Vol.2011-CH-91 No.6
2011/7/30

東洋学文献類目のセマンティック Wiki 化の試み

守　岡　知　彦[†1]

「東洋学文献類目」（類目）の EST を用いた再実装の試みについて述べる。類目はデータの蓄積の点でも電子化の点でも歴史が長く、レガシー問題を抱えている。このメンテナンス性の向上と Linked Data 的な性格の付与を試みる。

An experimental Semantic-Wiki service for the Bibliography of Oriental Studies

MORIOKA Tomohiko[†1]

This paper explains a new implementation of the Bibliography of Oriental Studies (Ruimoku) system based on EST. Ruimoku has long history of editing, publishing and digitization, so it has serious legacy problems. New implementation of Ruimoku is designed as Linked Data to provide usability and improve maintenancability.

1. はじめに

「東洋學文獻類目」（以下では、「類目」と呼ぶことにする）は、京都大学人文科学研究所附属東アジア人文情報学研究センターが毎年刊行している、東洋学研究に関する書籍や論文等の文献の目録である。1980 年台に類目の編纂の電子化が始まって以来、[1] データベースとしての側面も備えるようになり、現在では、冊子体として毎年刊行される「東洋學文獻類目」とデータベース検索サービスとしての「東洋学文献類目検索」の両面からなる事業となっている。

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

類目は 1935 年に東方文化学院京都研究所の事業として 1934 年度版「東洋史研究文獻類目」の編纂が始まって以来、第二次世界大戦中・戦後の混乱期に一時複数年の合併号となったのを除けば、70 年以上にわたってほぼ毎年編纂・刊行し続けた歴史の長いデータベースである。電子化も 1981 年に始まっており、[2] 既に 30 年が経過している。このことは類目のデータの量的・質的特徴を示すと同時に、その（人的側面を含む）システムがレガシー・システムとしての問題を抱えているということを意味している。[3] つまり、その間に生じたさまざまな社会的・技術的変動を限られた人的・経済的リソースによって対処しながら長期にわたって編纂・刊行を続けてきたということは、各時代において生じた問題をさまざまな問題をその時々に ad hoc に対処し続けてきたということを意味しており、また、これまでの類目の歴史に生じたさまざまな事情の全てを理解している人が誰もいないということでもある。これは仕様が厳密には良く判らず形式化しづらいということであり、リファクタリングもしづらいということでもある。これはメンテナンスという観点で問題である。[4]

こうした歴史的事情により、類目の編纂体制は、大別して

(1)　電子化以前（1934〜1980 年度版）
(2)　汎用機時代（1981〜2000 年度版）
(3)　現行データ（2001 年度版以降）

の 3 期に分かれる。電子化以前のデータも後に遡及入力が行われ、一応、全ての時期のデータが電子化されているが、遡及入力は編纂作業に使われた元データからではなく類目の冊子体から行われたため、1981 年度版以降のデータに比べて情報が欠落しており、データ形式も異なっている。また、汎用機時代のデータは当時の技術的制約から、さらに 2 種類の異なるデータセットに分岐しており、結局、大別すると 4 種類の異なる形式が存在している。しかしながら、各形式の仕様は十分に形式化できておらず、全ての変種をサポート可能なスキーマを書くのは容易ではなく、結果として、2 種類までの統合しか実現できていないのが現状である。

そこで、類目データのリファクタリング支援のために、著者が開発している構造データのための Wiki である EST[5] を用い、類目の再実装（類目 Ver.7）を試みている。ここでは、各種メタデータのオブジェクト化に焦点を当てて類目 Ver.7 について概説する。

2. 類目のマスターデータ

類目のマスターデータの形式は、電子化時代以前に用いられていた編纂用のカードの影響を強く受けたものとなっており、最大 3 階層の親子関係を持った入れ子状のデータ構造と

なっている。また、批評関係等を表すためのカード間のリンクも存在する。このように、単純な 1 枚の表では表現できないようなデータ構造になっている。また、電子化のための形式としては書誌情報の交換用形式であった UNIMARC の影響を受けていると考えられる。

