字参照
語彙	:	
部首コード	:	181,151,030,025
音読み	:	トウ；ズ；ト
訓読み	:	あたま；かしら；こうべ
文字コード	:	982D
大漢和コード	:	43490
代表フラグ	:	☑ 代表文字

史料区分	:	000 イメージ表示
架	:	0071
番	:	13
枝番	:	
号	:	
丁頁	:	00000011
拡張子	:	.tif
画像ナンバー	:	00011075.bmp
切取座標開始（X）	:	
切取座標開始（Y）	:	
切取座標終了（X）	:	
切取座標終了（Y）	:	

文書名	:	一宮修正会勤行所作人注文
史料群名	:	蟇沼寺文書
差出	:	
宛所	:	
備考	:	
分類	:	☐ 公家 ☐ 武家 ☑ 僧侶 ☐ 庶民 ☐ 女性 ☐ 署名 ☐ 在方

セキュリティ	:	00
最終更新者	:	wada(和田幸大)
最終更新日付	:	20061005

文字データ:語彙(文字列)

管理番号 : 34128613
和暦年月日 : 応永３１年９月 西暦変換
西暦コード : 1424 年 0 09 月 0 55 日 0

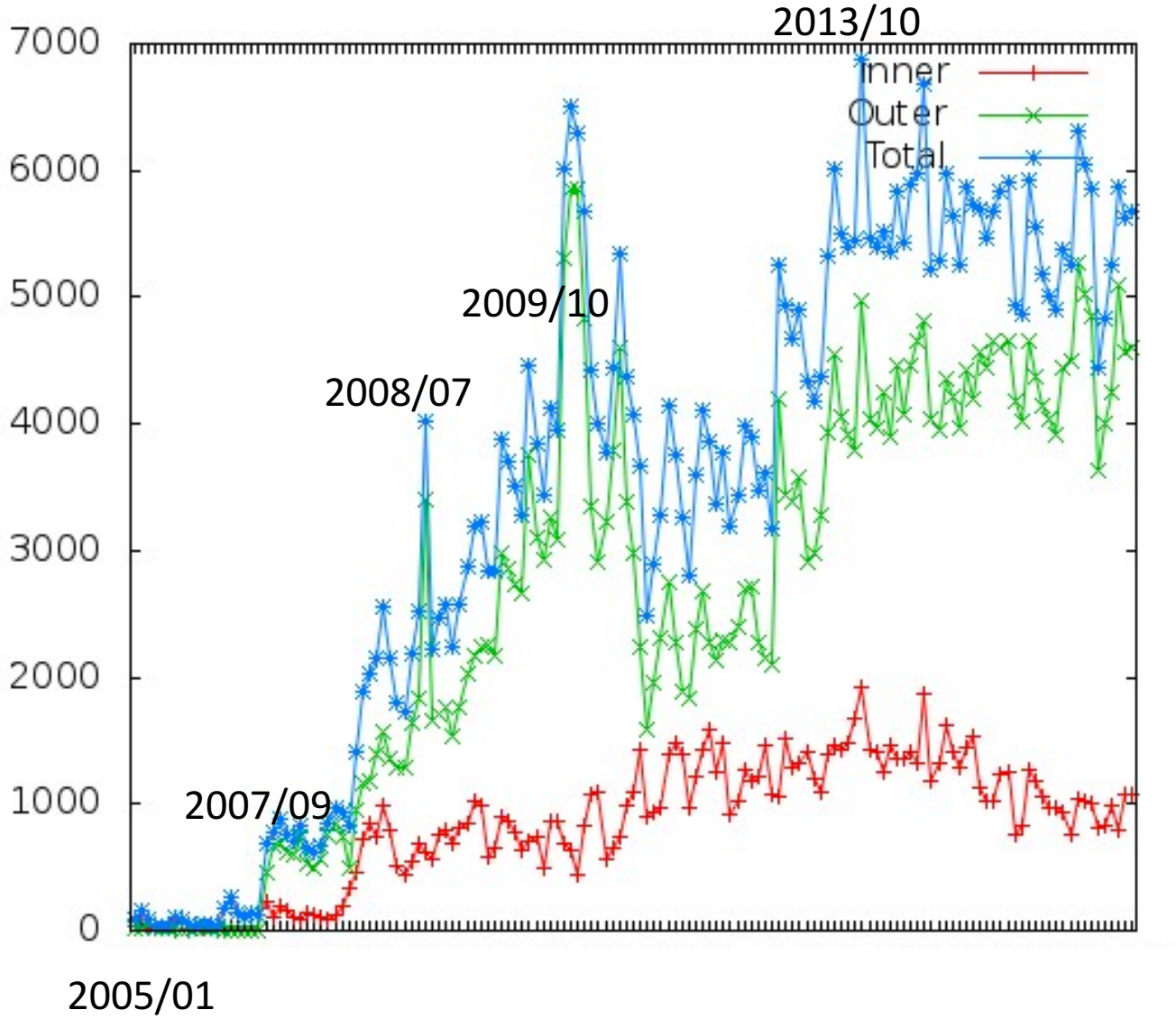
区分 : 語彙
親字 : 親字参照
語彙 : 内蔵頭
部首コード :
音読み : くらのかみ
訓読み :
文字コード :
大漢和コード :
代表フラグ : 代表文字

