文字メタデータを利用した文字学習コースウェアの開発

鈴木 優 中平勝子 三上喜貴
長岡技術科学大学
〒940-2188 新潟県長岡市上富岡町 1603-1
e-mail:053917@mis.nagaokaut.ac.jp

概要

本稿は，近年発展の著しいセマンティック Web の技術を利用し，文字に関するメタデータの策定及び語彙作成を行い，すべての国語学習の根源にある文字について RDF を用いて階層構造を構築することで，文字の電子的な体系化を行うことを提案する．また，これを用いて文字学習教材作成支援システムを作成した．このシステムは前述した文字メタデータを利用して，特定の文字に依存することなく教材開発を行うことを目指したものであり，また教材をメタデータによって自動的にソートし，簡単なコースウェアとしての形を整えることで，教材作成者の負担を減らすことを意図している．そして最後に，セマンティック Web を利用した文字の意味ネットワーク構築の将来像に関する考察を行う．

1. はじめに

セマンティック Web が 2001 年に具体的に提案[1]されて以来急速に技術として形を成しつつある．セマンティック Web とは，Web のコンテンツを機械で処理しやすい書式で表し，Web 上に意味ネットワークを構築して，エージェント等を用いてこれを利用しようという試みである．多くの研究者達の努力によって，これが実現可能な Web の未来像として具体化されつつある．

セマンティック Web で利用されている技術の一つである RDF(Resource Description Framework)は様々な形で既に我々の身近に存在している．サイトの最新情報などをメタデータとして記述，配信するための語彙である RDF Site Summary(RSS)1.0 や，主に Web 上のドキュメントの書誌的情報を記述する語彙である DublinCore[2]の DCMES，個人の特徴や人間関係を記述する語彙である FOAF[3]，また英単語の電子的な意味ネットワークである WordNet の一部も RDF にて既に定義されている．この他にも様々なジャンルにおいての様々な語彙が世界中の有志によって Web 上にて公開されている[4]．

しかし，こと「文字」というすべての国語学習の根源であり，また我々にとっても極めて身近で重要な情報単位についてはまだ確定的に定まってはいない．そこで，我々は文字に対してのメタデータを策定し語彙を作成することで，文字に関する情報を電子的な体系で表し，文字についての意味ネットワークを構築することを提案する．

また，このような文字メタデータを利用する分野として文字学習が考えられる．コンピュータを利用した学習システムについては，漢字などの分野において既に様々な研究がされており（例えば[5]），またインターネットを利用した効率的な学習システムについても研究されている（例えば[6]）．そこで今回我々はこの文字メタデータを利用した文字学習教材作成支援システムを作成した．以下，本論文では 2.で文字メタデータの策定について述べる．3.では文字メタデータの具体的な利用方法を明らかにし，文字学習教材作成支援システムについて説明する．そして 4.で文字メ

Development of a Character Learning Courseware Using Character Metadata.
Yu Suzuki, Katsuko Nakahira, Yoshiki Mikami
Department of Nagaoka University of Technology

	漢字	カタカナ/ひらがな	アルファベット	アラビア	ハングル
文字名	不明	不明	有	有	不明
形	有	有	有	有	有
音	有	有	有	有	有
義	有	無	無	無	無
文字源	有	有	有	有	有
筆順	有	有	有	有	有
翻字	有	有	有	有	有
順序	不明	有	有	有	有

表 1 文字による特徴の比較

タデータの構築と利用の将来像，課題についての考察を行う．

2. 文字のメタデータ

2.1. 文字とは何か

情報処理用語としての「文字」には，アルファベットや数字のほかに各種記号や制御文字までもが含まれており，極めて広義に定義されている．

本稿の主題として取り上げる「文字」とは，自然言語を記述するための書記体系を表すものを指す．よって以下で説明する「文字」については，上記の情報処理用語における記号などは含まないものとする．

2.1. Character クラスの設計

世界には 6000 を超える言語が存在していると言われており[7]，言語コードの規格である ISO639 が対象としているものだけでも 440 余りが存在している[8]．「文字」については言語数ほど多くはないものの，ISO 15924 Code Lists によれば 105 種類が存在している[9]．

これら多くの種類の文字に適用可能なメタデータを策定するには，まず始めに文字の諸特性を記述するのに必要と思われる共通要素について把握する必要がある．そこで，例として 5 つの文字をいくつかの特徴について比較した表を表 1 に示す．それぞれの共通要素は以下の通りである．

(1) 文字名　Name
(2) 形　Figure
(3) 音　Sound
(4) 義　Sense
(5) 字源　Origin
(6) 筆順　Stroke Sequence
(7) 翻字　Transcription
(8) 順序　Order in an Alphabet

