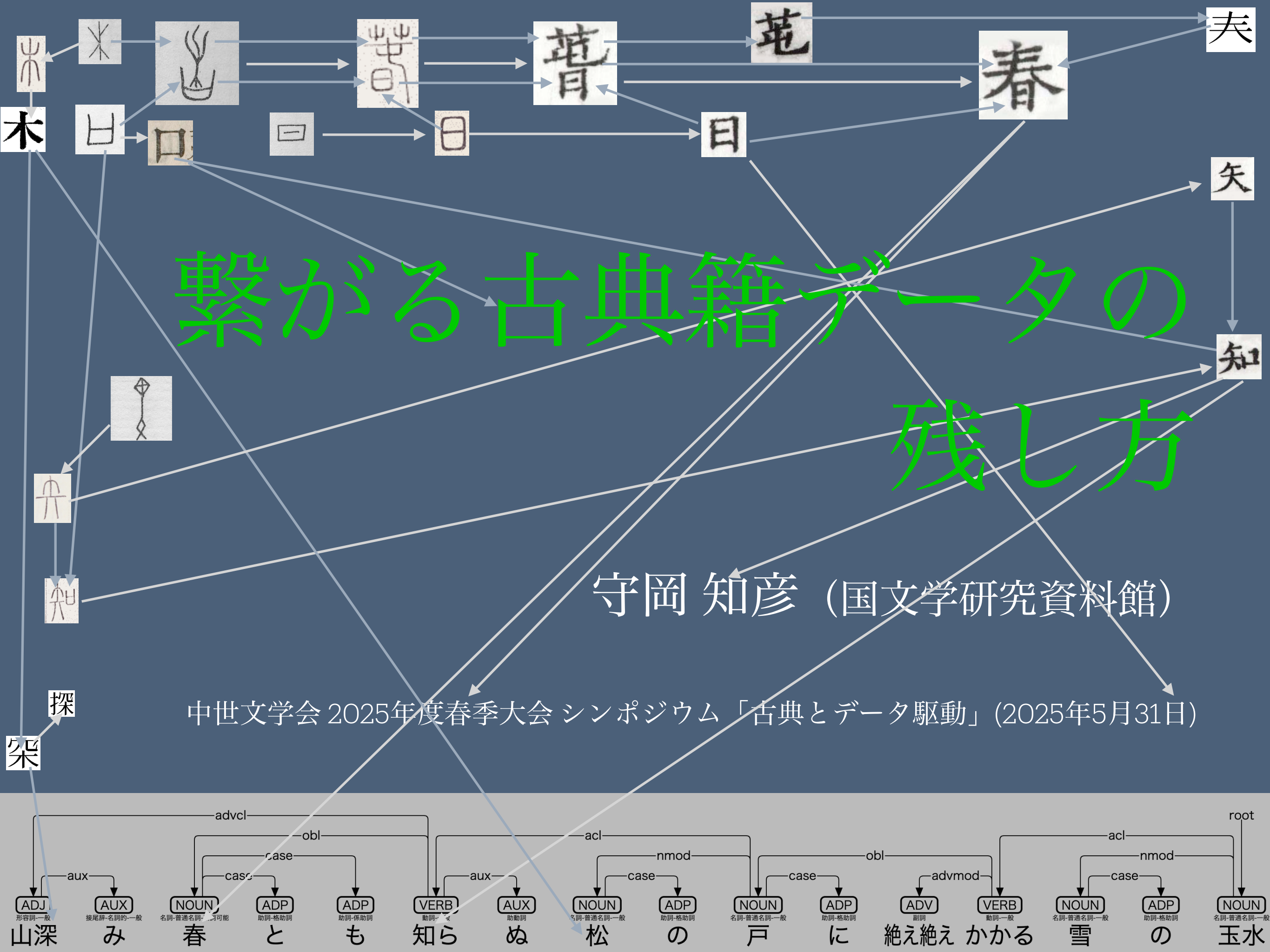




自己紹介

守岡知彦

- ・ 1999年頃から漢字情報処理に関する研究を行なっている
 - ・ CHISE project

<https://www.chise.org>

文字に関するさまざまな知識を直接利用することで、汎用文字符号に制約されない次世代文字処理環境の実現を目指すオープンソース型研究・開発プロジェクト

CHISE IDS 漢字検索：部品で漢字を探す



<https://www.chise.org/ids-find>



<https://www.chise.org/ipns/ids-find.chise.org/index.ja.html>

(Qme94UnoCa5fXGUNFnVByyf9sWdbSpbUQDhVsjHLvds4DA)

