Vol.2017-CH-114 No.1
2017/5/13

CHISEのRDF化の試み

守岡 知彦[1]

概要：文字情報サービス環境 CHISE では文字に関する知識を機械可読化した文字オントロジーに基づいて各種処理を行なっているが、そのデータベースバックエンドでのデータモデルは RDF と同様な有向グラフ型であるものの独自のものであり、その完全な RDF 化は断片的・形式的なものに留まっていた。本発表では Chaon モデルや多粒度漢字構造モデルといった CHISE で用いている文字モデルの RDF での表現について検討し、CHISE 文字オントロジーの意味論的な情報も含めた自然な RDF 化を試みる。

1. はじめに

CHISE project [1] では、特定の汎用文字符号に制約されることなく、文字や文字の性質を自由に表現・処理するための環境を実現するために、文字、特に、漢字に関するさまざまな知識を機械可読な形式で記述した「CHISE 文字オントロジー」とその処理系である CHISE 実装の開発を続けている。CHISE は素性の集合によってオブジェクトを表現する『Chaon モデル』に基づき、文字をオブジェクトとして扱うことで、文字の内部表現を隠蔽するとともに、文字オブジェクトを扱うためのさまざまなインターフェースを提供できるようにしている。[2]

CHISE は集合ベースのデータモデルを採用しているが、オブジェクト間の関係を表すための『関係素性』と呼ぶ素性を用いることにより名前付き有向グラフを構成することができる。つまり基本的なデータモデルの構造としては Resource Description Framework (RDF) [3] と同様であるといえ、CHISE 文字オントロジーは情報を落すことなく RDF のグラフで表現できるはずであり、Turtle [4] や JSON-LD [5], RDF/XML [6] といった形式でシリアル化可能なはずである。

しかしながら、CHISE では階層的素性名方式に基づいて素性名に階層構造を導入することで情報の文脈やドメイン、あるいは、出典情報等のメタデータを表現する仕組みや、Chaon モデルに基づきオブジェクトに複数の ID を与え名前解決する仕組みや、包摂粒度を ID 素性の接頭辞で表現する仕組み等があり、CHISE の素性名をそのまま RDF の述語に対応させるのが簡単ではなかったり単純に対応させるだけでは問題があるケースがあった。例えば、[7] では CHISE の文字情報を RDF/XML 形式で表現するために、CHISE の階層的素性名や包摂粒度情報付き ID 素性を XML で使用可能な文字の範囲内に収まるようにエスケープするとともに文の『具体化』(reification) を用いてメタデータ素性の情報を表現することを試みたが、この手法では CHISE 文字オントロジーでのセマンティクスを十分に表現したグラフになっておらず、また、文の『具体化』を用いることによって必要以上の複雑さをもたらしているといえ問題があった。

本稿では CHISE 文字オントロジーを形式的に RDF にマッピングするのではなく、CHISE 本来のセマンティクスを RDF で適切に表現するとしたらどうなるかという観点で分析し、RDF へのマッピング方を考察する。

2. CHISE のデータモデル

2.1 Chaon モデル

CHISE では『Chaon モデル』と呼ぶ方法によって文字を表現するようになっている。これは汎用符号化文字集合に依存することなく自由に文字を表現するために我々が提案しているもので、表現したい文字に関する知識（文字の性質の集合）の機械可読な表現によって文字を表現し操作する方法である。

Chaon モデルでは、文字を説明するための要素（文字の性質や用例など）を『文字素性』(character feature) と呼ぶ。文字素性としては、部首、画数、部品の組合せ方に関する情報（漢字構造情報）、発音、意味、用例、その他文字処理で必要となる各種情報などが考えられる。

Chaon モデルでは、この文字素性の集合によって表現される文字のことを『文字オブジェクト』（文脈に応じて、文字もしくはオブジェクトと略される）と呼ぶ。

文字素性は素性の名前と値の対で表現することができ

[1] 京都大学人文科学研究所
Institute for Research in Humanities, Kyoto University

る。文字素性という用語は文脈によって素性の名前（で指されるもの）と素性名と値の対を指すことがあり、両者を区別するために、前者を『素性名』、後者を『素性対』と呼ぶことにする。