以上の 8 つの項目を例として挙げたが，必ずしもそのすべてが値を持つとは限らない．例えば，漢字などの表意文字は文字単位で意味を表すが，アルファベットなどの多くの文字は意味を持たない．

このように上記のたった 8 つの項目に関してですら，すべての文字の特徴を満たすことができない．各文字にはさまざまな特徴があり，すべての文字に対して普遍的でありまた必要不可欠なメタデータを策定することは難しい．

そこでまず，多くの種類の文字で必要とされているプロパティを定め，これを Character クラスとし，すべての文字のスーパークラスとする．今回 Character クラスに用意したプロパティは，表 1 においてすべての種類の文字で該当した「形 (Figure)」「音 (Sound)」「文字源(Origin)」に加え，漢字な

どの表意文字において重要な「義(Sense)」，今回の比較では該当が少なかったがキリル文字，アラビア文字など多くの文字において大切な「文字名 (Name)」の５つである．翻字，順序に関しては複数の表現が考えられるため，筆順は形の一部でありまた表現するのが難しいため，今回はそれぞれ除外した．このCharacterクラスをRDFSにて記述したグラフを図１に示す．

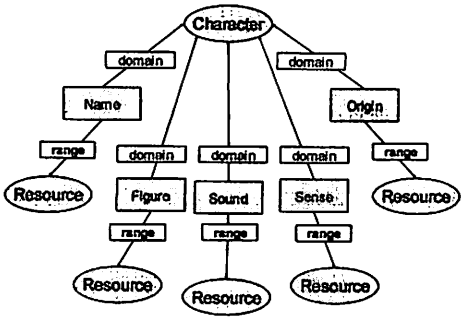

図１ Characterクラスの RDFS 表現

各プロパティの目的語となるリソースの値域（range）については，現状まだ詳細に定義できないので，ここではすべてのクラスの親クラスである Resource とする．

2.2. サブクラス設計

そして，このCharacterクラスを継承させ，それに各種類の文字ごとの特徴のプロパティを定めたサブクラスをそれぞれ作成していく．

ここでは例として次章にて説明する文字学習教材作成支援システムで利用することを考慮し，「ひらがな」のサブクラスを作成した(以下，Hiraganaクラスと呼ぶ)．

この Hiragana クラスで新たに定義したプロパティは以下の通りである．

(1)　ヘボン式ローマ字表現
(2)　カタカナ表現

それぞれのプロパティの意味は名前から容易に推察できるだろう．また，この Hiragana クラスのグラフ表現は次章にてインスタンスを用いて説明する．

スーパークラスとサブクラスの関係をRDFSにて記述したグラフを図２に示す．

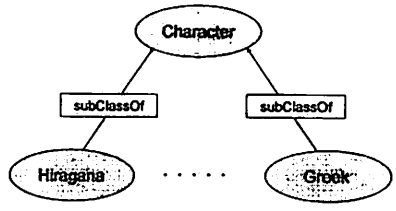

図２ それぞれの文字クラスの RDFS 表現

3. 文字学習教材作成支援システム

我々は前章にて定定義したCharacterクラスと Hiragana クラスを利用したシステムの一例として文字学習教材作成支援システムを作成した．本システムは教材作成者に対して文字に関する情報を予めメタデータとして提供することで，教材作成を容易にし，また特定の文字に依存しない教材開発環境の構築を目指したものである．ここでは最も我々に身近である母国語の「ひらがな」の文字学習教材を作成することにした．

3.1. 利用する文字メタデータ

ひらがなを学習する上で必要な情報とは何であろうか．これはひらがなに限らず全ての文字で言えることだが，何より必要な情報はその文字の形と音である．文字の形を視覚で確認し，音を聴覚で聞くことによって，初めて我々は自分でその文字を書き，その文字を発音することができる．これら二つの項目を満たす素材として，我々は今回フジノン株式会社の授業支援ツールで作成された MPEG の動画データを用いることにした[10]．これにより，ひらがな学習においてもう一つの重要な特徴である「筆順」をもフォローすることができる．この他に，先の Hiragana クラスの２つのプロパティ，(1) ある種発音記号とも取れるヘボン式ローマ字による表現，

(2) カタカナ表現，以上 2 つの項目を用意することにした．

これらの情報を元に Hiragana クラスにてインスタンス化した例を図 3 に示す．

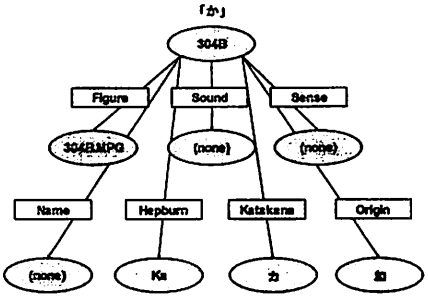

図 3 「か」の RDF 表現

主語となるのは Unicode である．これにより曖昧性が排除されて，文字は一意に決定される．Character クラスのプロパティである Figure の値はリソース (MPEG ファイル) であるが，Sound については動画のほうにまとめてあるため値を持たず，また Sense，Name もひらがなは値を持たない． Origin についてはリテラルの値を持つ．また Hiragana クラスのプロパティ，Hepburn，Katakana もそれぞれリテラルの値を持つ．