時 時

結合：

- + U+E0100 : 時

画数：6

漢字構造：  日 寺

Web韻圖 (廣韻) : [/時/](#)

= UCS : U+6642 (26178)

= **Big5** : 0xAEC9 (44745) - +

← 字種：**時**

→ [HDIC] 篆隸萬象名義： 是之反。是也，善也，中也。

→ [HDIC] 宋本玉篇： (反切)市之切。春夏秋冬四時也。

→ [HDIC] 新撰字鏡： (晡時)  市之之喜二反。四時也，是也，此也。  之喜反。在日部。

→ [HNG] 中国写本: 時 時 時 時 時 時 時 時 時 時 時 時

→ [HNG] 中国版本：

→ [HNG] 日本写本：

→ [HNG] 日本版本：

→ [HNG] 韓国資料：

→ [HNG] その他:

→ 甲骨文字：

→ 說文小篆：

→ 古字：𡵓

→ 古字(說文)：

→ 包摂: 吋 吋

古典中国語形態素用例：微時 時晦 舊時 昔時 時論 時政 時忌 時望 相時 逢時 歲時 時氣 卽時 平時 或時 少時 異時 異時 日時 時人 時宜

東洋学文献類目用例（題名・キーワード等）：

- 李寅生著「論宋元時期的中日文化交流及相互影響」 巴蜀書社 (@成都), 2007年3月
- 守屋 美都雄編「四時纂要—中國古農書古歲時記的新資料」, 1961年11月
- 許雪姬訪問「日治時期在「滿洲」的台灣人」中央研究院近代史研究所口述歷史叢書;79, 中央研究院近代史研究所 (@臺北), 2012年11月
- 汪伯岩(au)「第二次國內革命戰爭時期的農村革命根嚴地」, 1955年2月
- 傅衣凌(au)「明清時代商人及商業資本」, 1956年7月
- 傅衣凌(au)「明清時代的商業」, 1956年11月

<https://www.chise.org/est/view/character/時>

はじめに

- Web の夢と現実
 - 人間が読む、機械が読む、機械に読んでもらって教えてもらう
 - 便利なサービスもいつかなくなる
 - （プラットフォームの光と闇）
 - （いきなり夢のない話ですみません）
- 人文情報学の理念と...

Web サービス化した古典籍

- 良い点（悪い点）
 - 場所を取らない（サービスしてる人が負担してる）
 - どこでも見れる（サーバーやネットワークが突然トラブる）
 - 大量のデータから検索可能→『遠読』
 - 複数のデータセットをリンク可能（↓リンクが腐っていく）
- 悪い点：長期間維持するのが大変

データベースはどうして消える？

- 数年に一度ぐらいコンピューターシステムのアーキテクチャーが大きく変化する
 - 大型計算機→UNIX サーバー・ワークステーション・パソコン→スマホ・タブレット→？
 - スタンドアローン→インターネット化
 - 固定長データ→(ISO 646 ベースのローカル文字コードの)テキストデータ→Unicode
 - （ある技術の普及が別のレイヤーにも波及して変化を促す）
- いつまでも古いシステムは使い続けられない
 - セキュリティ問題、ハードウェアの故障、クライアント側のソフトウェアとの互換性、etc.
- Web 固有の問題