2.2 素性の種類

文字素性は大別すると

基礎素性 数値や識別子（シンボル）といったアトミックなデータ、または、それらのリストや配列を値として取る素性

ID 素性 オブジェクトに対する ID（素性名においてユニークな数値または識別子）を値として取る素性。ある ID 素性において、その素性を持つ各オブジェクトとその素性値である ID は 1 対 1 対応していなければならない（これにより、値である ID をキーにオブジェクトを得るための索引を作ることができる）

構造素性 オブジェクトを要素として持つリストや配列を値として取る素性

関係素性 オブジェクトの集合を値として取り、値に取った各オブジェクトと素性を持つオブジェクトの関係を表す素性。これは、オブジェクトをノードとした有向グラフのリンクとなるものである。関係素性対を持つオブジェクト（主語）とその値の各要素（目的語）の間には、その主語を値の要素とする逆関係素性対がその目的語に付く

メタデータ素性 元になる素性に対するメタデータを記述するための素性

に分類することができる。

ID 素性、関係素性、メタデータ素性の名前は、CHISE における素性名の命名規則によって、特定の形式を用いることになっている。

ID 素性は ‘=’ から始まる名前を持つ。また、CHISE の『グリフ ID 素性名の命名規則』[8] に従い、例示オブジェクトの集合を指すための名前から、それを抽象化したオブジェクトの集合を指すための派生名を接頭辞（表 2 の「S 式」列に示す）によって機械的に決定できるようにしている。[2]

関係素性は ‘->’ もしくは ‘<-’ から始まる名前を持つ。両者は互いに逆関係となっており、あるオブジェクト A が関係素性 ‘->*foo*’ を持ち、その値が $(B_1\ B_2\ ...)$ である時、オブジェクト B_i は逆関係素性 ‘<-*foo*’ を持ち、オブジェクト A はその素性値の要素のひとつとなる。

メタデータ素性は、言及対象となる素性名の後ろに ‘∗’ から始まる文字列を付けた名前を持つ。ここで、‘∗’ の後に続く文字列を『メタデータ識別子』と呼ぶ。

3. RDF での表現

3.1 オブジェクトの表現

CHISE の文字オブジェクトは Chaon モデルに基づき素性対の集合で表現される。この各素性対は RDF のトリプルに対応するものといえ、文字オブジェクトを主語、素性名を述語、素性値を目的語とした RDF のトリプルとして表現することができる（図 1）。ここで CHISE の素性名に対応した述語を『素性述語』と呼ぶことにする。

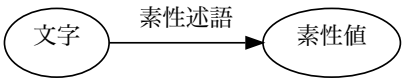

図 1 素性対の表現

ここで文字オブジェクトをどのように表現するかという問題が生じる。Chaon モデルは文字オブジェクトを素性対の集合という確定記述の束として、文字を特定の文字符号におけるコードポイントのような『固有名』に依らずに、表現するものといえるが、RDF では原則として IRI という固有名によって情報資源を指示する枠組を採っている。しかしながら、CHISE でも各種文字符号を表現するために ID 素性というものがあり、また、RDF にも空白ノードというそれ自体は IRI を持たず他の情報資源との関係性によって指示可能となるものもある。

Chaon モデルは集合論的なモデルであり、文字オブジェクトはそれが持つ素性対の指示対象の共通部分（論理積）であることを意味している。よって、素性対に対応する IRI を構成できるならば owl:intersectionOf を用いて表現することができる。

また、文字オブジェクトが複数の ID 素性を持つ場合、それらの ID 素性対は同じもの（即ち、その ID 素性を持つ文字オブジェクト）を指示するという性質がある。この場合、これらの ID 素性対に対応する IRI 構成できるならば、各 IRI 間の指示対象が同じことを owl:sameAs を用いて表現することができる。

Chaon モデル本来のセマンティクスを忠実に表現するという観点ではオブジェクトは原則として空白ノードにするべきであるかも知れないが、データへのアクセスという観点では繁雑であるといえ、CHISE-wiki [9] (EST [10]) で用いているような文字オブジェクトに含まれる ID 素性を用いて IRI を構成する方法が現実的であるといえる。この詳細は 3.6 節で述べる。