史料区分 : 000 イメージ表示
架 : 0673
番 : 5
枝番 :
号 : 2
丁頁 : 00000020
拡張子 : .jpg
画像ナンバー : 00384520.bmp
切取座標開始（X） : 893
切取座標開始（Y） : 675
切取座標終了（X） : 1035
切取座標終了（Y） : 958

文書名 : 応永三十一年九月条
史料群名 : 薩戒記
差出 : 中山定親
宛所 :
備考 :
分類 : 公家 武家 僧侶 庶民 女性 署名 在方

セキュリティ : 00
最終更新者 : g3401(電子くずし字字典画像切取ユーザ)
最終更新日付 : 20071107

アクセス数 :トップ画面へのアクセス数



ここ10年間
4,700件/月

奈良文化財研究所とのシステム連携

注）2009年

奈良文化財研究所とのデータベース連携公開に関する覚書の交換

5月29日（金）午前11時より史料編纂所中会議室（福武ホール地下1F）において、独立行政法人国立文化財機構奈良文化財研究所所長の田辺征夫氏ほか関係者をお招きし、データベース連携公開に関する覚書を交換しました。

覚書を交換する田辺征夫奈良文化財研究所長（右）と
加藤友康史料編纂所長（左）

今回連携を進めるのは、史料編纂所が公開する『電子くずし字字典データベース』（研究代表 久留島典子教授）と奈良文化財研究所の公開する『木簡画像データベース・木簡字典』（研究代表 渡辺晃弘都城発掘調査部史料研究室長）となります。来る10月を目処にそれぞれのホームページから連携検索画面を公開する予定です。

両データベースは、前者が古文書・古記録、後者が出土木簡と、それぞれ異なる対象からデータを集めていますが、文字を読むツールとしてその機能に共通する部分が大きく、連携公開を目指す運びとなりました。両データベースが連携すると、あわせて約17万件の画像データから任意の文字字形を表示できるようになります。ご利用の皆様には、古代木簡から江戸時代の古文書まで、1000年余にわたる文字の変遷を一画面のうちに御覧いただくことが可能になります。