Web サービスはどうして消える？

- Web は URL が命（全ての情報資源は IRI (URL) で参照される）
 - Web では情報をその内容そのものではなくて、それを置く場所（URL）で管理している（場所 ID 方式; location addressing）
 - ➡ URL が変わると参照できなくなる
 - ➡ URL が指している中身が変わる可能性がある
（Linked Data や IIIF マニフェストで使われている IRI が変化するとそれを使って作っていたデータが崩壊する）
- URL を長期間維持するにはお金と労力と強い意志が必要
 - 組織が変わってもドメイン名が変わらないようにするか、組織に依存しないドメイン名を使わないといけない
 - システムリプレイスの際に URL（パス名など）が変わらないようにしないといけない
 - 汚い URL を綺麗にしたい欲望をグッと堪える（変えるなら古い URL を新しい URL にちゃんと飛ばす→古いドメイン名や URL とずっとお付き合いしないといけない）

データセット化・内容ID化

便利だけど儚いWebサービスを紙の本のようにするには？

- 計算機アーキテクチャの大きな変化の波を超えない
 - ➡データの引越しをしやすくする
 - 標準的なデータ形式の利用（ベンダーロックインされないようにする）
 - データベースシステムからデータを分離（データセット化）
 - 誰でもコピーして動かせるようにする（オープンソース、オープンデータ）
- URLを長期間維持するのは大変
 - ➡データの内容そのものでIDを構成（内容ID方式）
 - ➡IDの場所依存性が消える → 分権化 (decentralization)
 - Git, IPFS, (DOI)

データセット化

- Web サービスを安全に動かし続けるためには常に攻撃に備えバグを修正し防御の穴を改修し続ける必要がある（大変！）
- これに比べて、データベースの中身をデータセット化して保存する方が労力は少ない
- 例：「漢字字体規範史データセット」(HNG)

<https://www.hng-data.org/>

- オープンデータ
 - 誰でも自由に(再)利用・再配布可能なデータ（改変の自由の重要性：メディアコンバートやデータ形式等を変更する自由がないといずれ使えなくなる）
- オープンソース・自由ソフトウェア
- 複数人での編集・管理→Git などの版管理システムの重要性

Git

分散版管理システム：共同で編集・管理するためのインフラ

- 版管理システム：リポジトリ（ファイルの集合）の編集履歴を記録
 - いつ（どこで）誰が何をどのように変更したかを記録
 - 過去の版に戻したり、複数の版を比較したり
 - ➡変更の安全性を担保
 - 枝 (branch) を使って、異本の系統樹を表現（例：漢籍リポジトリ）
- 分散型：どこかに唯一の原本があるのではなく、各コピーリポジトリが対等で、互いに同期可能
 - GitHub が止まってもローカルで編集したり別のリポジトリに push 可能；その内容を後で GitHub に同期可能
 - 内容 ID：データの内容そのもの（のハッシュ値）を ID にすることで実現
 - ➡ID における場所依存性が消える（引越してもデータの意味が不変）

Git の利用例：

- 漢籍リポジトリ <https://github.com/kanripo>
- 平安時代漢字字書総合データベース <https://github.com/shikedada/HDIC>
- 漢字字体規範史データセット <https://gitlab.hng-data.org/HNG>
- 資料横断的な漢字音・漢語音データベース <https://github.com/daikaku/DHSJR>
- CHISE <https://gitlab.chise.org/CHISE>
- CBETA <https://github.com/cbeta-git>
- 古典中国語形態素コーパス <https://corpus.kanji.zinbun.kyoto-u.ac.jp/gitlab/Kanbun/>
- などなど

国書データセット（仮称）

国書データベースのデータセット化の試み

- 国文研所蔵（寄託分を除く）のパブリックドメインで公開されている画像をNDL古典籍OCR ver.3で処理したテキストと座標データを Git 化
- 書誌ID(BID)版：<https://gitlab.nijl.ac.jp/Kokusho/>
 - 国書データベースの書誌頁 <https://kokusho.nijl.ac.jp/biblio/<BID>> のテキストデータを https://gitlab.nijl.ac.jp/Kokusho/<BID>_text に格納
 - 現在、約25000件を収録しパブリックドメインで公開
- 著作ID(WID)版：<https://gitlab.nijl.ac.jp/Kokusho-Works/>
 - 国書データベースの著作頁 <https://kokusho.nijl.ac.jp/work/<WID>> のテキストデータを https://gitlab.nijl.ac.jp/Kokusho/W-<WID;10進8桁>_text に格納
(BID 版を Git submodule として収録)

