Vol.2013-CH-99 No.7
2013/8/3

RDB と CMS を用いた アノテーション付与型画像データベースシステムの構築 −データ構造とインターフェイスの標準化を目指して−

和氣愛仁[†1]　永井正勝[†1]

筆者らが構築した古代エジプト語神官文字のパピルス画像データベースシステムでは，内部データストアとしてリレーショナルデータベース（MySQL）を，ユーザインターフェイスとしてオープンソース CMS（Drupal）を，それぞれ採用している．言語学的な文字・語情報を扱う本システムにおいて，このような構成を採用することの利点や拡張性について述べる．あわせて，今年度採択された科研費課題（基盤研究（C）25330395「アノテーション付与型画像データベースシステムのための汎用プラットフォーム構築」）において目指している，多種資料のデータベース化を視野に入れた，データ構造・ユーザインターフェイスの標準化の目論見について述べる．

The Construction of a Database System of Photographs with Attached Annotations Using RDB and CMS: Moving Toward Standardization of Data Structures and Interfaces

Toshihito WAKI[†1]　Masakatsu NAGAI[†1]

In the database system of photographs of ancient Egyptian hieratic papyri that they have developed, the authors have employed a relational database (MySQL) as an internal data store and an open source CMS (Drupal) as a user interface. For the main system, which uses linguistic information on glyphs and words, they will discuss the advantages and expandability of having adopted such a structure.
In addition, they will introduce an outline of their new research that is being conducted now with the aid of a Grant-in-Aid for Scientific Research (C) 25330395 “The Construction of a General Purpose Platform for a Database System of Photographs with Attached Annotations” and discuss their desire to achieve standardization in data structure and user interface by inserting a range in the conversion of many types of materials for the database.

1. はじめに

筆者らは,古代エジプト語の神官文字(ヒエラティック)を対象とした，言語情報システム（以下「本システム」とする）を構築した[1][2]．これは，三千年以上前のパピルス資料を高解像度デジタルカメラによって撮影し，その画像の任意の部分に多角形領域を定義して，その領域に文字・単語等のデータをリンクさせることで，画像上から直接，文字やテキストデータへアプローチすることが可能になっている（このようなシステムを以下「アノテーション付与型画像データベースシステム」とする）．本システムでは，内部のデータ保存のための仕組みとして，リレーショナルデータベース（以下 RDB とする）を採用した．以下本稿では，まず，筆者らが構築したシステムについて，対象とした資料の紹介およびシステムの技術的な側面についての紹介を行い，ついで，言語学的なデータを保存するために RDB を利用することのメリットについて述べる．さらに，今後の計画として，他の言語資料や，さらには非言語資料までを対象とした，より汎用的なアノテーション付与型画像データベースシステムの開発計画について述べる．

†1 筑波大学人文社会系
Faculty of Humanities and Social Sciences, University of Tsukuba

2. 資料について

2.1 筆記体研究の重要性

古代エジプトの地で文明が開化してからおよそ二千年間,ブロック体の文字の「聖刻文字（ヒエログリフ)」と筆記体の文字の「神官文字（ヒエラティック)」が，エジプト社会で使用されてきた．これらの文字のうち，人口に膾炙しているのは聖刻文字であるが，本研究では，世界的に見ても研究者数の少ない神官文字を対象としている．

本研究では，２つの理由から，神官文字を資料とした．その１点目は，神官文字の字形研究が世界的に停滞しているからである．たとえば，研究の基礎となる神官文字の字形リストは，今から百年以上も前に出版された文献[3]が，現在でも基本となっているほどである．[3]は，その当時に知られていた神官文字を時代順に並べて示した好著であり,それゆえ確かに便利なリストになってはいるが,今日では,主に，文字の認定・トレースの仕方・文字の番号付けという点において,再検討を要する箇所が少なからず存在する．そこで， [3]に代わるリストを作成するための前段階の作業として，本研究では，神官文字の字形データベースを作成することとした．

そして，神官文字を研究対象に選んだ２つ目の理由は，

エジプト学に蔓延している聖刻文字至上主義への代案を提示するためである．図 1 に示したように，古代エジプト社会には，聖刻文字で書かれた原資料[図 1-(a)]と，神官文字で書かれた原資料[図 1-(b)]とが存在していた．

