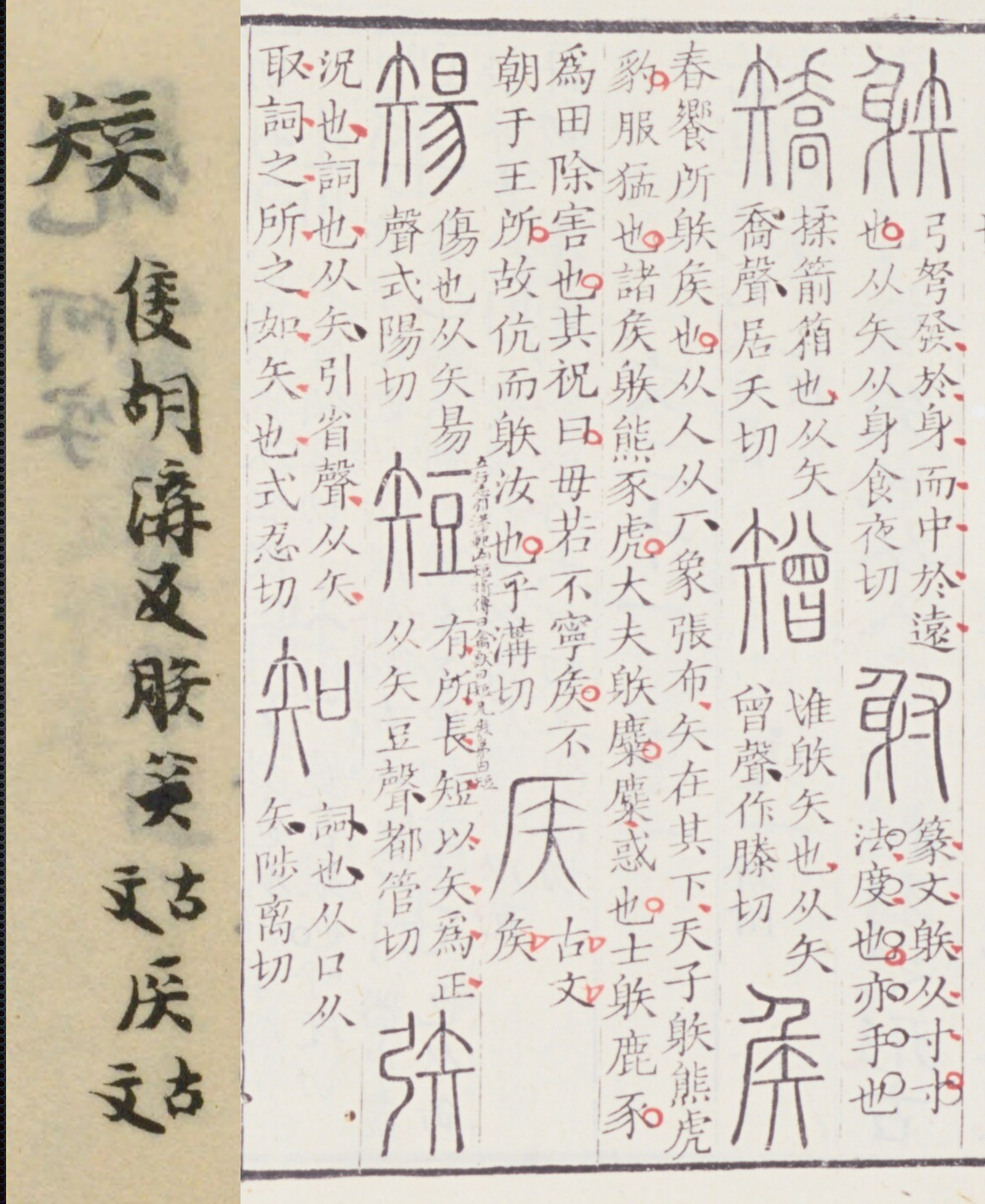
文字オントロジーにおけるマークアップに関する試論

守岡 知彦（国文学研究資料館）

東洋学へのコンピュータ利用 第36回研究セミナー (2023-07-28)