CLI 環境での使い方

(Mac なら Homebrew, Windows なら WSL を入れて、Git を入れる)

- 例：新古今和歌集 <https://kokusho.nijl.ac.jp/work/4499>

```
% git clone https://gitlab.nijl.ac.jp/Kokusho-Works/W-00004499_text.git --recursive  
% cd W-00004499_text  
% git submodule update --remote
```

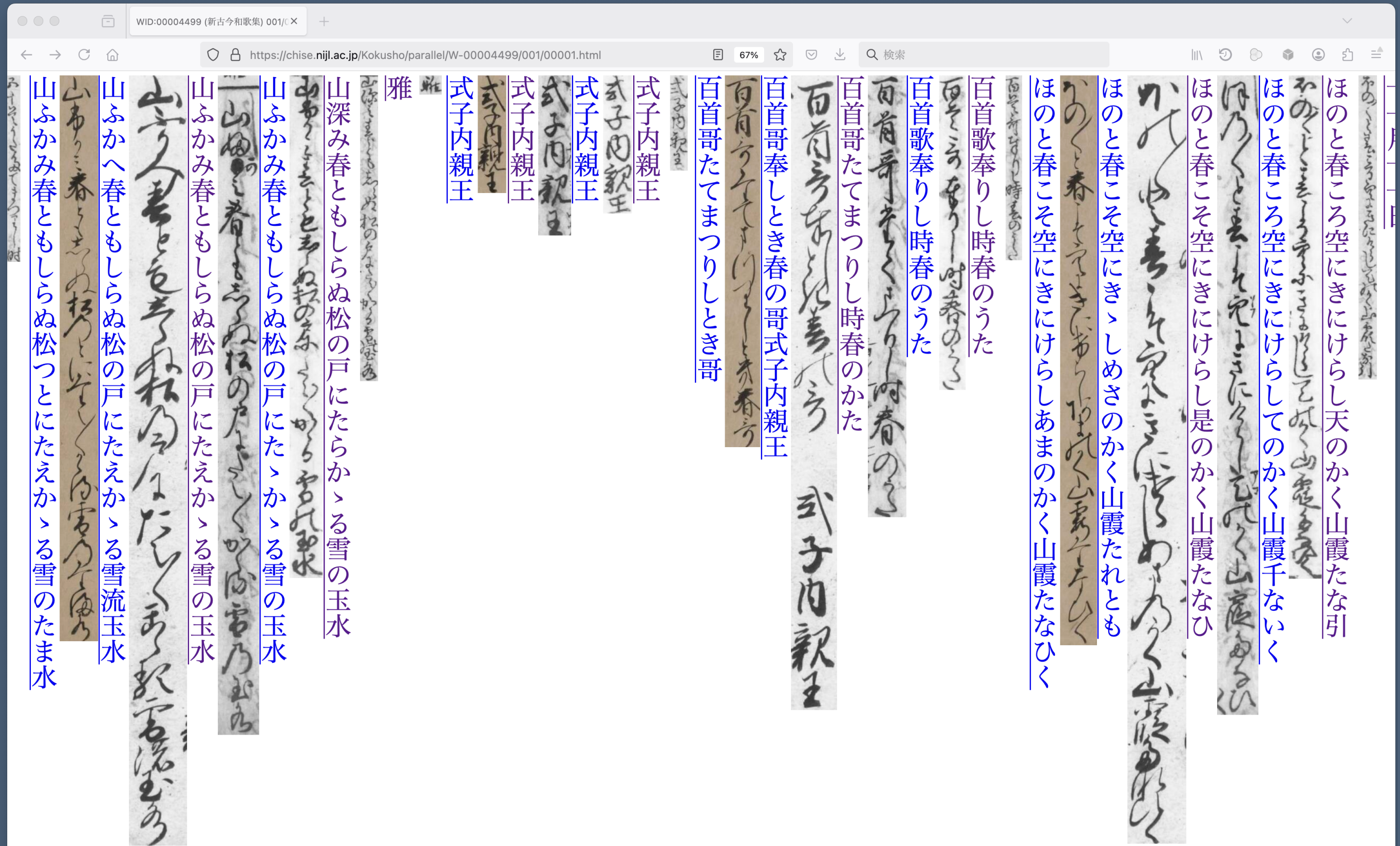
これで、W-00004499_text/ の中に83文献のテキストファイルが入る