文字素性名	素性述語
ideographic-radical	ideo:radical
ideographic-strokes	ideo:strokes
total-strokes	ideo:total-strokes
->subsumptive	:subsume
<-denotational	:denotation-of
<-formed	:form-of
name	:name

表 1　主な素性名と述語

3.2　素性名の表現

素性名は RDF の述語に対応するものであり、IRI で表現することができる。

CHISE において、各素性名は文字オブジェクトの世界の中での固有性が保証されているので、CHISE 用の述語を置くための名前空間 IRI を定め、その下に素性名を付ければ良い訳であるが、CHISE の素性名で使用可能な文字の範囲の中には IRI でエスケープしなければならないものが含まれるため、可読性を考慮し CHISE-wiki における素性名の URL 表現 [9][*1]（以下では、この変換を『(素性名の）CHISE-wiki 符号化』と呼ぶことにする） を用いることにする。

CHISE の述語用名前空間は 1 つあれば十分であるが、Turtle 記法等において接頭辞を使用した場合にローカルパートが短く判り易いものになることと対象や用途の種類毎に述語の集合をまとめることを意図して、基本述語集合と漢字関連述語集合、漢字構造記述用述語集合の 3 種類の述語集合を設け、それぞれの IRI を `http://rdf.chise.org/rdf/property/character/main/`, `http://rdf.chise.org/rdf/property/character/ideo/`, `http://rdf.chise.org/rdf/property/character/isd/` とすることにする。文字定義の Turtle 文書ではそれぞれの名前空間接頭辞を `:`, `ideo:`, `isd:` とする（CHISE 関連以外の一般の文書で用いる場合は、その前に `chise` を付けることにする)。また、RDF の世界で標準的に使われる述語が存在する場合、適宜それを用いることとする。なお、他に適切な述語集合が存在しない場合、CHISE の基本述語集合の IRI 以下に CHISE-wiki 符号化した素性名を置いたものを用いるものとする。表 1 に主な素性名と述語の対応表を示す。

3.3　階層的素性名の表現

CHISE の階層的素性名では “@” の前にベースとなる素性名が付き、“@” の後にその情報が使われる文脈やドメインが付く。後者もまた “/” で区切られた複数のドメイン識別子からなる階層構造を採り得るものになっている。

[7] ではこのドメイン識別子の階層構造をエンコードして XML の名前空間のプレフィックス部に押し込むことによって表現していたが、この方法ではベースとなる素性名の指示対象と文脈やドメインが示す情報の関係が RDF のグラフとして表現されず問題である。そこで、今回は空白ノードを用いて表現する方法を採ることにする。

階層的素性名が使われている場合、ベースとなる素性名に対応した述語がとる目的語として空白ノードを設け、その空白ノードの述語 `:context` で階層的素性名におけるドメイン情報を表現する。また、述語 `:target` で素性値を表現する（図 2)。

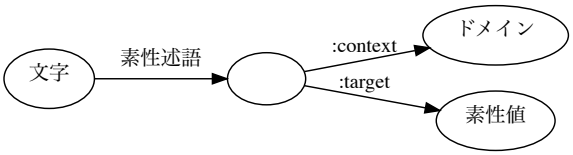

図 2　ドメイン付き素性名の表現

例えば、文字「𠂉」の UCS における部首を示す素性 ideographic-radical@ucs の値が 4 である時、この素性のベースとなる素性名 ideographic-radical に対応する素性述語 ideo:radical と空白ノードを使って図 3 のように表現することができる。

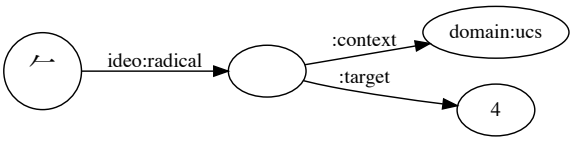

図 3　ドメイン情報付きの RDF の例

複数のドメイン識別子が存在する場合に空白ノードの述語 `:context` の値を RDF コレクションやコンテナ等を使って構造化することも考えられるが、当面、ドメイン情報記述用の名前空間 IRI として `http://rdf.chise.org/data/domain/` を設け、この下に CHISE-wiki 符号化したドメイン名を付けた IRI で表現することにする。ドメイン間の階層関係等の情報はこの IRI を起点とした RDF グラフで表現することにする。なお、文字定義の Turtle 文書ではこのドメイン情報記述用の名前空間の接頭辞として `domain:` を用いることにする（CHISE 関連以外の一般の文書で用いる場合は `chisedomain:` とする)。