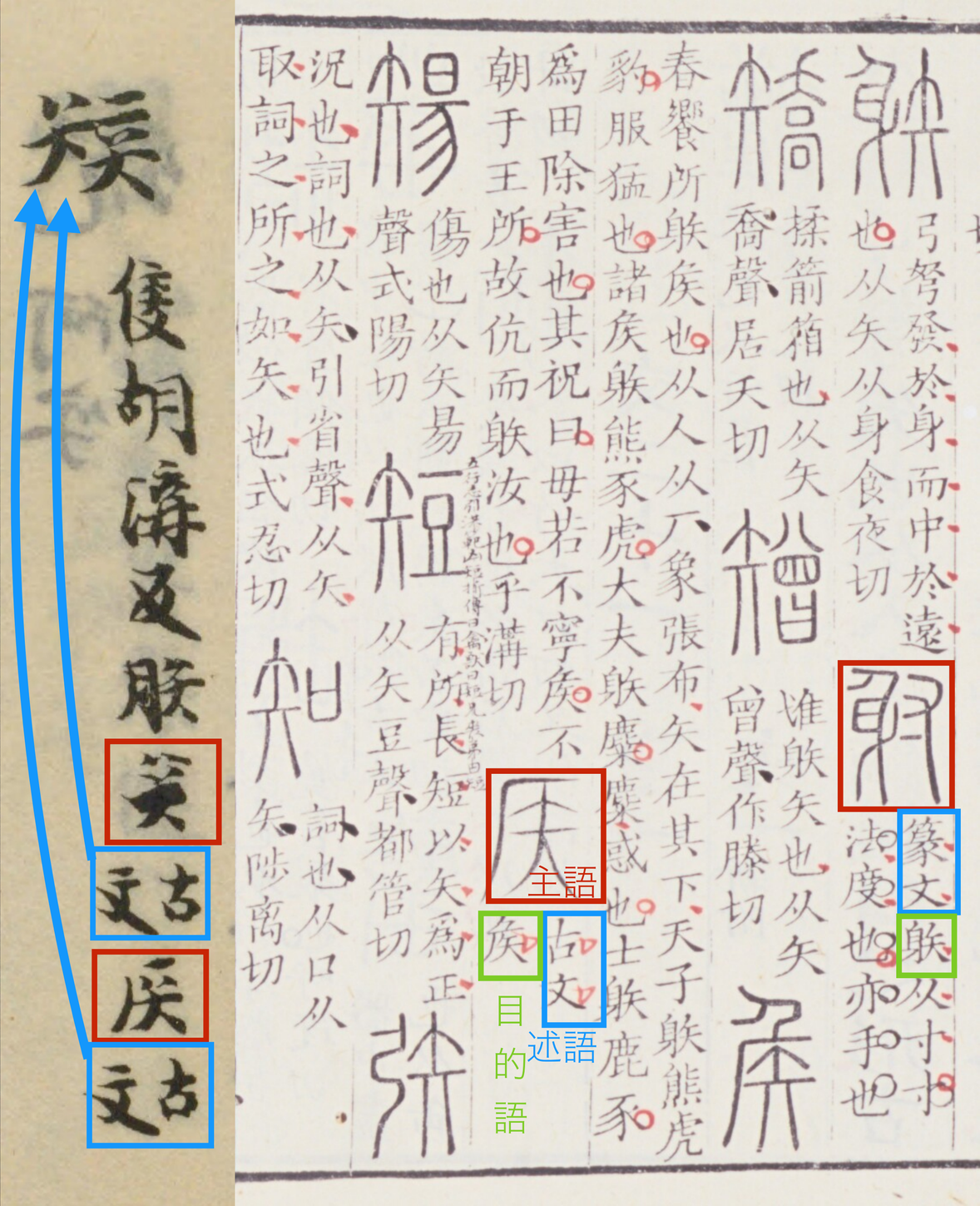
はじめに

- 文字データは典拠が命
  - 原本画像・テキスト公開の広がり
    - これらをリンクしたい
  - アノテーション、マークアップ、…
    - 本質的に Linked Data なのだが