(以降、W-00004499_text の中で3行目を実行すれば最新版に更新される)

```
% diff -u 200000176_text/all.txt 200000201_text/all.txt | less
```

などとすれば、BID:200000176 と BID:200000201 のテキストが比較できる

Web での比較のイメージ



Linked Data

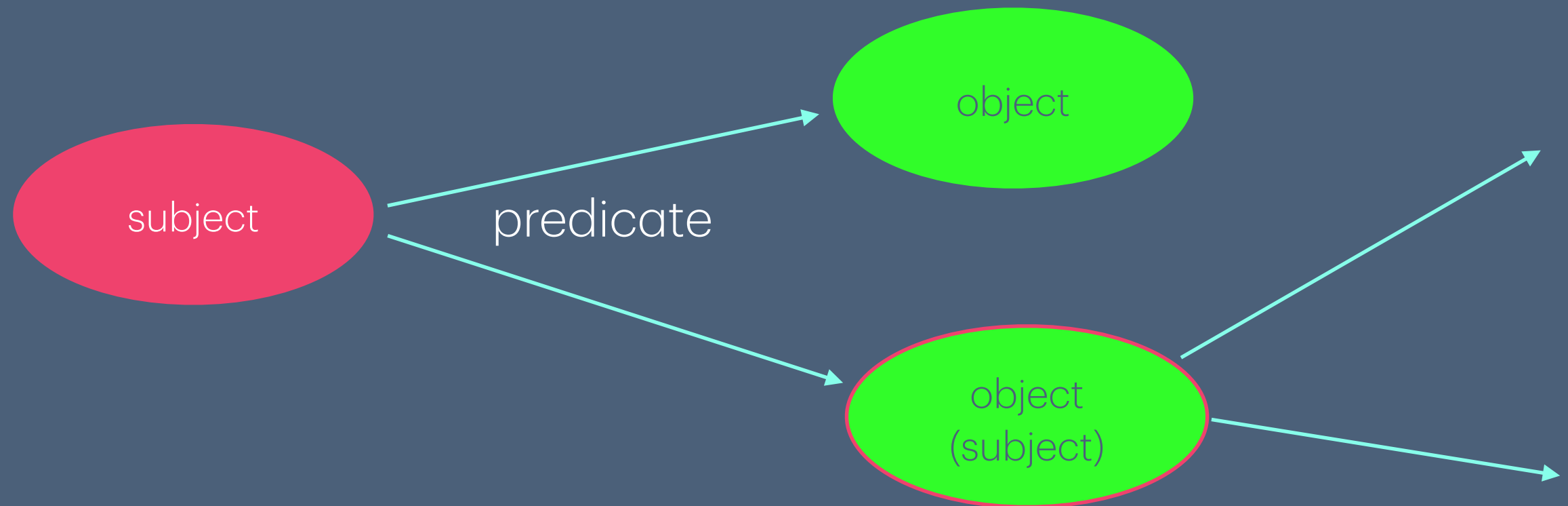
インターネット全体を巨大なデータベースに

- RDF (Resource Description Framework) : インターネット上の情報資源に関する意味論的記述のための標準となるデータモデル
 - XML, JSON, Turtle 等の複数のシリアル化形式
 - IIIF では JSON でシリアル化した RDF (JSON-LD) を採用
- Linked Data: RDF でデータセットを記述する技術
 - 複数のデータセットを繋ぐことができる
- Linked Open Data: Linked Data + オープンデータ
- セマンティック Web : Linked Data でオントロジー（概念記述；対象物の存在論）を書いて（数理論理的な）推論可能な Web を構成する技術