各文献は、単行本や雑誌、論文、細目論文、書評論文などの種別に応じたカードで表される。各カードはカードの種類によって決まる幾つかのフィールドからなる。

汎用機時代のデータでは、各フィールドは、行番号、タグ、サブフィールドという 3 つの欄からなる。また、サブフィールドは、親カードにおける子番号（以下、単に『子番号』と呼ぶ)、$ + 英字 1 文字 からなるサブフィールド名、および、サブフィールド値（文字列）からなる（図 1)。なお、実際には、子番号はサブフィールド名/値とは別の欄になっている。そして、行番号欄、タグ欄、子番号欄は EBCDIC, サブフィールド名/値欄は JEF 漢字コード (=jef-china3) で符号化されている。また、サブフィールド値中の漢字はいわゆる『康熙体』に正規化されている。

雑誌-論文-細目論文 や 単行本-書評論文 等の親子関係やリンクを表現するために、階層構造やリンク構造を持ったデータ形式となっており、雑誌カードの後には幾つかの論文カードが続き、論文カードの後に複数の細目論文カードを置くこともできる（図 1)。書評論文カードや批評論文カードでは対象となる文献に対するリンクを `$1` *ID* という形式で表現している（図 2)。

現行類目のデータも基本的に汎用機時代のデータ形式を踏襲しているが、UTF-8[6] のプレイン・テキストによって符号化されており、欄という概念はない。論文・細目論文を除きタグは廃止されており、原則として、サブフィールドの情報だけが独立したフィールドとして表現されている。また、論文・細目論文レコードのタグは子番号を後ろに付けた形で独立した行として表現されている（図 3)。また、漢字の正規化はやめて、原則として、原表記で入力することになった。また、現行類目に移行後の途中から、中国語著者名にピンインを入力することになり、著者名欄として、原表記とともに任意の言語タグと必要に応じて対応する漢字/カナ/ラテン文字表記を付与可能な新形式を導入した。この他、段階的に幾つかの拡張が加わっており、幾つかのバリエーションが存在する。

一方、1980 年度版以前の類目を遡及入力したデータは、編纂作業に使われたカードが既に破棄されていたために、冊子体の情報を入力することとなったため、汎用機時代のデータや現行類目と比べて情報が足りないことと、汎用機時代のデータ形式への理解不足から、類目のカード間の階層・リンク構造が崩れたデータ構造になっており、汎用機時代のデータや現行類目とデータモデル上の不一致が大きい上、さまざまなフィールドが十分に形式化さ

	行番号	タグ	サブフィールド		
雑誌カード	00021070	010		Ａ８２０９１５	ID
	00021080	020		＄ｊ１３２８００	雑誌名コード
	00021090			＄ｋ２０	巻
	00021100			＄ｄ１９８２（3）	出版年（月）
	00021110	030		＄ｎ創立二十周年記念特集號	雑誌・著者の注記
論文カード	00021120	100	01000	＄ｔ佛教の體系と展相の研究（２）	タイトル
	00021130			＄ａ武邑 尙邦	著者名
	00021140			＄ｆタケムラ ショウホウ	著者名のよみ
	00021150			＄ｚｌ	著者名の排列種別
	00021160			＄ｐｌ２８－１６１	ページ
	00021170			＄ｂ０７４３	分類コード
細目論文カード	00021180	110	01001	＄ｔ佛教の體系的同一性と歴史的多様性について	
	00021190			＄ｕインドの場合	副題
	00021200			＄ａ武邑 尙邦	
	00021210			＄ｆタケムラ ショウホウ	
	00021220			＄ｚｌ	
	00021230			＄ｐｌ２８－１３９	
細目論文カード	00021240	110	01002	＄ｔ說一切有部における遍知（承前）	
	00021250			＄ａ加藤 宏道	
	00021260			＄ｆカトウ ヒロミチ	
	00021270			＄ｚｌ	
	00021280			＄ｐｌ３９－１５２	
細目論文カード	00021290	110	01003	＄ｔ龍樹における知の問題（1）	
	00021300			＄ｕ動詞ｊñā とその派生語の使用の検討を通して（上）	
	00021310			＄ａ田丸 俊昭	
	00021320			＄ｆタマル トシアキ	
	00021330			＄ｚｌ	
	00021340			＄ｐｌ５３－１６１	