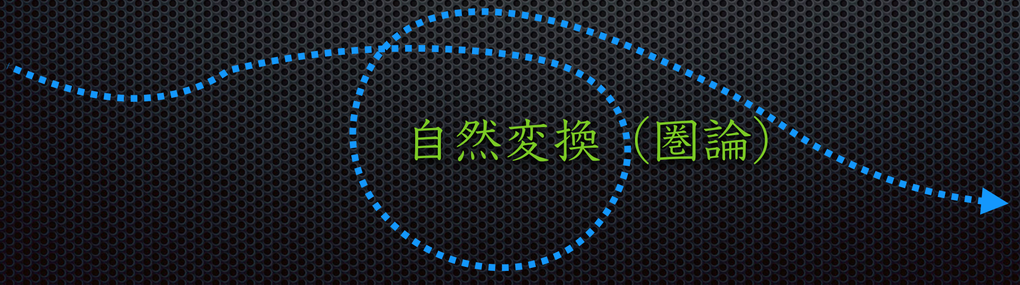
字書の中の記述

- 引用、注釈、コピペ、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



マークアップの形式と意味

- プレインテキスト（文字コード列）を所与の存在として仮定
  - １次元配列→木構造→指示対象（もっと複雑な構造?）
- アノテーションという観点では叙述対象の指示方法は（IRI のような ID が付与できれば）任意（メディアタイプごとに決まるような何か）
  - 通常、有向非循環グラフ(DAG)を想定
- 同じものを記述しているならば同様な情報を持っていて互いに対応しているはず？



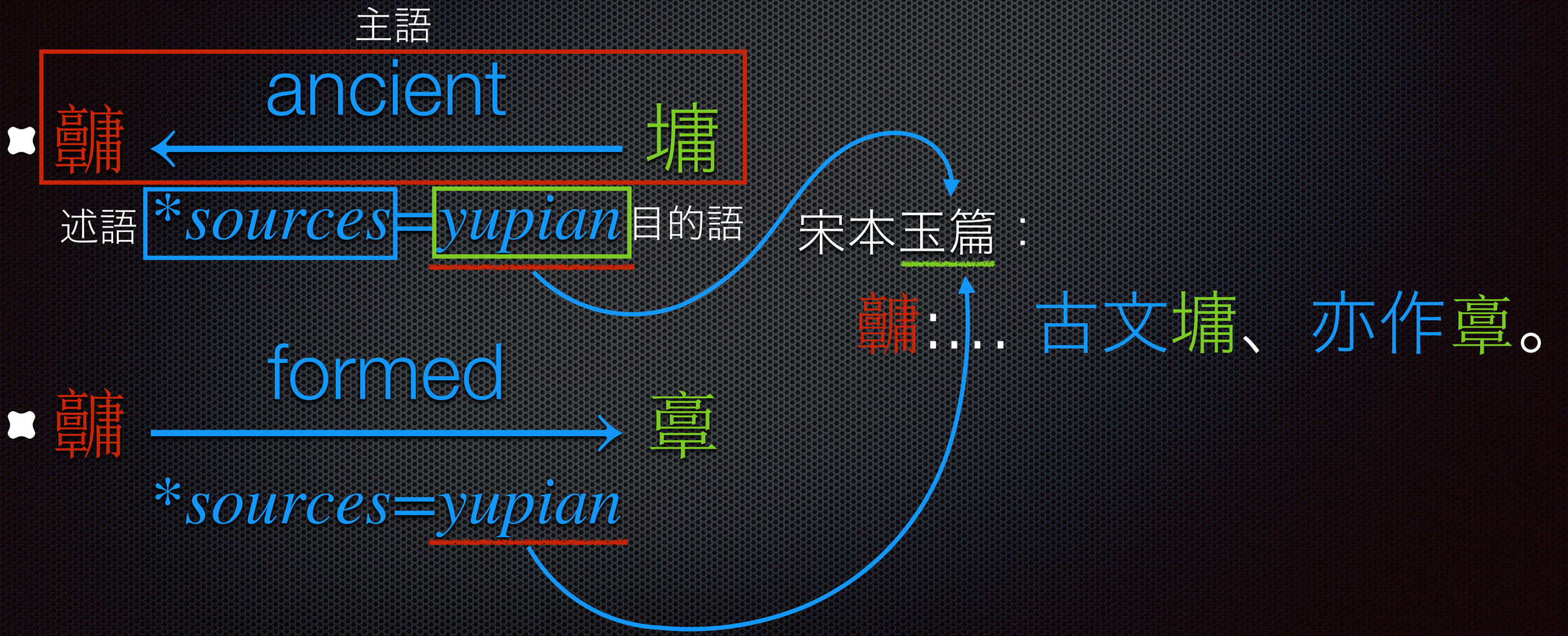
文字オントロジーでの記述と典拠

- 𩫖 ←(ancient)— 墉
- 𩫖 —(formed)→ 墉

宋本玉篇：

𩫖:… 古文墉、亦作墉。

メタデータ素性


主語
ancient
嚞庸
墉
述語
*sources
yupian
目的語
宋本玉篇：
嚞庸：……古文墉、亦作亶。
formed
嚞庸
亶
*sources=yupian