[*1] ID 素性の表現に関しては、包摂粒度の分類体系が後に変更されたために、[9] での記述から変更されており、現在は表 2 の「IRI」列に示すものを用いている。

3.4 メタデータ素性名の表現

CHISE のメタデータ素性名では「*」の前に言及対象となる素性名が付き、「*」の後にそれに対するメタデータの種類を示す識別子やそのメタデータのドメイン識別子が付く。

このメタデータ素性で表現される情報も、3.3 節で述べた階層的素性名の表現と同様に、素性述語の目的語に空白ノードを設けて表現することにする。

階層的素性名の場合、空白ノードの述語 `:context` で階層的素性名におけるドメイン情報を表現し、同じく述語 `:target` で素性値を表現したが、メタデータ素性の場合、「*」以降のメタデータ識別子を CHISE-wiki 符号化したものをその空白ノードの述語として表現することにする。ここで、このメタデータ識別子に対応する述語のことを『メタデータ述語』と呼ぶことにする。

もしメタデータ識別子にドメインが付いている場合、階層的素性名の表現方法に従いその目的語を空白ノードとしてドメイン情報をその空白ノードの述語 `:context` で表現する。

メタデータ素性の値の RDF での表現はメタデータ識別子に依存する。出典情報を示す `:sources` の場合は、素性値の各要素のシンボルを CHISE-wiki 符号化したものの前に `http://rdf.chise.org/data/bibliography/` を付けた IRI を目的語とする。また、ここで、この名前空間 IRI `http://rdf.chise.org/data/bibliography/` に対する接頭辞を `chisebib:` とする。また、デフォルトでは次節で述べる基礎素性の表現に準じるものとする。

3.5 基礎素性の表現

基礎素性の値（基礎素性値）は概念的には RDF リテラルと同様なものといえるが、シンボルのような一意性を持ったもの、あるいは、リストや配列といった構造を持ったものもあるので、単純に全てを RDF リテラルで表現する訳にはいかない。

基礎素性値のうち、文字列や数値、真理値としての t は RDF においてもリテラルで表現できる。nil は真理値として使われている場合はリテラル、それ以外の場合は IRI を持ったエンティティーになる（空集合として使われている場合は RDF コレクションの rdf:nil とする）。t と nil 以外のシンボルは文字列で表現することもできるが、ユニーク性を示したい場合は IRI を持ったエンティティーとする。リストは RDF コレクションで表現できる。

3.6 ID 素性の表現

ID 素性は

(1) 文字符号の符号位置を示す
(2) 包摂粒度を示す
(3) 複数の ID 素性対を持つ文字オブジェクトにおいて、それらの素性対の等価性（指し示すものが同じであること）を示す

という 3 種類の機能を持っているといえる。

最初の機能は文字符号 (CCS) の名前を *name* とし、符号位置を *cpos* とする時、`http://rdf.chise.org/data/ccs/`*name*`/code-point/`*cpos* を IRI とするエンティティーを主語とし、符号位置を示す述語と整数値とのトリプルによって表現できる。ここで `http://rdf.chise.org/data/ccs/`*name*`/code-point/`*cpos* は *name* という名前の CCS の *cpos* という符号位置を示す IRI であり、こうした IRI を『符号位置 IRI』と呼ぶことにする。

2 番目の機能は包摂粒度を示す述語を使って文字オブジェクトに対応する IRI（これを『文字 IRI』と呼ぶことにする）と符号位置 IRI を関係づけることによって表現できる。文字 IRI としては、当面、CHISE-wiki [9] (EST [10]) で用いている `http://www.chise.org/est/view/character/`*prefix*`.`*name*`=`*cpos* というものを用いることにする。ここで *prefix* は包摂粒度を示す接頭辞である。この接頭辞を表 2 の「IRI」列に示す。また、包摂粒度を示す述語をこの表の「RDF 述語」列に示す。[*2]