図 1 汎用機時代のマスターデータの例（細目論文）

Vol.2011-CH-91 No.6
2011/7/30

	行番号	タグ	サブフィールド	
雑誌カード	00024590	010	Ａ８１０２４０	ID
	00024600	020	＄ｊ２７７０００	雑誌名コード
	00024610		＄ｋ５５	巻
	00024620		＄ｇ３	号
	00024630		＄ｄ１９８２（３）	出版年（月）
書評論文	00024640	200	00001 ＄ｒ千葉 徳爾	評者名
	00024650		＄ｆチバ トクジ	日本語著者名のよみ
	00024660		＄ｚ１	著者名の排列種別
	00024670		＄ｐ１９８－１９９	ページ
	00024680		＄ｌＢ８１０１１９	リンク
単行本カード	00035090	510	Ｂ８１０１１９	ID
	00035100	520	＄ｔ神話と傳說の旅	タイトル
	00035110		＄ａ川喜田 二郎／加藤 千代	著者名
	00035120		＄ｆカワキタ ジロウ／カトウ チヨ	
	00035130		＄ｚ１１	
	00035140	540	＄ｅ古今書院／／東京	出版者//出版地
	00035150		＄ｄ１９８１	出版年
	00035160		＄ｐ２４６	
	00035170		＄ｓネパール叢書	シリーズ名
	00035180	550	＄ｍ圖１７表２	ページの注記
	00035190	560	＄ｂ１４４Ｘ	分類コード
	00035200		＄ｃ４２Ａ０ＸＸ１	排列のための手がかり

図 2 汎用機時代のマスターデータの例（リンク）

雑誌レコード	A2007-00471	ID
	$j040200	雑誌名コード
	$k59	巻
	$g2	号
	$d2007(9)	出版年（月）
	$i通巻第569号	巻号の注記
論文レコード	10000001	タグ+子番号
	$t前漢皇帝陵の再検討:陵邑, 陪葬の変遷を中心に	タイトル
	$A 村元　健一;ja(ムラモト　ケンイチ)//著	著者名（日本語）
	$p38-60	ページ
	$b12XX	分類コード
	$c111000B111EXX1	排列のための手がかり
論文レコード	10000002	タグ+子番号
	$t韓半島南海岸新石器時代の埋葬遺構	
	$A 任鶴鐘;zh[ren he zhong]//著	著者名（中国語）
	$A 平郡　達哉;ja(ヒラゴオリ　タツヤ)//訳	訳者名（日本語）
	$p127-145	
	$b12XX	
	$c18B010EXX1	
	$q東三洞貝塚, 金谷洞栗里貝塚, 欲知島, 山登貝塚, 礼安里, 煙台島, 凡方貝塚	内容の注記

図 3 現行類目のマスターデータの例（現行形式）

れていないデータ形式になっているために、データのバリデーションという観点でも幾つかの問題を抱えている。

3. データモデル

類目 Ver.7 は EST とそのバックエンドである Concord[7] を用いて実現されており、さまざまな情報は、意味を持つまとまり毎に、素性の集合からなるオブジェクトとして表現される。オブジェクトの素性は関係データベースの属性（列）と異なり容易に追加することが可能であり、試行錯誤がしやすい。また、異なる ID 体系やデータ形式毎に固有の素性を対応させることで、オブジェクトを参照するための複数の方式を容易に共存させることができる。

Concord / EST ではオブジェクトは『ジャンル』というオブジェクトの種類（共通するインターフェースを持つ同様な種類のオブジェクトの集合；名前空間のようなもので、一般的なオブジェクト指向言語におけるクラスに似ている）を持つが、類目 Ver.7 では

Vol.2011-CH-91 No.6
2011/7/30

creator@ruimoku 作者（著者、訳者、編者、etc.）のオカーレンス[*1]
person-name@ruimoku 作者の名前
journal@ruimoku 雑誌（定期刊行物）
journal-volume@ruimoku 雑誌（定期刊行物）のある号
article@ruimoku 論文
book@ruimoku 単行本
classification@ruimoku 類目における分類
region@ruimoku 地域
period@ruimoku 時代

というジャンルを設けている。

雑誌レコードは journal-volume@ruimoku ジャンル、論文・細目論文レコードは article@ruimoku ジャンル、単行本は book@ruimoku ジャンルのオブジェクトになる。

サブフィールドはオブジェクトの素性として表現する。

雑誌名 ($j) や分類 ($b) のように値がコード化されているサブフィールドはそのコード値を ID 素性値としたオブジェクトに対する関係素性（例：雑誌名の場合は <-volume, 分類の場合は->classification）とし、オブジェクト間の関係として表現する。

