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

1　はじめに

漢字は 1 字が古典中国語の 1 形態素に対応する表語文字と考えることができ、伝統的に形音義にかかわる諸要素の組合せや対応関係によって分析・整理されてきた。こうした漢字に関する知識は説文解字や玉篇、康熙字典といった近代以前に編纂された古字書や各種注釈書等で前の時代に成立した文献や注釈書等への引用・注釈という形で積み重なってきたといえ、このことは近代以降の漢字辞書や JIS X 0208 や ISO/IEC 10646 = Unicode のような漢字を収録する汎用的な文字符号においても同様であり、こうした知識を電子化した文字データベースや文字オントロジー等においては漢字に関する諸属性の典拠情報をどのように記述するかは重要な問題であるといえる。特に、昨今、さまざまな文献の原本画像や翻刻テキストのインターネット上での公開が進展しており、電子化された漢字辞書や文字オントロジーにおける記述の典拠として元テキストを紐づけることで引用元のテキストをすぐに確認できるようにしたり、古字書や現代の漢字辞書の引用関係を把握できるようにすることが技術的に可能になりつつあり、その有用性からさまざまな試みが行われている。

しかしながら、こうしたリンクの多くは特定のシステムやテキスト・版に依存したアドホックな形で行われており、現在・将来に進みつつある多数のバリエーションのテキストが複数のサイトで公開されていく現状を鑑みれば、Linked Open Data の世界を前提に、引用や引用箇所の記述やその名前解決の問題について整理しておく必要があると思われる。本稿では CHISE 文字オントロジーにおける典拠情報の記述や HDIC の古字書データの取り込みを例にこの問題について検討したい。

2　古字書のテキストとその指示対象の関係の記述

異体字関係の記述は前近代の漢字字書においても見られ、[5] 干禄字書のように異体字の整理・弁別に特化した字様書と呼ばれるジャンルの字書も編纂されている。[3] 字様書は複数の異体字を列挙してその種別（正、俗、通、等）を記述した表のようなものと捉えることができるが、説文解字や玉篇等では自然言語（古典中国語）で記述された文字の説明の中で異体字関係の記述も行われている。

例えば、説文解字では「射」に関して、

 (躰): 弓弩發於身而中於遠也。从矢从身。〔食夜切〕

 (射): 篆文躰从寸。寸、法度也。亦手也。

のように、「躰」を見出し字として挙げ、「射」をその異体字として挙げ、「射」が「躰」の篆文であり、「射」が（「矢」ではなく）「寸」に従っていて、その差分となる部品「寸」について解説している。これらの記述のうち、異体字関係に関る部分は「  (射): 篆文躰」であるが、これは

 (射) $\xleftarrow{\text{篆文}}$  (躰)　(1)

という関係を示していると考えられる。即ち、「  (射): 篆文躰」という文は式 (1) に示した RDF のトリプルを表現しているといえる。ここで「  (射)」はこのトリプルの主語で、「篆文」は述語、「躰」は目的語を表現しているといえる。

このように、古字書の[1)]テキストにおける領域（部分文字列）の世界と RDF のトリプルで表現されるような文字に関する意味グラフ（文字オントロジー）の世界の対応関係を考えることができ、テキストの世界の領域に対して文字オントロジーの世界の要素（文字オブジェクト（主語･目的語）、関係やその他文字素性（述語）、その他属性値（目的語））との対応関係を注記することをマークアップという。実際には、マークアップは RDF のトリプルで表現されるような Linked Data との対応関係を示すものに限られずテキスト構造化一般に用いられるものであるが、本稿では Linked Data の世界との適切な対応関係を表現可能なマークアップテキストの用件について考察したいので、テキスト世界の圏と文字オントロジーの世界の圏が自然変換をなすようなマークアップを主な対象とすることにする。このような観点でのマークアップテキストにおいては、Linked Data の世界の実体 (entity) に対応する部分文字列が過不足なく単一の部分木になり、マークアップされ

1)　とは限らないが

た部分文字列もしくはマークアップテキストの構文木における部分木とそれが表現する Linked Data 上の実体（意味グラフ上のノード) がきちんと対応することが肝要である。