3.2. システムの概要

始めに，このシステム全体の構造を図 4 に示す．

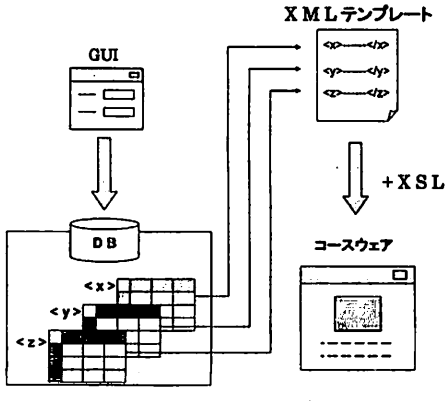

図 4 システムの構造

まず教材製作者は GUI によって教材に必要なデータを入力することになる．これは図 3 で示したそれぞれのプロパティの値を入力する作業である．この入力データはメタデータごとにテーブルを用意し，RDB に格納される．そして教材として出力する時にあらかじめ用意してある XML 文書のテンプレートの各タグ欄にデータを貼り付けてオブジェクトごとに XML 文書を構築する．そして XSL を用いてブラウザ上で確認できる形に変換する．

今回本システムは PHP と MYSQL を用いて実装した．

3.3. 出力教材のコースウェア化

また，このとき同時にそれらオブジェクトに対しての目次ページも出力する．この目次ページはメタデータごとにテンプレートを用意しておき，任意のメタデータによってソートされた結果を出力する．今回の「ひらがな」オブジェクトについてはひらがな順，Unicode 順，カタカナ順ぐらいしか目次の種類がなく，学習者にとってあまり実用的な効用は期待されないが，例えば将来「漢字」についてのオブジェクトを作成した場合，メタデータとして部首，画数，音読み，訓読み，があればそれをもとに画数順の目次，音読みの五十音順の目次などを作成することができる．こうして教材が意味のある順序を持って配置されることで，初めて教材はコースウェアとなる．このように，サブクラスの設計の時点で目次の仕様が決まるので，出力されるときには自動的にコースウェアが作成されるのである．よって，教材作成者の負担は教材の取捨選択及びコースデザインのみに集中することができる．

3.4. 出力例

作成したコースウェアの教材と目次ページの出力例を図 5，図 6 で示す．

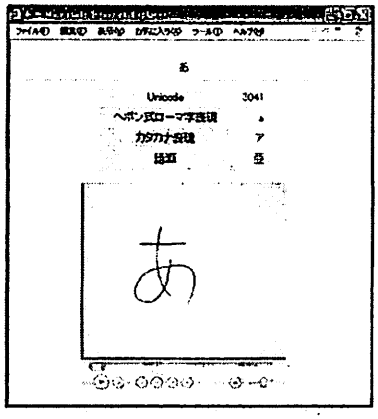

図 5 教材の出力例

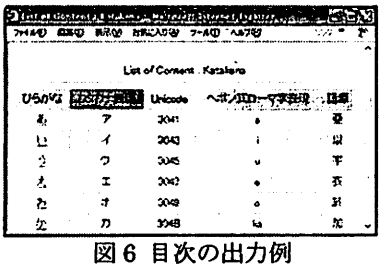

図 6 目次の出力例

4. 考察

4.1. 文字クラス設計の考察

現在のクラス構造ではCharacterクラスの下に各言語の無数の子クラスがぶらさがっている状態である．しかし，これら言語ごとのサブクラスを作成していく過程で，地域などによって言語間で共通の特徴を見ることができるだろう．そこで各言語を共通の特徴によるクラスタリングをし，それぞれに対応したサブクラスを作成することで構造化することが可能になる．こうした文字体系の分類は当然のことだが既に多くの研究がなされている[8]．このように既知の研究，学問との連携が，メタデータによる電子的な文字の体系化への近道と成り得る．

こういった研究をもとにすべての親クラスとなるCharacterクラスのプロパティについて，より効率的な再設計を行う必要がある．今回表1にて5つの文字について比較，検討を行ったが，ここにはインド系の文字なども含めておらず，普遍的な妥当性を検討するには検討対象が不十分と言わざるを得ない．

また，アルファベットには大文字小文字があり，漢字については音読み訓読み，簡体字，繁体字があるなど表1のカテゴリーの中にもまたサブクラスが存在しえる．今回作成したHiragana クラスのプロパティ，Hepburn，Katakana もまた，Character クラスのサブクラスとして翻字クラスを用意しておけば，そのプロパティと言えるだろう．こうした場面において，普遍性を犠牲に Character クラスを広く設計するか，サブクラスでのプロパティとして列挙するかも議論すべき問題である．