親子関係は ->included（親から子）, <-included（子から親）で表現し、批評対象へのリンク（$l）は ->reviewed で表現する。

4. 作　　者

類目では、現在の所、各文献の著者・訳者・編者等は単に名前で表現されており、同姓同名の別人を区別して人（法人）を同定することは行われていない。

一方、現行類目では漢字表記を原則として原文のままにしているので、同じ漢字名が繁体字・簡体字・日本新字で別文字列になってしまうことがある。また、翻訳の際等に、漢字名がラテン表記されたり、ラテン表記の名前に漢字表記が付けられたりする場合もある。また、こうした転写に複数の変種があることもある。

名前表記の変種の問題を鑑みれば、人オブジェクトが 1 つ以上の名前オブジェクトを持つという風にモデリングした方が良いといえるが、人の同定は手間であるし判らないこともあるので、現状では人オブジェクトを設けるのは難しい。

[*1] occurrence; 著作行為のインスタンス

そこで、文献毎にその著者・訳者・編者等の役割を担った人をオブジェクト化することにした。これが creator@ruimoku ジャンルのオブジェクトで、これは人の『出現』(occurrence) に相当する。

この作者ジャンルのオブジェクトには、役割に関する素性と名前オブジェクトへのリンクとなる関係素性を持つ。

役割に関する素性としては、現在の所、役割の名前を示す素性 role*name と役割のタイプを示す素性 role*type を設けている。ただ、これは役割オブジェクトにして、そのオブジェクトへのリンクとなる関係素性 ->role にした方が良いかも知れない。

一方、名前オブジェクトへのリンクは関係素性 ->name で表現している。

また、名前オブジェクトは、名前の表記 (文字列)、言語 (あるいは、書記系等)、種類 (個人名、団体名等)、別表記へのリンクからなるものとしている。表記毎に名前オブジェクトを立て、別表記へのリンクを張ることで、CHISE 文字オントロジー[8] における異体字・類字関係の表現と同様に扱うことができる。

5. 地域と時代

類目のマスターデータでは、文献が対象とする地域や時代等の情報を「排列のための手がかり」($c) というサブフィールドに格納している。これは文献の内容に関するメタデータの一種であり、地域コード、時代コード、事項コード、内容コードという 4 種類の情報を結合したものであり、この $c の値でソートすると類目の冊子体の排列に近い結果になるよう工夫されている。このことは、言い替えれば、検索の容易さのことはあんまり考えてないということを意味しており、$c 内の各コードや分類コード ($b) との間で依存関係が生じる扱いづらいものとなっている。

時代コードは、西暦年や世紀による表現の他、古代・中世・近世といった時代区分（この時代区分は対象地域毎に異なる）や王朝名等（この時代区分も対象地域毎に異なる）も使えるようになっており、また、これらの形式を使った時間表現に対して、初期・中期・末期・前半・後半といった修飾子を付けることもでき、更に、このようにして表現された（修飾子付き）時間表現を 2 つ使って、開始時期と終了時期からなる期間を表現することもできる。

このように類目の時代コードは複合的な時間表現を重ねたようなものとなっており、また、同じ時期・期間を複数の表現形式で表すことができる。これは西暦年に正規化したようなシステムに比べて非常に複雑であるが、その代わり、さまざまな時代概念を複数の異なる抽象レベルで表現できるという利点がある。

Vol.2011-CH-91 No.6
2011/7/30

そこで、類目 Ver.7 では類目の時代コードに対応するような時間オブジェクトを導入した。開始時期と終了時期からなる期間を表す時間オブジェクトの場合、開始時期を示す時間オブジェクトへのリンクと終了時期を示す時間オブジェクトへのリンクを張る。修飾子付き時間オブジェクトの場合、修飾子の無い基本的な時間オブジェクトへのリンクを張るとともに、開始年を示す時間オブジェクトと終了年を示す時間オブジェクトへのリンクを張る。また、時代や王朝名等による時間オブジェクトには地域オブジェクトへのリンクを張る。このように、オブジェクト間のネットワーク構造によって時代コードを表現する訳である。

類目の時代コードの表現には文献の内容・種類・分野等に対応した偏りがあり、こうした複合的な時間オブジェクトを用いることにより、単に西暦年に正規化しただけでは見えなかったような構造が見えることがある。また、西暦年や年代・世紀オブジェクト等を介在して、国・地域をまたいだ時代表現や時代区分の差異等を可視化するのも容易である。