3　引用文献の表現

前節とは逆に文字オントロジーの側からその記述の典拠となる文献へのリンクの問題について考えてみる。例えば、

$$\text{鼞} \xleftarrow{\text{ancient}} \text{墉} \tag{2}$$

$$\text{鼞} \xrightarrow{\text{formed}} \text{亹} \tag{3}$$

という２組の異体字関係が存在し、その出典として HDIC の宋本玉篇の

鼞：餘恭切。古文墉、亦作亹。

というものがあるとする。但し、$A \xleftarrow{\text{ancient}} B$ は「A は B の古文」、$A \xrightarrow{\text{formed}} B$ は「B は A の異体字（「A はまた B にも作る」)」という関係を表すもの（関係素性）とする。

CHISE ではこのような場合にメタデータ素性 [7] というものを用いて関係素性に対してメタデータを付与することができ、出典情報示すために `*sources` というメタデータ素性を用いている。例えば「鼞」は、CHISE の S 式 (Lisp) 表現 [1] では

```
(define-char
  '((ideographic-radical . 189) ; 高
    (ideographic-strokes . 17)
    (total-strokes       . 27)
    (ideographic-structure ?⿰ ?亹 ?庸)
    (=ucs                . #x29AF1)      ; 鼞
    (<-ancient           ?墉)
    (<-ancient$_1*sources yupian@hdic-syp)
    (->formed            ?亹)
    (->formed$_1*sources yupian@hdic-syp)
    ))
```

のように表現することができる。ここで、「鼞」(U+29AF1) の素性対

```
(<-ancient ?墉)
```

は素性名が `<-ancient` でその値が (?墉) という文字のリストであり、式 (2) のトリプルを表現している。また、素性対

```
(<-ancient$_1*sources yupian@hdic-syp)
```

は素性名が <-ancient$_1*sources で、これは素性 <-ancient の最初の値に対するメタデータ素性 *sources であることを示している。

メタデータ素性 *sources の値には引用文献を示すシンボル（文献 ID）のリストが格納される。例えば、玉篇に対しては yupian というシンボルが割り振られており、これに対応する bibliography ジャンルのオブジェクト（文献オブジェクト）

https://www.chise.org/est/view/bibliography/rep.chise-bib-id=yupian

が設けられている（図 1）。

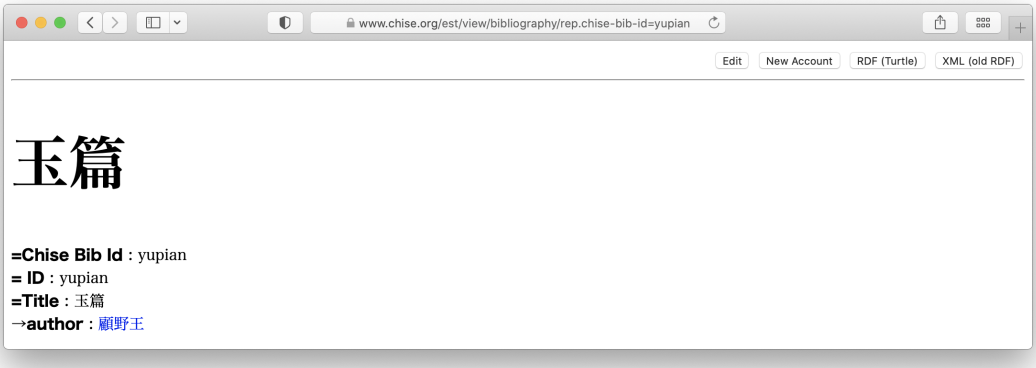

図 1　玉篇の文献オブジェクト

文献 ID においても CHISE の階層的素性名が使えるため、前述の例における yupian@hdic-syp という文献 ID はドメイン @hdic-syp における玉篇と解釈でき、全体として HDIC に収録された宋本玉篇を示す。

CHISE の文献オブジェクトに割り当てられるシンボルを表現するために ID 素性 =chise-bib-id が設けられており、文献を表現するための別の ID 体系に対して別の ID 素性を振ることで複数の体系間の橋渡しが可能なようになっている。また、関係素性 ->author を用いることにより creator ジャンルのオブジェクト（作者オブジェクト）へのリンクを付与することも可能である。例えば、玉篇の文献オブジェクト（図 1）には作者を示すための関係素性 ->author 素性があり、これは作者オブジェクト（図 2）

https://www.chise.org/est/view/creator/rep.id=u9867u91CEu738B

に対するリンクの役割を果たしている。この作者オブジェクトの =name 素性には「顧野王」という文字列が格納されており、文献の作者名を得ることができる。作者オブジェクトも素性名を割り当てることで別の ID 体系や別名に対応できるようにしている。