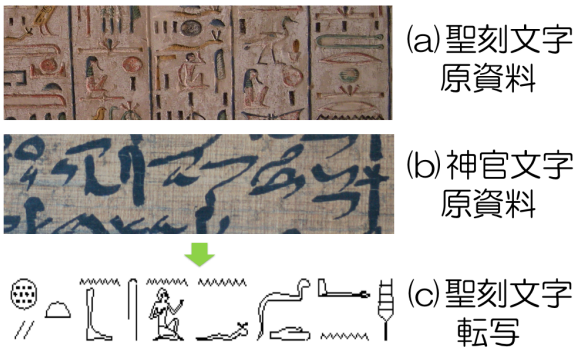

図 1　聖刻文字と神官文字の例

言うまでもなく，聖刻文字と神官文字はともに古代エジプト人の残した原資料の文字である．ところが，古代エジプト語の研究では，もう 1 つ，別の種類の聖刻文字が存在する．それは，神官文字資料を学者が聖刻文字に改めた聖刻文字転写 [図 1-(c)] である．

エジプト学の研究では，神官文字資料の写真やトレースが公開されずに，神官文字資料の聖刻文字転写が「資料」と称されて刊行されることが少なくない．ところが，聖刻文字と神官文字はそれぞれ異なる体系を形成する文字であるので，十全な意味で両者は 1 対 1 に対応しないのである．それゆえ，現代の学者が作成している聖刻文字転写は，原資料の聖刻文字を，近似していると学者が判断した聖刻文字に置き換えることによって作成されたものとなっている．このように，聖刻文字転写は二次的に作成された「代案物」でしかないはずだが，エジプト学の世界では，すでに述べたように，聖刻文字転写が「資料」として堂々と使用されているのである．筆者らは，このような状況に警鐘を鳴らす意味でも，神官文字を対象としたデータベースを作成することとした．

2.2 BM EA10221 ("Papyrus Abbott")

データベースに使用する資料としては，手始めに，大英博物館に所蔵されている神官文字パピルス写本 BM EA10221（通称"Papyrus Abbott"）の 1 点を扱うこととした．この資料は古代エジプト第 20 王朝時代に書かれた墓泥棒の裁判記録である（紀元前 1100 年頃）．本資料はエジプト学の研究で有名な歴史資料（史料）であり，エジプト史の概説書で頻繁に引用されている．その出版は[4]で行われているが，ここには神官文字の手書きトレースが掲載されているのみで，写真の添付はない．加えて，手書きトレースの宿命として，原資料の字形と細部において異なる部分があり，そのような違いがときに文字の解釈の違いを生じさせている．したがって，本研究によるカラー写真の学術的な公開は，本写本に対する世界初の試みとなる．

2.3 写真撮影

デジタル画像のデータベースを作成する場合，高精細な画像データを得る必要がある．画像サイズが大きいとデータ処理に時間を要するという問題点もあるが，字形の細部を確認するという学術目的を勘案すれば，元の画像はできるだけ高精細であるのが望ましい．そこで本研究では，中判カメラ Mamiya AFD III とマクロレンズ Mamiya Sekor Macro MF 120mm f/4 に，約 4000 万画素のデジタルバック Phase One P45 を装着して写真撮影を行った．その際，画像現像ソフト Capture One ver.5 を用い，MacBook Pro にカメラを連結させて撮影した．

なお，資料の撮影は，大英博物館学芸員の Richard Parkinson 博士の許可と協力のもと，2011 年 5 月に永井が実施した．

3. システムについて

3.1 概要

以下に本システムのスクリーンショットを示す．

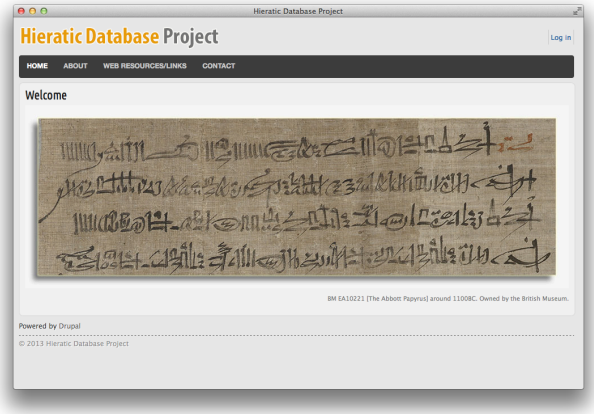

図 2　トップページ

図 2 はログイン前のトップページ画面である．大英博物館との申し合わせにより，現在のところ許可を得たユーザのみがパピルス画像およびデータにアクセス可能なようになっている．