なお、現在の類目 Ver.7.0 では時間オブジェクトと地域オブジェクトを別々の素性として論文・単行本オブジェクトに付与しているが、文献の分類項目という観点で考えた場合、両者を組み合わせた『時空間オブジェクト』として（本来の $c に近い形として）扱い、時間オブジェクトや地域オブジェクト間のネットワーク構造によって扱った方が良いかも知れない。

6. 検　　索

現在の所、EgT 自体には検索機能がないため、類目 Ver.7 専用の検索用ページを設けている（図 4）。これには、現在の所、タイトルおよびキーワードを検索するための入力窓と、分類のトップオブジェクト（図 7）や幾つかの地域オブジェクトへのリンクを用意している。

検索用の入力窓には検索したい文献のタイトル、もしくは、キーワードに含まれる文字列を入力し、検索開始ボタンを押すと検索が実行され、検索結果が表示される。複数の文字列をスペースで区切って並べた場合、AND 検索が行われ、これらの全てを持つ文献のリストが返される（図 5）。検索結果の表示画面には EgT による文献オブジェクトへ表示用ページへのリンクが張られており、それをクリックすることで文献の詳細情報を見ることができる（図 6）。

7. Linked Data 的側面

現在の所、EgT は RDF を出力できないので、Web 標準に則った Linked Data にはまだなっていない（将来的には、EgT に RDF 出力機能を追加することを計画している）が、概