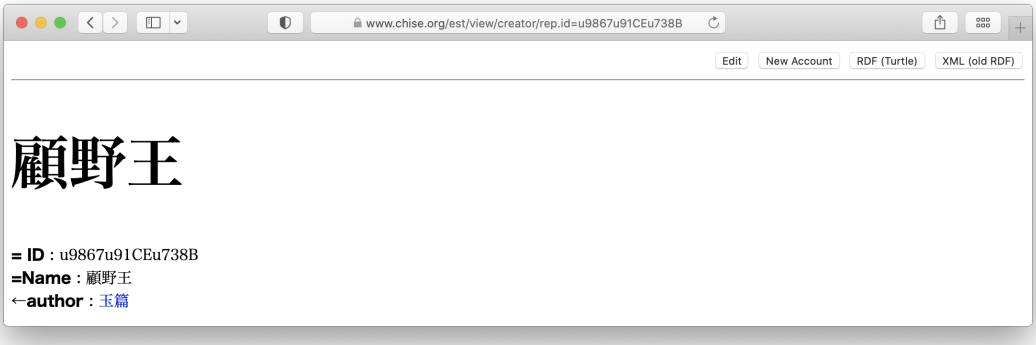

図 2　顧野王の作者オブジェクト

但し、この従来の CHISE の出典記述の仕組みは最低限のものといえ、元テキストの該当個所へのリンクを実現するためには不十分なものといわざるを得ない。例えば、

$$\text{鼆} \xleftarrow[*sources=yupian]{\text{ancient}} \text{塘} \tag{4}$$

という出典情報付き異体字関係はこの異体字関係の典拠が玉篇であるという情報しかなく、玉篇のどこにあるという情報が存在しないからである。また、現実的にはどの玉篇という情報も必要だろうし、公開されている場所の URL も必要だろう。

しかしながら、複数の版の画像やテキストが各地で公開されている場合、特定のサイトで公開されているデータのみに依拠して、即ち、その URL を直接出典情報として記載してしまうと、他のサイトの画像やテキストが利用しづらくなってしまう。よって、もう少し抽象的な参照情報から具体的な URL を得る（名前解決する）方法を用いることが望ましいといえる。そのためには、両者の橋渡しをするためのデータが必要となる。

字書の場合、通常、文字が項目となっているので、各項目と項目の見出しが適切にマークアップされた構造化テキストや原本画像における掲出字の切り出し座標データなどがあれば文字をキーにして名前解決することが可能である。但し、幾つかの古字書では同じ文字が重複して別項目になっているケースがある。また、説文解字では掲出字が小篆で示されるが、楷書体ベースの現代漢字とのマッピングは通常 1 対多関係であり一般には多対多関係になる。[2)] この際、掲出字の小篆と現代漢字の対応関係を適切に扱うための仕組みが必要である。[6]

[2] では HDIC と CHISE を統合する際に、HDIC に収録されている古字書の掲出字に対応する字形オブジェクト（HDIC 代表字形オブジェクト）を辞書項目に対応するオブジェ

2)　例えば、「右」に関して説文解字は口部と又部の両方に項目を立てており、これらの説文小篆に対する楷書体としては「右」の他に「𠂇」も存在する。

クトの代用品として用いている。本来、これは文献オブジェクトを構成する部分としての辞書項目オブジェクトにすべきであるが、字形オブジェクトとすることで同じ文字が重複して別項目になっている場合に別の文字オブジェクトにしつつ、包摂関係やその他関係素性を用いて、同じ文字が重複しているケースと別字衝突しているケースを適切に記述することができる。また、仮に異なる部分がある異本を収録する場合も包摂階層の仕組みを用いて配置することが可能である。[3)] もっとも、漢字の包摂階層を用いるのはあくまで代用であり、この方法では掲出字の字体は同じだが注文が異なるケースにおいて、注文の系統に基づいた分類を適切に表現することができない。よって、やはり字形オブジェクトで代用するのではなく、辞書項目オブジェクトを別に立てるのが望ましいといえるが、ここで重要なのは、文献オブジェクトやその要素である辞書項目オブジェクトも、多粒度漢字構造モデルのような漢字のグリフの包摂階層と同様な、抽象的な文献・項目と具体的な文献・項目の包摂関係を記述可能なモデル化を行うべきだということである。もちろん、もし文献や項目の系統関係が判っているならばそうした中間階層を設けた多粒度化を行った方が良い。また、文献と項目の関係とそれぞれの抽象-具象関係の間に自然変換が成り立つように整理することが望ましい。