包摂粒度名	S 式	IRI	RDF 述語
超抽象文字	`==>`	`a2`	:super-abstract-character-of
抽象文字	`=>`	`a`	:abstract-character-of
統合字体	`=+>`	`o`	:unified-glyph-of
抽象字体	`=`	`rep`	:abstract-glyph-of
詳細字体	`=>>`	`g`	:detailed-glyph-of
抽象字形	`==`	`g2`	:abstract-glyph-form-of
字形	`===`	`repi`	:glyph-image-of

表 2 包摂粒度の表現

3 番目の機能は異なる CCS の符号位置 IRI に対応した文字 IRI が存在する時にその等価性を `owl:sameAs` によって示すことで実現できる。但し、当面、文字定義の Turtle 文書では `owl:sameAs` と等価な述語`:eq` を用いることにする。

例えば、

```
((=ucs@jis . #x6F22)
 (=adobe-japan1-0 . 01533)
 (=jis-x0208 . #x3441)
 (=jis-x0213-1 . #x3441)
 (=gt . 22918)
 (=gt-pj-1 . #x3441)
 (=daikanwa/+p . 18068))
```

という ID 素性対の集合は図 4 のように表現することができる（但し、このグラフでは符号位置 IRI と符号位置の関係は省略している）。

[*2] 但し、空接頭辞「:」は `http://rdf.chise.org/rdf/property/character/main/` の省略記法とする。

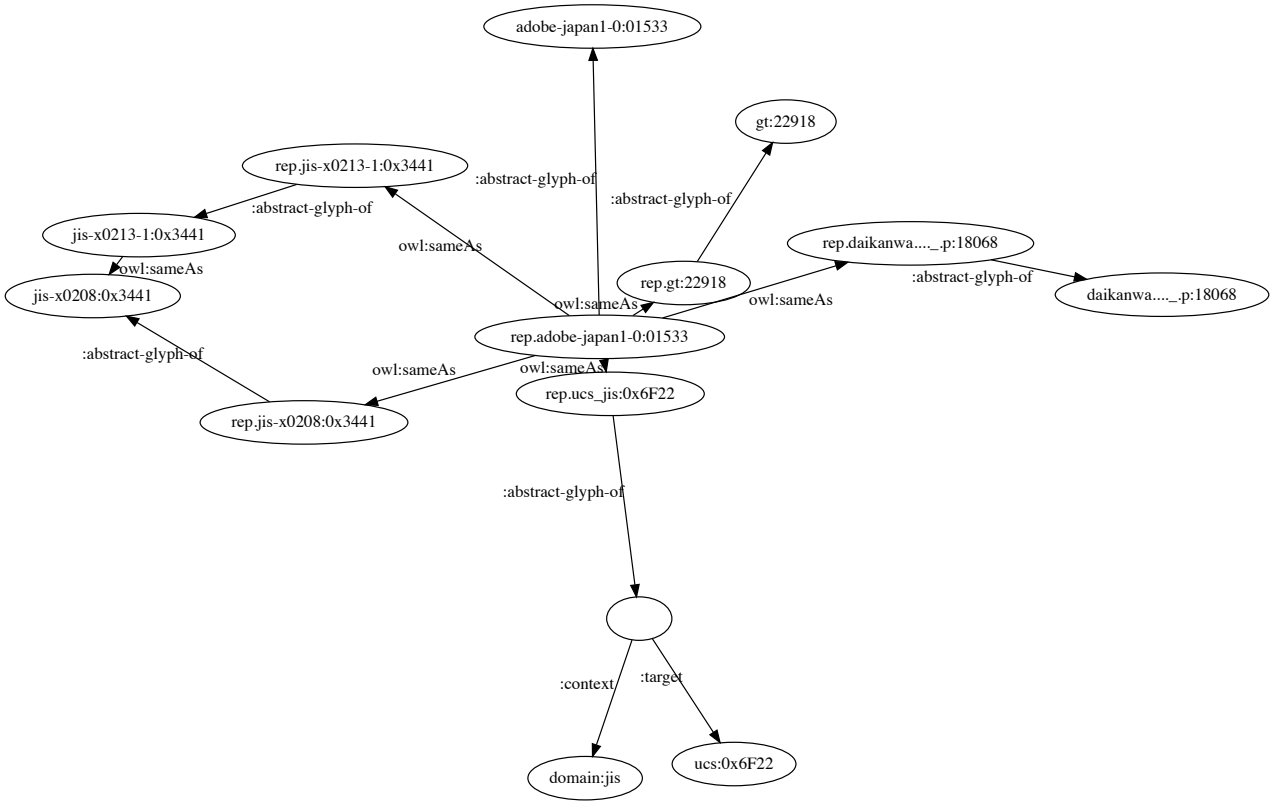

図 4　ID 素性の集合の RDF での表現の例（「漢」）

このように、CHISE 文字オントロジーにおける ID 素性はこの 3 種類の記述に分解することで RDF 化することができる。このため、ある文字オブジェクトが n 個の ID 素性対を持つ場合、通常、$3n-1$ トリプルの記述が必要となる。

3.7　漢字構造情報の表現

多くの漢字は複数の部品の組合せからなっており、字義カテゴリーを示す部品（意符）や音カテゴリーを示す部品（音符）などを用いた造字法によってさまざまな漢字が構成されている。このため、ある漢字がどのような部品をどのように組み合わせているかという情報は検索や処理、分析を行う上で大変有用な情報であり、これを『漢字構造情報』と呼ぶ。また、この情報の記述を『漢字構造記述』と呼ぶ。漢字構造記述の標準形式としては ISO/IEC 10646-1:2000 [11] で定義された IDS (Ideographic Description Sequence) 形式があり、CHISE ではこの形式に基づき UCS 統合漢字をほぼ網羅する「CHISE 漢字構造情報データベース」を提供している。