RDF (Resource Description Framework)

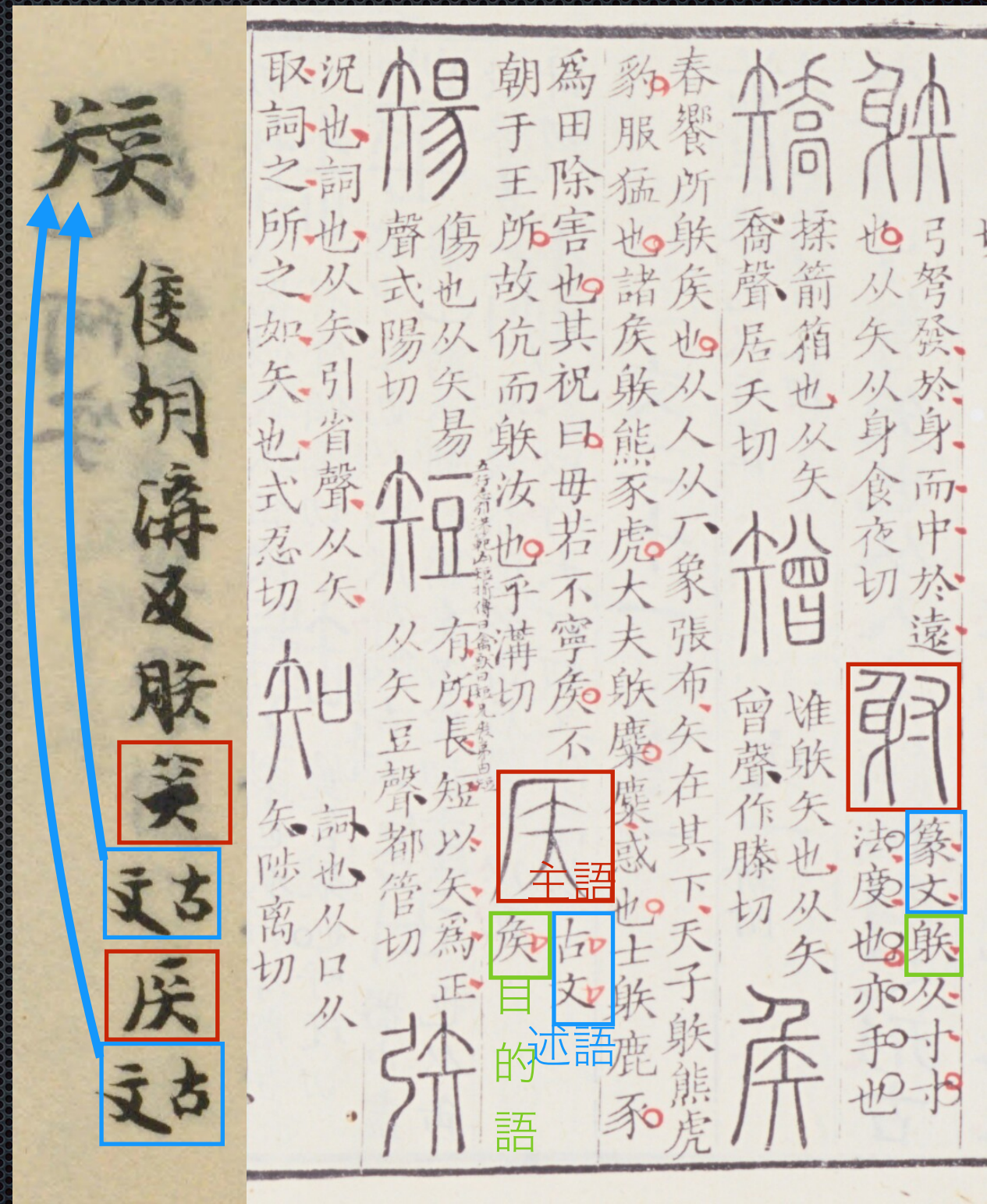
- 主語 (subject)
- 述語 (predicate)
- 目的語 (object)