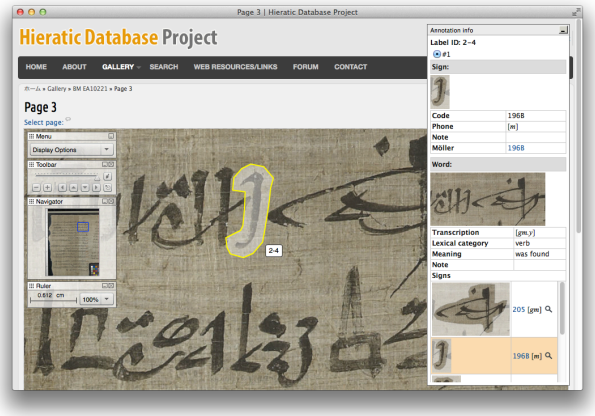

図 3　ギャラリー画面

図 3 はパピルス画像を拡大し，ひとつの文字要素をクリックして，その文字要素に関する情報および単語に関する情報を表示させたところである．単語中に含まれる文字から，その文字を含む別の単語一覧を検索することもできる．

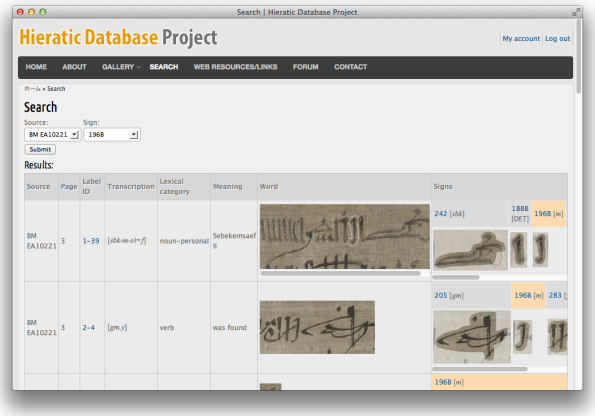

図 4　検索画面

図 4 は検索画面である．ここでも，単語中に含まれる文字から，その文字を含む別の単語一覧を検索することができる．現在のところ文字（文字コード）による検索しか実装していないが，およそ SQL により記述が可能なデータはすべて検索可能であるので，将来的には品詞や意味等による検索機能も実装する予定である．

その他本システムでは，研究者のためのディスカッションフォーラムや，ウェブ上のリソースへのリンク集等も実装してある．このあたりの機能は，Drupal（後述）を採用したおかげで非常に柔軟に構築が可能になっている．

3.2 データベース

本システムでは，言語学的なデータを保存するために RDB を採用した．具体的な実装として用いたのは MySQL である．これは，ひとつには，Drupal がデータベースを利用するということもあって都合がよかったということもあるが，それ以上に，言語学的なデータ構造設計に対する親和性が非常に高いということが大きな理由である．本稿執筆時点でデータベースに保存されているテーブルの数は 28 であるが，今後さらに増える予定である．データベースの構造については後で少し詳しく述べる．

3.3 ユーザインターフェイス

3.3.1 Drupal

本システムでは，ユーザインターフェイスの基礎部分を構築するために，オープンソース CMS（コンテンツマネージメントシステム）のひとつである Drupal[5]を採用した．Drupal は，極めて柔軟性に富んだ拡張モジュールシステムの採用によって，シンプルな構成のウェブサイトやブログサイトから，複雑な構成をもつウェブアプリケーションまで，様々なタイプのウェブサイトを高速に作成することができるようになっている．Drupal の一次配布元で公開されている拡張モジュールの数は，本稿執筆時点で実に 22000 を越えており，またサイトの見かけを設定する「テーマ」も 1700 を越える数のものが公開されている．これらを組み合わせることで，基本的にはプログラミングをすることなしに，様々な動的仕掛けを持ち，見た目にも美しいウェブサイトを構築できる．筆者らは，今後，古代エジプト語のみならず，古い時代の日本語や，非言語データについても同様のアノテーション付与型画像データベースシステムを構築することを計画しており，そうした目標からみても，システムの基本部分を標準的な手段により大きな手間をかけずに構築できることのメリットは非常に大きいと言える．

3.3.2 Zoomify

本システムでは，高解像度画像の拡大・縮小・スクロール操作，およびデータへのリンク部分の実装に関して，Zoomify[6]という画像処理ライブラリを採用した．Google Maps のようなものを想像すればわかりやすいだろう．Zoomify は商用ソフトウェアであるが，Enterprise Developer 版にはソースコードが付属しており，自由にプログラムを改編することができる．本システムでは Flash（言語としては ActionScript）版の Zoomify ビューアを用い，一部ソースコードに手を入れた上で利用している．前述の Drupal とあわせ，ユーザインターフェイスの基礎部分の構築にかかった時間は，約 1 か月である（実際に構築作業を行ったのは和氣ひとり）．このことからも，CMS やその他の既成ソフトウェアを活用することの有効性は理解されるであろう．