CHISE 文字オントロジーでは漢字構造情報は IDS 形式をパースした結果の構文木を S 式にしたものとして表現しており、その文字素性として ideographic-structure を用いている。RDF ではこのデータ形式をそのまま表現するのではなく、[12] で述べた『IDC の述語化』や IDS 用コンテナを使ったモデルを用いることにした。但し、記述の簡潔さや検索時の利便性を考慮して、以下に述べるような修正を加えている。

まず、漢字構造記述のための名前空間 IRI として `http://rdf.chise.org/rdf/property/character/isd/` を用い、その接頭辞として `isd:` を用いる。また、IDC のための型を表現するための名前空間 IRI として `http://rdf.chise.org/rdf/type/character/idc/` を用い、その接頭辞として `idc:` を用いる。そして、IDC のための型は idc: の後に IDC 文字を付けたものとする。例えば、「⿰」の型は idc:⿰ となる。

漢字構造情報を表現するための述語として `isd:structure` を用い、その値として型が前述の IDC 型の空白ノードを取ることにする。ここで、この IDC 型の空白ノードを『漢字構造コンテナ』と呼ぶことにする。このコンテナを主語として、IDC の種類に従い、その引数にとる部品を表 3 に示す述語の目的語とする。

例えば、「漢」の漢字構造記述「⿰氵 莫」は図 5 のように表現することができる。

4.　実装

XEmacs CHISE で動作する Emacs Lisp 言語で CHISE

IDC	述語 1	述語 2	述語 3
⿰	isd:left	isd:right	————
⿱	isd:above	isd:below	————
⿲	isd:left	isd:middle	isd:right
⿳	isd:above	isd:middle	isd:below
⿴	isd:surround	isd:filling	————
⿵	isd:surround	isd:filling	————
⿶	isd:surround	isd:filling	————
⿷	isd:surround	isd:filling	————
⿸	isd:surround	isd:filling	————
⿹	isd:surround	isd:filling	————
⿺	isd:surround	isd:filling	————
⿻	isd:underlying	isd:overlaying	————

表 **3** IDC 述語一覧

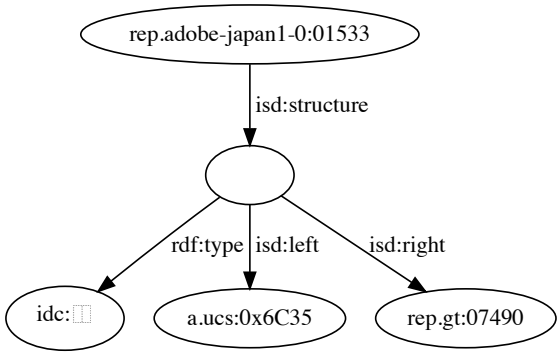

図 **5** 漢字構造記述の例（「漢」）

文字オントロジーを前節で提案した RDF に変換するプログラムを作成した。なお、シリアル化形式としては Turtle [4] を用いている。

この Turtle 版データの各ファイル名は XEmacs CHISE 附属の文字定義ファイルの拡張子を `.ttl` に変えたものになっており、今の所、同内容の情報を収録している。