の3つ組（これを「トリプル」という）で情報を記述
それぞれ、IRI で表現



古字書の情報

- 引用、注釈、コピー、etc.
- アノテーション、マークアップ、...
- Linked Data なのだが
 - 掲出字を主語とし、形・音・義などについて述べている



古字書の記述の Linked Data 化

▪ ancient

𪛗



壙

宋本玉篇：

𪛗：... 古文壙、亦作𪛗

▪ formed

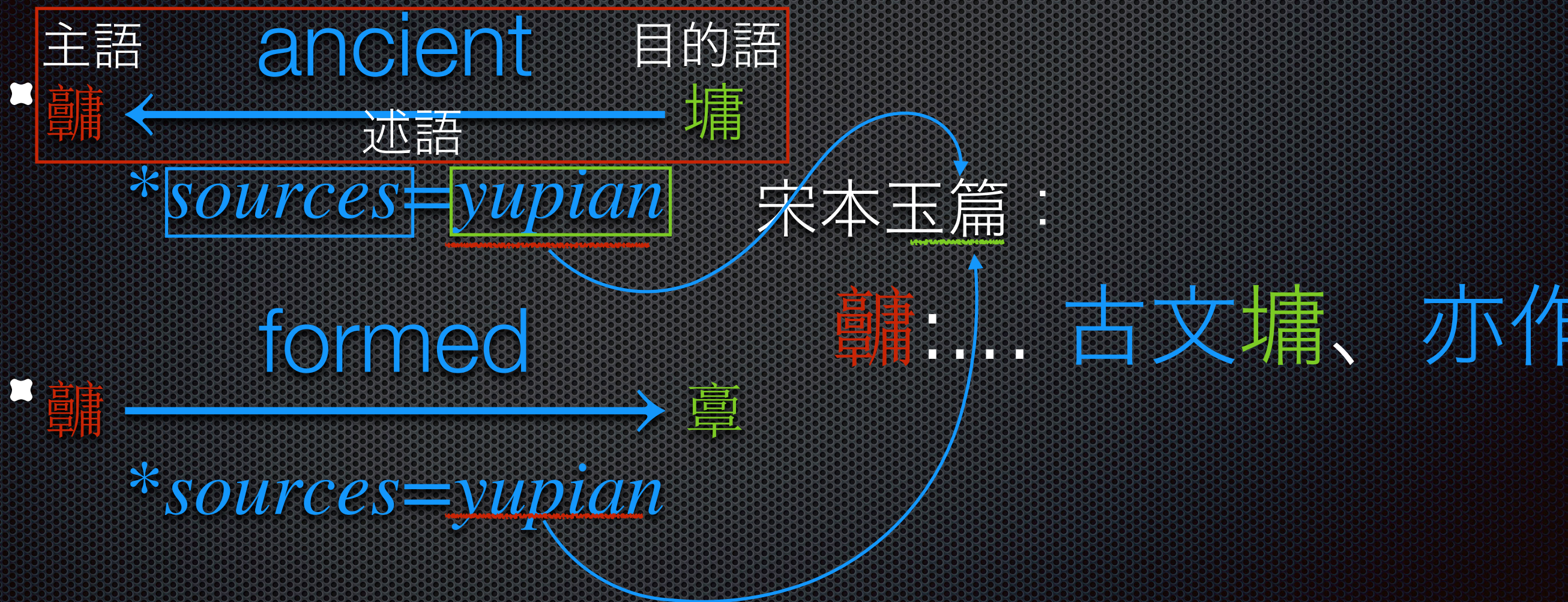
𪛗



𪛗

- CHISE ではこれらの知識を文字オントロジー(データベース)に蓄積

典拠情報のLinked Data化



古典語や少数言語と(いまどきの)AI

- 大規模言語モデル(Large Language Model; LLM)は大量のコーパスを必要とする (数兆~10兆語?)
- 話者数の多い現代の言語の場合、インターネット上のさまざまなテキストが利用できるが、古典語や少数言語の場合は大変
 - 例えば、四庫全書でもたった10億字しかない(!)
- ➡他の言語の知識を流用する必要がある