このような文献やその項目における抽象-具象関係の記述をマークアップの側から見た場合、TEI を用いたマークアップにおいて `<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がそれに相当すると考えられる。但し、通常こうしたマークアップは異本間の差異の記述を目的としており、そこで構成される部分構文木とそれ（その個々のテキスト及びそれらをまとめた基底テキストに代表される抽象的な項目）に対応する意味グラフにおける実体 (entity) との対応関係を同時に表現することはあまり意識されていない。言い替えれば、ここに文字列レベルでの差分（対応）と概念レベルでの差分（対応）が自然変換を為すようなマークアップ上の制約を科すことで、その両方の世界が対応したマークアップが可能になると考えられる。

4　おわりに

CHISE 文字オントロジーのような linked data で表現されるようなデータセットにその記述の典拠となる古字書などの元のテキストをつなげることを目的とした場合に求められるマークアップテキストの用件について議論した。

マークアップテキストはしばしばプレインテキスト（文字列）の側からその構造化という観点で捉えられるが、RDF や IPLD といった有向非循環グラフ (Directed acyclic graph,

3)　CHISE では説文小篆や篆隷万象名義の篆書掲出字の収録において既にこのような試みを行っている。

DAG) で表現されるようなデータから眺めた場合、木構造や文字列はその特殊な形であるはずである。そして、より重要なのは異なる記述・形式のデータが対応していることである。

文字の性質に関する記述に対してその典拠（さまざまな古字書やテキスト、研究書、論文等）がすべて適切に紐づけられた理想的な文字オントロジーが存在するとしたら、それをある文献という断面で切ったものはその文献の理想的なマークアップテキストと等価な情報を持っているはずであり、その情報をマークアップテキストとして記述することができるはずである。このような意味グラフの断面としてのマークアップテキストの要件を考察することは TEI で記述されたマークアップデータを RDF や IPLD の世界における意味グラフとして解釈したり処理する上でも有益ではないかと思われる。

また、本稿で取り上げたような問題を扱う場合、典拠となる文献を指示する際の粒度の問題やある粒度で指示したものに該当する具体的なリソースの集合の中からどのように代表を見つけ例示するべきかといった課題が生じるが、これに関しては別稿で検討したい。

本稿はやや抽象的で解像度の低い議論に終始してしまったことは残念であるが、今後、[4] などの TEI に基づく古字書のマークアップに関する既存の研究をベースにより具体的な記述のあり方についても検討したいと考えている。

参考文献

[1]守岡 知彦. *CHISE* **のデータ形式** *(Ver.0.1)*. `http://git.chise.org/~tomo/character/chise-format.pdf`. Aug. 2017.

[2]守岡 知彦. “CHISE における HDIC 統合の試み”. In: **情処研報** 2022-CH-129.12 (May 2022), pp. 1–6.

[3]池田 証壽. “漢字字体史の資料と方法 : 初唐の宮廷写経と日本の古辞書”. In: **北海道大学文学研究科紀要** 150 (Dec. 2016), pp. 201–236.

[4]岡田 一祐. “日本平安期古辞書の符号化モデル: TEI をもとにした符号化”. In: **デジタル・ヒューマニティーズ** 2 (2020), p. 26. DOI: 10.24576/jadh.2.0_26.

[5]李 媛. **空海の字書—人文情報学から見た篆隷万象名義**. 楡文叢書, Mar. 2023.

[6]守岡 知彦. “説文小篆に対する漢字構造記述の試み”. In: **東洋学へのコンピューター利用第** *34* **回研究セミナー**. July 2021, pp. 17–24.

[7]秋山 陽一郎, 守岡 知彦, and 浦田 衣里. “階層的素性名を用いた異体字記述の試み”. In: **情処研報** 2005.76 (July 2005). 2005-CH-67, pp. 55–61.