こうして議論の後に構造化することで，文字間の語彙の再利用を行えるメリットができ，また体系的に表すことによってより効率的な文字学習方法が浮かび上がってくると思われる．

また，これらクラスの仕様を決めた後に実際に文字ごとにオブジェクトを作成していけば，そこには一字一字の文字オブジェクトによる意味ネットワークが形成される．例えば図3の説明にて，今回は Katakana, Origin, などの値はリテラルとして扱ったが，「カ」や「加」のオブジェクトが作成されていればそれを値に持つことができ，ここに文字同士の繋がりが生まれる．こうして文字の体系化を電子的に可視化することで新たなる発見を助長し，言語学へのフィードバックも期待される．

4.2. セマンティック Web とコースウェア作成の将来像

普遍性のあるクラスが設計され，それが幅広く使われるようになり，多くのリソースが同じ語彙によって作成されることで，来るべきセマンティック Web 時代に対応したコースウェア作成を実現することができる．各研究者が同じ語彙を用いて文字オブジェクトを作成し，それがインターネットによって繋がることによって意味ネットワークが構築され，これを目的別の様々なエージェントが利用す

ることによって，同じリソースから異なるコースウェアをより少ない労力で作成することが可能になる．

また，RDFを使用した技術としてAnnoteaという機能がある[11]．これはRDFを使用して自分のWebコンテンツのみに限らず，第三者のWebコンテンツに注釈（アノテーション）を付与することを可能にし，それによってWeb上でのオープンなディスカッションを実現する仕組みである．まだ対応しているWebブラウザやWebコンテンツは数少ないが，将来的に普及したときには，文字クラス作成者や文字オブジェクト作成者の公開しているクラス，オブジェクトに関して多くの人の意見を取り込むことを可能にするだろう．「文字」という多くの人にとって重要な分野の仕様について，このような技術によって様々な意見が交換されることは必要不可欠であり，これによって教材作成のスタイルを協調的なものへと大きく変えていくことができるだろう．

5. おわりに

本稿では，文字メタデータの策定及び語彙作成の提案と，それを利用した文字学習教材作成支援システムについて述べた．文字メタデータは文字という概念を電子的な体系で表し，セマンティックWeb時代の意味ネットワークの構築を目的とする．文字学習教材作成支援システムは教材作成を容易にし，また自動的なコースウェア作成を実現する．

今後は，親クラスの再検討及び文字別の様々なサブクラスの設計といった概念的仕様についての研究と文字学習教材作成支援システムの改良といった実用面についてとの両側面から研究を続けていく必要がある．現状サブクラスについてはHiraganaクラスしかなく，また文字学習教材作成支援システムについてもそれについてにしか対応していない．そこで，様々なサブクラスの設計を行い，システムはそのサブクラスを換装させることで，GUI部分を自動的に変化させることができ，そこで初めて文字に依存しない教材開発を行えるというこのシステムの有効性が現れるだろう．

謝辞 本研究は，フジノン株式会社より依頼された受託研究「技能者育成のためのオンデマンドプレゼンター併用型指導実践及びその教育効果測定」の補助を受けて行われたものである．

参考文献

[1] Tim Berners-Lee, James Hendler, Ora Lassila, "The Semantic Web", Scientific American, 284 May 2001

[2] Dublin Core Metadata Initiative(DCMI), http://dublincore.org/

[3] the friend of a friend(foaf) project, http://www.foaf-project.org/

[4] Schema Web – RDF Schemas Directory, http://www.schemaweb.info/

[5] 前田和昭，龍岡亮二，押木秀樹，"漢字CAIのための漢字情報管理システムの開発"，情報処理学会論文誌，Vol. 40 No. 3 Mar. 1999

[6] 越智洋司，矢野米雄，脇田里子，林敏浩，"ユーザのブラウジングから学習漢字を選定する漢字学習環境の構築"，情報処理学会論文誌，Vol. 40 No. 2 Feb. 1999

[7] Raymond G. Gordon, Jr., "Ethnologue Languages of the World Fifteenth Edition", International Academic Bookstore

[8] 三上喜貴，"世界の文字と文字符号（前編）"，情報処理，Vol. 46 No. 8 2005

[9] Unicode Home Page , http://www.unicode.org/

[10] 横山淳一，松田信一，中平勝子，福村好美，"マルチメディア取り扱いが容易な授業支援ツールの開発"，情報処理学会研究報告 Vol. 2004 No. 117 pp. 61-66

[11] 斎藤信男，萩野達也，"セマンティックWeb入門"，オーム社