Universal Dependencies (UD)

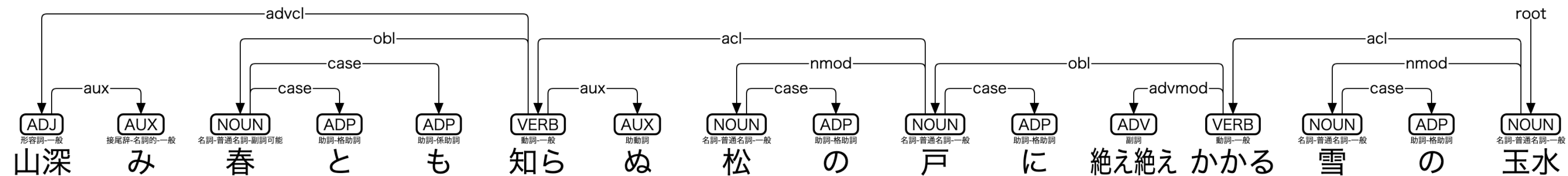
<https://universaldependencies.org/>

- 依存文法的一种を使って、単語間の依存関係を記述
- 世界のさまざまな言語のコーパス（ツリーバンク）を構築する国際的な共同プロジェクト
- 約半年毎にデータ論文（紀要）として出版
- 最新版は UD Version 2.16 (May 15, 2025)

<http://hdl.handle.net/11234/1-5901> 179言語のツリーバンクを収録

- 京大人文研コーパス共同研究班が開発した古典中国語のコーパスも
 - https://universaldependencies.org/treebanks/lzh_kyoto/index.html
 - 最近は変体漢文にも挑戦中（例：[日本書紀](#) [日本靈異記](#)；未収録）

新古今和歌集の UD 化の試み



Git リポジトリ : <https://gitlab.nijl.ac.jp/UD/ud-shinkokin>

- 京大人文研の安岡さんが作った古典中国語用 UD エディターを改造して

<https://chise.nijl.ac.jp/UD-Kokusho/conllusvg/ShinKokin.html>

に Web UI を設置

- Git リポジトリの内容を可視化
- 国文研 GitLab にアカウントがあれば、Web ブラウザーで動く編集画面から上記 Git リポジトリに push 可能
- 原文と現代語訳の統語構造が比較できる

古文UDの作り方

1. とりあえずUD用の自動解析器にかける

- 例えば、SuPar-UniDic

<https://github.com/KoichiYasuoka/SuPar-UniDic>

- 国語研が開発した形態素コーパス UniDic が利用可能性
- 和歌の場合、中古和文UniDic（平安時代の仮名文学作品用）

https://clrd.ninjal.ac.jp/unidic/download_all.html#unidic_wabun

があってるっぽい（他にも「中世口語UniDic」（室町時代の口語資料用）、「中世文語UniDic」（鎌倉時代の文語文用）なども利用可能）

2. 結果を直す

3. 疑問点を議論してガイドラインに記載し 2. へ

おわりに

- データセット化・オープンデータ化・Git化（内容ID化≡分権化）
 - 古典籍データに紙の本のような性質を持たせることで寿命を伸ばす
（プラットフォームに依存しないことの重要性；「学問の自由」の技術的保証）
- データを繋ぐための技術
 - Linked (Open) Data: 機械可読な Web の世界
 - （IPLD: 内容IDベースの Linked Data（Bluesky の基礎技術の一つ））
 - Universal Dependencies (UD) 汎言語的な依存文法コーパス
 - 言語類型論研究のためのツールだが、コーパスの量が少ない古典語や少数言語の自然言語処理に、巨大コーパスを持つ言語の知識（モデル、ツール等）を流用可能にするための基盤になる
- 「方法論の共有地」(Methodological Commons)
 - デジタル人文学の理念を示す標語
 - 人文学における「知の雲」(Clouds of knowing)と特定の研究プロジェクトの
間で共有された技術を媒介としての人文情報学
➡サイロに閉じこもらないことの意義