典拠文献の名前解決(1)

- さまざまな異本・版の問題
  - 内容や文字の異同、体裁の差などなど
    - 同じ版本でもサイト毎に異なるデジタル化
      - 特定の公開画像に特化（URI依存）すると簡単だが別のバージョンで利用できない（比較できない）

全文画像やテキストの公開が進んできたが故に…

典拠文献の名前解決(2)

- 引用箇所の探索
  - 文献の中の実際の引用箇所を表示したい
    - 現行の CHISE のメタデータ素性には通常典拠文献 ID (e.g. yupian) しか入っていない
      - ➡ そのままでは実際の引用箇所にたどりつけない

典拠文献の名前解決(3)

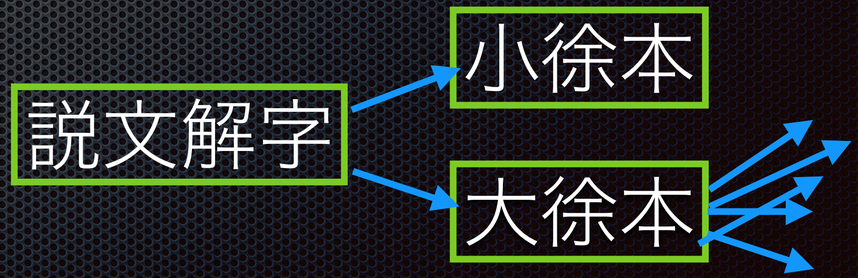
- 字書の特性
  - 引用箇所は掲出字のことが多い
  - 異体字関係の記述の場合、掲出字かその異体字の可能性が高い
    - 但し、字体にはバリエーションがあり、対応する別字体を引用することもあり（他の文献との合わせ技（比較・推論等）も）
  - なので、掲出字の場所がわかれば大体なんとかなる

探したい文字ほど見つからない問題

- ctext.org や漢リポなどの全文テキストを使えば探せる（はず）
  - しかし、見たい掲出字ほどOCRの誤認識や欠落、外字等になってることが多くて探せないことが多い
    - 前後のテキスト（異体字関係の記述や音注等）を使って探す

典拠文献の名前解決(4)

- いろんな文献が出てきてしまう問題
  - （何でも良いから!?）代表を決める必要がある
  - 抽象的な文献と具体的な版（サイト）の継承関係
    - FRBR モデルの Work / Expression 等
    - CHISEの階層的ドメイン方式
  - テキスト内での抽象的な場所の表現



マークアップ再考

- TEI の <app>, <lem>, <rdg> タグなどを用いて異なるテキストを一つにまとめてその差分を表示するという行為は対応する異本の抽象↔具象関係の記述を行なっている
  - 但し、通常のインラインマークアップでは『抽象（的な記述）』を底本等を用いて具体的に書く
  - 文字の一次元配列（テキストストリーム）に基づいて書く
    - でも、多分、アノテーションという観点では本質的ではない
    - （本発表での観点における）本当に必要だったもの：
      典拠リンクの名前解決のための情報記述

典拠リンクの名前解決に必要な情報

- 字書の場合、最低限、掲出字の位置情報があればなんとかなりそう
  - リンク先の種類の情報：リンク先は主語か目的語か複合型か?
  - 本文の情報があるとうれしい
- 文献の名前、別名、ID 等
- 文献の包摂（抽象↔具象）関係
- 文献間の関係（注釈、引用等）
- 文献の属性（作者、種類等）

HDIC はこれに近い
情報・構造を持っている
ような気がする

インラインマークアップの合成

- 文字オントロジー内の記述（あるいはHDIC的なもの）を十分に精緻にしていけばそこからインラインマークアップを合成することができるはず
  - （スタンドオフマークアップ）
  - エンティティーをどう表現・名前解決するか
    - 包摂関係
    - 複数のID体系の併用
    - などなど

おわりに

- 典拠リンクの名前解決のための情報としてのマークアップテキスト
  - 一次元の文字列ではなく有向非循環グラフ(DAG; e.g. RDF, IPLD)を基礎に実体間の関係を記述（プレーンテキスト中の位置関係に依存しない；必要ならIDを振り、位置関係依存性はコンテナで表現）
- 字書の場合、掲出字の情報があればだいたい OK（だが、現状、不十分。補完するには結局本文が要る）
- 抽象↔具象関係や文献間の関係を考慮したID体系や文献オントロジーの必要性