4. RDB の採用

4.1 言語学的データと RDB との親和性

ここでは，本システムが RDB を採用したことの理由として，言語学的データと RDB との親和性の高さということについて述べてみたい．

4.1.1 関係と集合

古代エジプト語は死滅した言語であり，文字にせよ文法

にせよ，現時点ですべてのことが明らかになっているわけではなく，例えばどういったものを「文字」として扱うかということさえも確定的ではない．漢字の場合にたとえていえば，漢字の部首を，独立した漢字として扱う可能性もあるようなものである．しかも古代エジプト語の場合，「文字のようなもの」が日本語や中国語のように必ずしも同じような大きさの矩形枠の中に収まるわけではなく，また並び順も，一条的なラインが一方向に向かって流れるように記述されるわけではない．しかし，そのような場合でも，実際の資料上における「文字のようなもの」同士が互いにどのように関係を持ち，どのように配列されているかということ自体は，RDB を使えば保持可能である．むしろ，データの線条性に拘束されることなく，データを集合として扱えることは逆に大きなメリットと言えるだろう．

同様のことは「語」の構造についても当てはまる．例えば「国立大学法人筑波大学」は，形態素解析における「長単位」としてはひとつの語と見なすのが妥当であろうが，しかしその内部構造を

(1) [[[[[国][立]][大学]][法人]][筑波大学]]

のように分析することは可能である．

このように分析してデータを保持しておけば，「大学」「法人」といった語を単位としたエンティティ（実体）に分解・整理した上で，それぞれの語と「国立大学法人筑波大学」との関係性を記述できるので，データの冗長性排除・メンテナンス性の確保という意味で非常にメリットが大きい．

さらに，多様な解釈をデータとして保持できることも極めて重要である．前述の通り，古代エジプト語は一度死滅した言語であるという事情から，文字にせよ文法にせよ，現在のところ研究者によって見解が異なる部分が多々あるが，こうした複数の見解について，データの個数の制約を受けずに，しかも必要に応じて各解釈の優先度を付与した上で保存できるというのは大きなメリットである．

このように，データを集合として扱い，かつ，集合内の要素に対して一対多あるいは多対多の関係を記述できることは，こうした研究において RDB を利用することの非常に大きな動機となり得る．

4.1.2 階層

従来，集合論的概念をベースとしてデータを扱う RDB においては，「階層」という概念はデータ的に表現しにくいもののひとつとされていた．現在まで比較的よく用いられているのは「隣接リストモデル」と呼ばれるデータ表現方法であるが，これは，各ノードが自分自身の「親」を保持しておくことによって，ツリー構造を表現するものである．以下に簡単な例を示す．図 5 は組織における上司 - 部下の関係を表している．

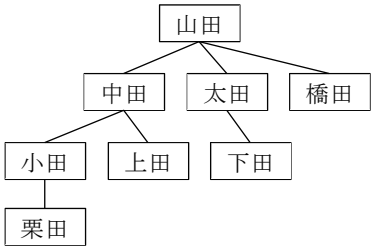

図 5　階層の例

名前	上司
山田	NULL
中田	山田
太田	山田
橋田	山田
小田	中田
上田	中田
下田	太田
栗田	小田

表 1　図 5 をテーブルの形で表したもの

隣接リストモデルによるデータ化の場合，基本的な SQL のみではツリー構造を再現することができず，外部プログラムによってループ処理する必要があり，また，データ更新の失敗により孤立したサブツリーが生じてしまうなど，いくつかの問題があった．そのような中，セルコ（[7]ほか）により提唱されたのが，「入れ子集合モデル」と呼ばれるデータ表現方法である．これは，ひとことで言えば，「階層を集合論的に表現する」データモデルということになる．以下に簡単な実例を示す．