2009/10/14よりシステム連携開始

奈文研「木簡画像データベース・木簡字典」

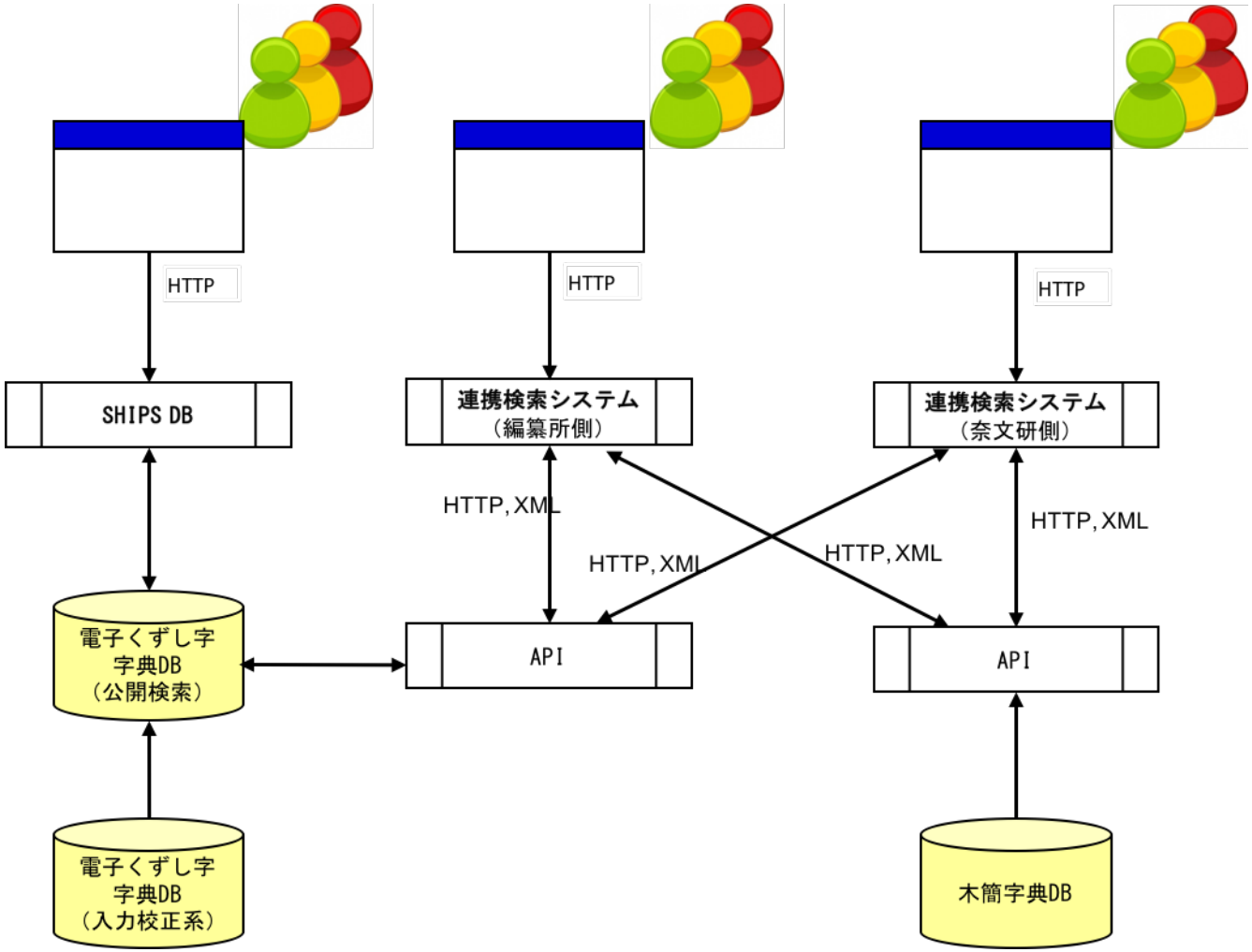
- 出土文字史料を読み解くためのツール
- 釈読のノウハウを学界として共有するためのツール