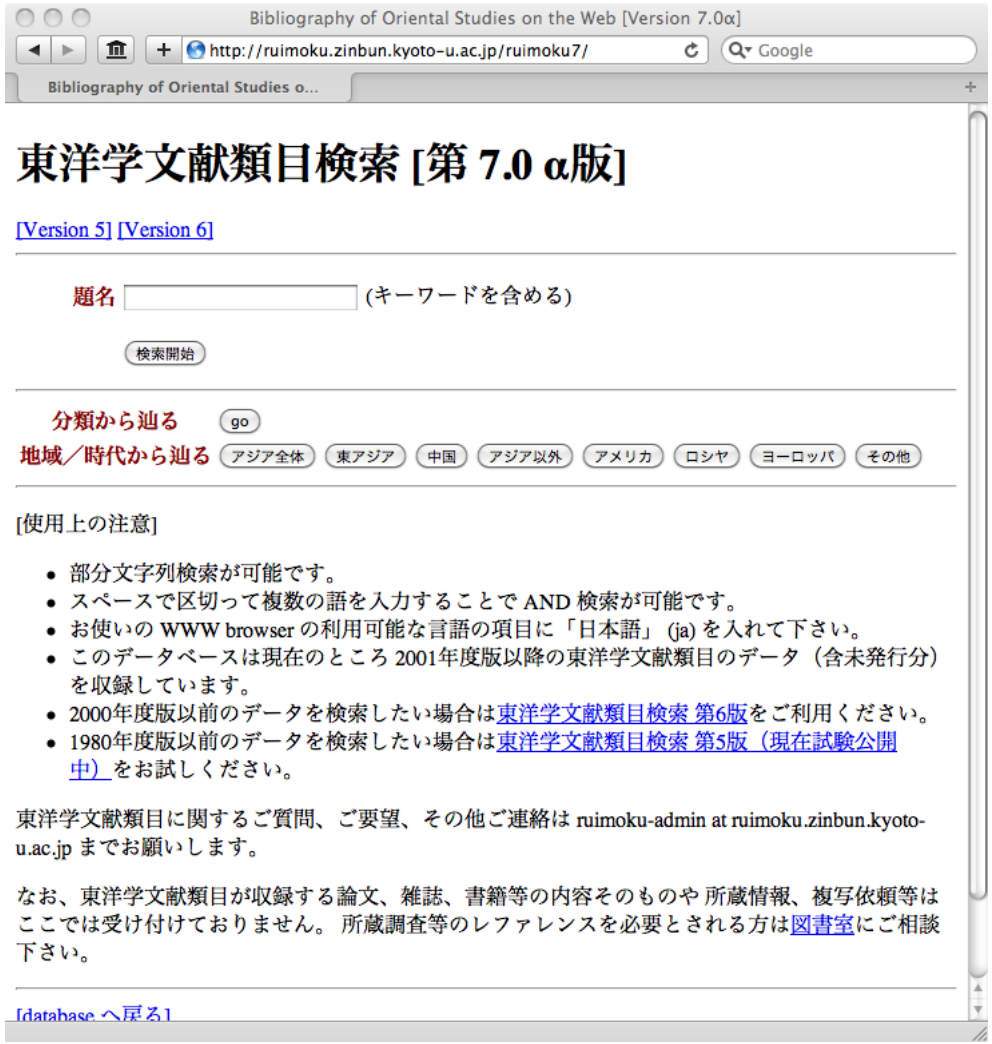

図 4　類目 Ver.7 の検索画面

図 5　類目 Ver.7 の検索結果の例

図 6　論文表示画面の例

図 7　類目 Ver.7 の分類のトップ

念的には類目 Ver.7 は Linked Data 的な性格を有していると考えられる。

EsT では全てのオブジェクトに固有の URI が与えられるので、その枠組の上で表現された類目 Ver.7 のオブジェクトもまた全て固有の URI を持つ。これらは類目 Ver.7 システムの外からも参照・リンク可能であるので、他のデータベース、オントロジー、サービス等から類目の情報を比較的深いレベルで利用可能となる。

8. Wiki 的側面

EsT ではオブジェクトの素性値や素性の定義などを Wiki 的に編集することが可能であり、類目の情報を修正したり、本来の類目に含まれていなかった情報を追加したり、情報の表示の仕方を変更したりといったことが可能である。これは類目のデータをリファクタリングする上でも有用であり、類目の情報を別のサイトの情報と連係させる上でも有用であると考えられる。

9. おわりに

東洋学文献類目（類目）の EsT を用いた再実装について概説した。

類目は階層構造やリンク構造を持った構造化データになっており、また、長期にわたって蓄積が続けられたため形式や運用面で変遷があり、構造化データでありながらその構造の詳細について全てのデータを網羅する形での十分な形式化ができていないという問題があり、一般的な関係データベースで扱う上で問題があった。また、従来の関係データベース (PostgreSQL) を用いた実装（類目 Ver.5, Ver.6）では類目が本来有するさまざまなメタデータを十分に活用することができていなかった。

Concord / EsT を用いることで、類目本来の構造に近い形でデータを格納できるようになり、また、各種コードのような具体的なデータ表現をオブジェクトとして隠蔽することで、異なるデータ表現を共存させやすくしたり、リファクタリングしやすくすることが可能になったといえる。また、EsT によってオブジェクトを可視化することで、データ構造の見通しが良くなり、リファクタリングの方針が立てやすくなった。また、全てのオブジェクトに対応する URL が与えられ、外部から利用しやすくなったといえる。このことは基盤データとしての類目を第三者が活用する上で重要なことだといえる。

現在の所、類目 Ver.7 は現行類目のみをサポートしており、汎用機時代の情報や電子化以前の情報の遡及入力データをサポートしていない。今後、実際にこうした異なる形式のデータの取り込みを試み、利用者の利便性の向上を計るとともに、EsT を用いたレガシー・データベースのリファクタリングのケーススタディーとしても研究を進めて行きたいと考えている。

参 考 文 献

1) 星野　聰，勝村哲也：東洋学文献類目データベースの研究と開発，情報処理学会論文誌，Vol.25, No.2, pp.187–193 (1984).
2) 安岡孝一：東洋学文献類目』の編纂の歴史 — CHINA3，センター所蔵資料の活用と人文社会科学，13, pp.63–70 (2003).
3) 守岡知彦：レガシーとの付き合い方 — 東洋学文献類目の場合，漢字文献情報処理研究，No.11, pp.82–95 (2010).
4) 守岡知彦：データを生み出すデータのために，人文科学とコンピュータシンポジウム論文集 — サービス指向のデジタル技術へ〜人文科学のポテンシャル〜，情報処理学会シンポジウムシリーズ，Vol.2008, No.15, 情報処理学会，情報処理学会，pp.13–18 (2008).
5) 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).
6) International Organization for Standardization (ISO): *Information technology — Universal Multiple-Octet Coded Character Set (UCS)* (2003). ISO/IEC 10646:2003.
7) 守岡知彦：Concord: プロトタイプ方式のオブジェクト指向データベースの試み，Linux

Vol.2011-CH-91 No.6
2011/7/30

Conference 抄録集, Vol.4 (2006).
8) Morioka, T.: CHISE: Character Processing based on Character Ontology, *Large-scale Knowledge Resources (LKR2008)*, LNAI, No.4938, pp.148–162 (2008).