このプログラムと Turtle 化したデータは版管理システム Git を使って管理・公開しており、この Git リポジトリの内容や変更履歴は `http://git.chise.org/gitweb/?p=chise/chiset.git;a=tree` で閲覧することができる。[*3] また、ローカルで Git の機能が利用可能な場合、

% git clone http://git.chise.org/git/chise/chiset.git

とすることで、このリポジトリをローカルにコピーして git コマンドを使って操作することも可能である。

また、SPARQL エンドポイントの提供も予定している。

5. おわりに

CHISE 文字オントロジーのセマンティクスをなるべく適切に表現できるような RDF 化へのマッピング手法と Turtle での実装について概説した。今回の試みでは以前のもの（例えば、[7]）に比べてより自然な RDF グラフが構成できたと考えられる。しかしながら、RDF における漢字の情報記述という観点では文字や形態素といった複数の種類の情報にまたがった記述に再編することが望ましいといえ、これは今後の課題としたい。

[*3] 現在の所、RDF データベースの一つである 4store への取り込みが成功したものだけを git commit / git push するようにしている。

参考文献

[1] : CHISE Project, `http://www.chise.org`.
[2] Morioka, T.: Multiple-policy Character Annotation based on CHISE, *Journal of the Japanese Association for Digital Humanities*, Vol. 1, No. 1, pp. 86–106 (2015).
[3] World Wide Web Consortium (W3C): *Resource Description Framework (RDF): Concepts and Abstract Syntax*, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ (2004).
[4] Beckett, D., Berners-Lee, T., Prud'hommeaux, E. and Carothers, G.: *RDF 1.1 Turtle*, World Wide Web Consortium (W3C), `https://www.w3.org/TR/2014/REC-turtle-20140225/` (2014).
[5] Sporny, M., Longley, D., Kellogg, G., Lanthaler, M. and Lindström, N.: *A JSON-based Serialization for Linked Data*, World Wide Web Consortium (W3C), `https://www.w3.org/TR/2014/REC-turtle-20140225/` (2014).
[6] World Wide Web Consortium (W3C): *RDF/XML Syntax Specification (Revised)*, `http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/` (2004).
[7] 守岡知彦：CHISE の階層的素性名の RDF 化の試みについて，情処研報，Vol. 2013-CH-97, No. 3, pp. 1–6 (2013).
[8] 守岡知彦：CHISE に基づくグリフ・オントロジーの試み，じんもんこん 2009 論文集，情報処理学会シンポジウムシリーズ，Vol. 2009, No. 16, 情報処理学会，情報処理学会，pp. 9–14 (2009).
[9] 守岡知彦：CHISE のセマンティック Wiki 化の試み，情処研報，Vol. 2010-CH-87, No. 8, pp. 1–8 (2010).
[10] 守岡知彦：Wiki 的手法に基づく構造化データの編集について，人文科学とコンピュータシンポジウム論文集—人文工学の可能性〜異分野融合による「実質化」の方法〜，情報処理学会シンポジウムシリーズ，Vol. 2010, No. 15, 情報処理学会，情報処理学会，pp. 33–40 (2010).
[11] International Organization for Standardization (ISO): *Information technology — Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane (BMP)* (2000). ISO/IEC 10646-1:2000.
[12] 守岡知彦：漢字構造情報の RDF 化の試み，すべてをコンピュータの中に—繋がってしまったデータとその未来 (山崎直樹，編)，京都大学人文科学研究所共同研究プロジェクト: 情報処理技術は漢字文献からどのような情報を抽出できるか—人文情報学の基礎を築く，全国共同利用・共同研究拠点「人文学諸領域の複合的共同研究国際拠点」，pp. 3–22 (2013).