- 2003年度開発開始，2005年度公開

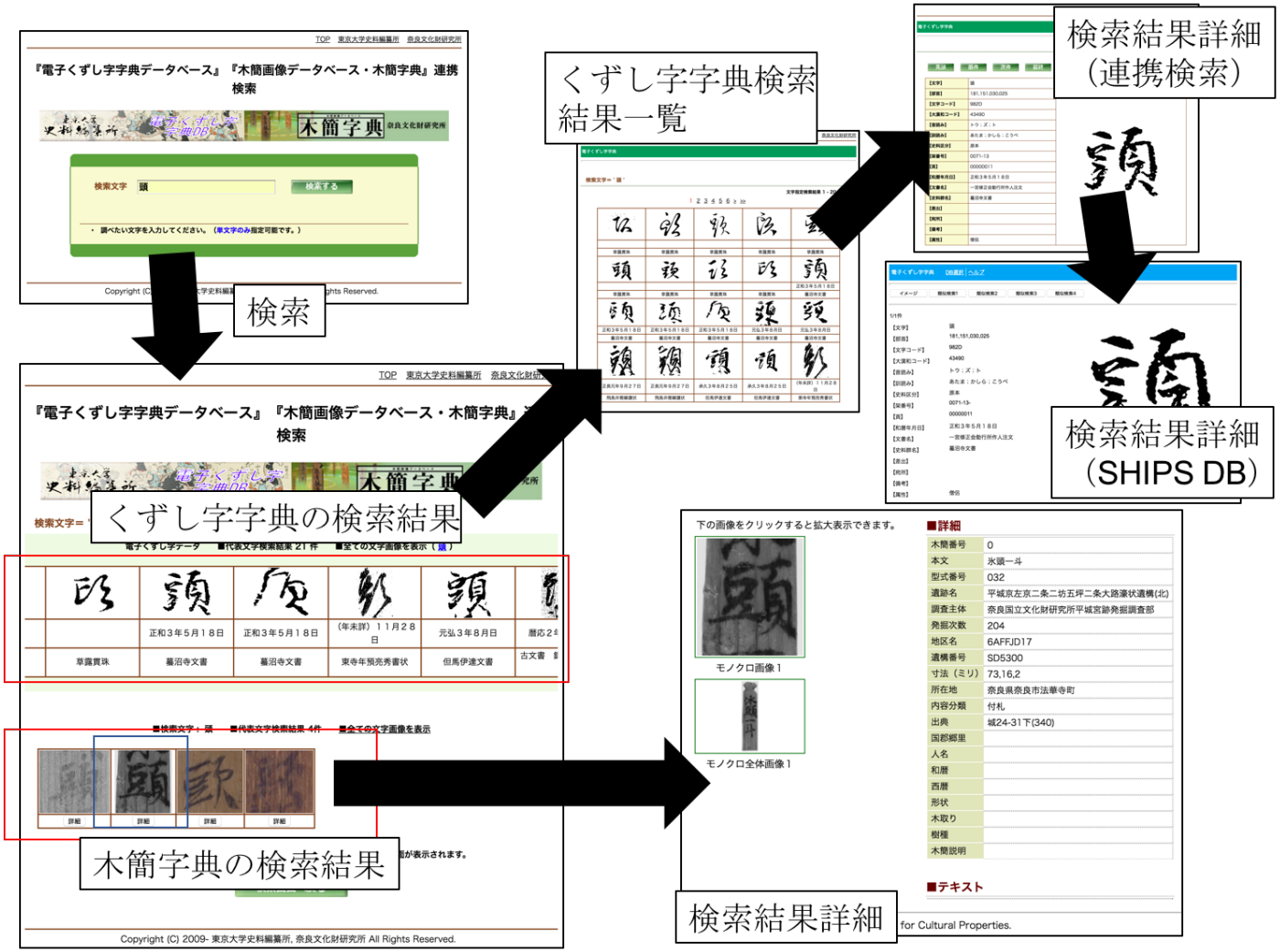
- 13,822点の木簡
- 86,347点の文字

- 木簡の出土先
  - 平城京ほか飛鳥藤原京・多賀城跡・伊場遺跡群・飯塚遺跡・大宰府跡など

システム


HTTP
HTTP
HTTP
SHIPS DB
連携検索システム
（編纂所側）
連携検索システム
（奈文研側）
HTTP, XML
HTTP, XML
HTTP, XML
HTTP, XML
電子くずし字
字典DB
（公開検索）
API
API
電子くずし字
字典DB
（入力校正系）
木簡字典DB

『電子くずし字字典データベース』『木簡画像データベース・木簡字典』連携検索

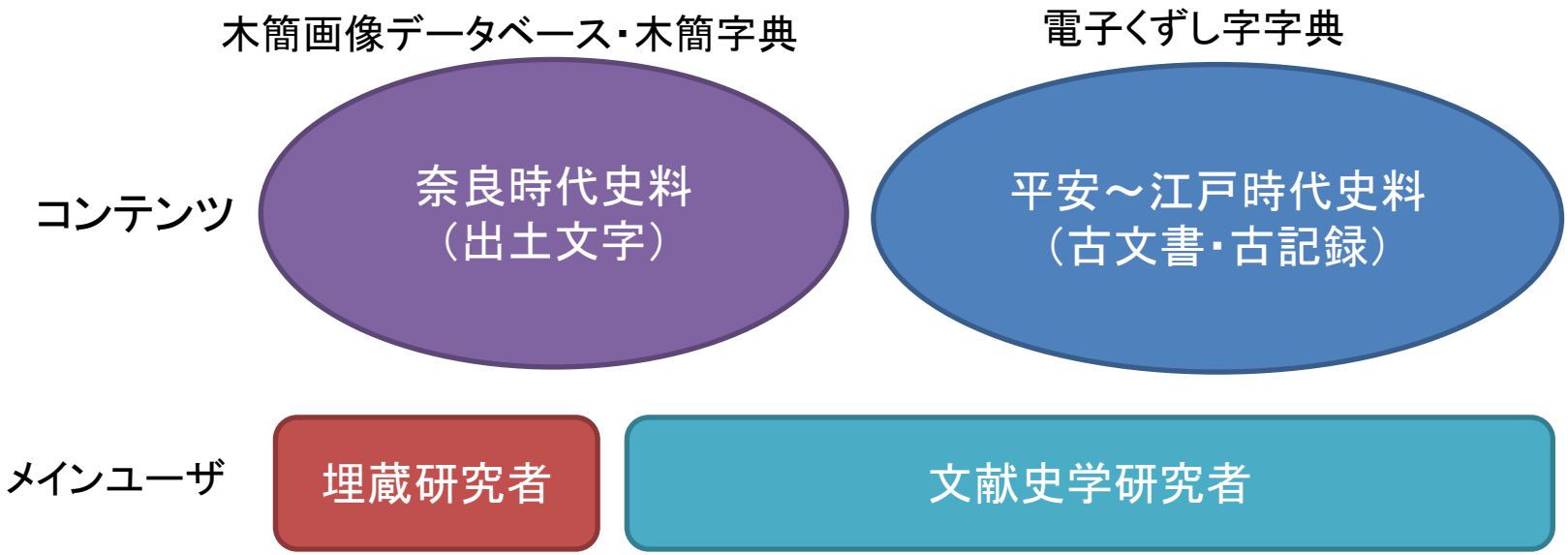


連携用のデータ定義等

No.	項目名	電子くずし字字典	備考
1	dc:title	文字（親字）	
2	dc:creator		東京大学史料編纂所
3	dc:subject	分類	
4	dc:description	音読み 訓読み 文字コード 大漢和コード 備考	
5	dc:publisher		東京大学史料編纂所
6	dc:date	和暦年月日 西暦年月日	
7	dc:type	代表フラグ	
8	dc:identifiter	管理番号	
9	dc:language		ja
10	dc:contributor	登録者	
11	dc:format		
12	dc:source	文書名 史料群名 差出 宛所	
13	dc:relation	史料区分 架 番 枝番 号 頁 画像ナンバー	
14	dc:coverage		
15	dc:rights		東京大学史料編纂所

No.	フィールド	説明
1	xml	xml定義
2	results	実行結果
3	return_code	リターンコード
4	error_parameter_list	エラーパラメーター覧
5	error_parameter	-
6	parameter_name	エラーパラメータ名
7	error_code	エラーコード
8	rdf:RDF	RDF定義
9	search_total_count	検索結果総件数
10	display_count	代表表示一覧件数
11	titleLinkList	検索文字リンク情報リスト
12	titleLink	検索文字リンク情報
13	title	検索文字
14	url	URL
15	representative_list	代表表示一覧
16	rdf:Description	
17	dc:title	文字(親字)
18	dc:date	和暦年月日
19	dc:source	史料群名
20	dc:relation	文字画像ファイルURL
21	dc:identifier	識別子

コンテンツとユーザ



ノウハウの共有、新たな方法論にむけた協同

展望（1）

- データの追加
  - コツコツとデータを作成していく
  - 対象史料を拡大していく
    - 正倉院文書（奈良時代）：これだけで約1万件
  - 連携先の増大
    - 奈文研以外も検討していました…
    - 文字データへパーマリンクを付与していく

展望(2)

- 積極的に利用してもらうために…
  - 文字を検索しやすくする
    - 文字画像を用いる：MOJIZO
    - 文字を用いる：文字n-gramを用いた文字検索
      - そのうち実装
  - オープンデータに向けて
    - 史料や史料画像のオープンデータ化は結構たいへん
    - しかし，文字データ・文字画像は可能性が高い（かも）

おわり