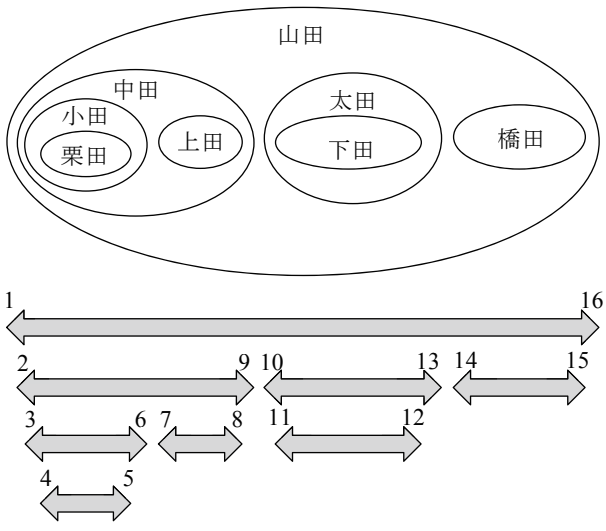

図 6　入れ子集合モデル

名前	lft	rgt
山田	1	16
中田	2	9
太田	10	13
橋田	14	15
小田	3	6
上田	7	8
下田	11	12
栗田	4	5

表 2　図 6 をテーブルの形で表したもの

この方式は，直感的に理解しやすく，また SQL もシンプルに記述できるというメリットがある．ただその一方で，新たなノードを挿入する際に，自分よりも右側にあるすべてのノードの座標値を更新しなければならないため，更新の影響範囲が大きく，パフォーマンス的な面で不利というデメリットがあった．そこで，入れ子集合モデルのメリットを残しつつ，更新の際のデメリットを大幅に軽減する方法が考案された．それが「入れ子区間モデル」[8]である．これは，基本的な原理は「入れ子集合モデル」とまったく同様であるが，集合の右端と左端をあらわす座標値を整数でなく実数値とすることにより，自分よりも右にあるすべてのノードの値を更新する必要をなくしたものである．以下, 特定の区間にノードをひとつ挿入する例を以下に示す．ここでは, 図 6 の上田に部下を配置することを考えてみる．

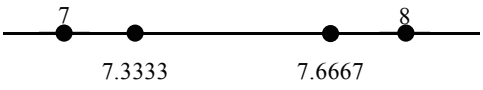

図 7　入れ子区間モデルによるノード挿入

新たに挿入するノードの左右の座標値は，以下のようにして計算可能である．挿入する区間の左側の座標値を lft, 右側の座標値を rgt とすると，定義により，

lft < rgt

したがって，

lft * 3 < lft * 2 + rgt
lft * 2 + rgt < lft + rgt * 2
lft + rgt * 2 < rgt * 3

以上より

lft * 3 < lft * 2 + rgt < lft + rgt * 2 < rgt * 3

各辺を 3 で割り，

lft < (lft * 2 + rgt) / 3 < (lft + rgt * 2) / 3 < rgt

ゆえに，

挿入ノードの左座標 = (lft * 2 + rgt) / 3
挿入ノードの右座標 = (lft + rgt * 2) / 3

この関係は常に成り立つので，ノードを挿入する際は，あたらしいノードの左座標と右座標のみを，上記計算式により挿入区間の左右の座標値から算出すればよい．これはテーブル内のレコード数の影響を一切受けないので，パフォーマンス的に非常に優れている．また，後から挿入できるノードの数は，理論上は無限である．ただし実際には計算機上の浮動小数点の有効桁数の上限に依存するが，それでも事実上まったく問題ないといって良いであろう．

言語学的見地からいうと，文は複数階層からなる深層構造をもつということは常識であり，「階層」という概念は，言語学的なデータを扱うにあたって欠くことのできないものである．またその階層の深さは，分析の際に依って立つ文法理論にもよるが，いずれにせよ可変であり，場合によってはかなり深い階層構造を想定せねばならないこともある．例えば，いささか極端な例ではあるが，

(2)　太郎が病院で会った花子が食べていた食事を作った山田さんの奥さんは美しい

のように，非常に複雑な埋め込み構造をもつ文は（発話をリニアに聴取したときの解釈の困難さはおくとしても）原理的に生成可能である．このような文の分析において，中間階層の要素（例えば「花子が食べる」の部分）をどのような用語で呼ぶか（どのような概念として定義するか）という言語学的問題はあるにしても，RDB を用いれば，文全体の階層構造をデータとして記述することは可能である．もし仮に，理論上さらに多くの中間階層の認定が必要になったとしても，RDB 上で上記「入れ子区間モデル」を利用すればあとから中間構造をデータ的に追加することも容易であり，階層構造を非常にスマートに RDB 上で扱うことが可能になる．

4.2 本システムのデータ構造

ここで，本システムのデータベース構造について簡単に説明する．まず本システムの ER 図を以下に示す．

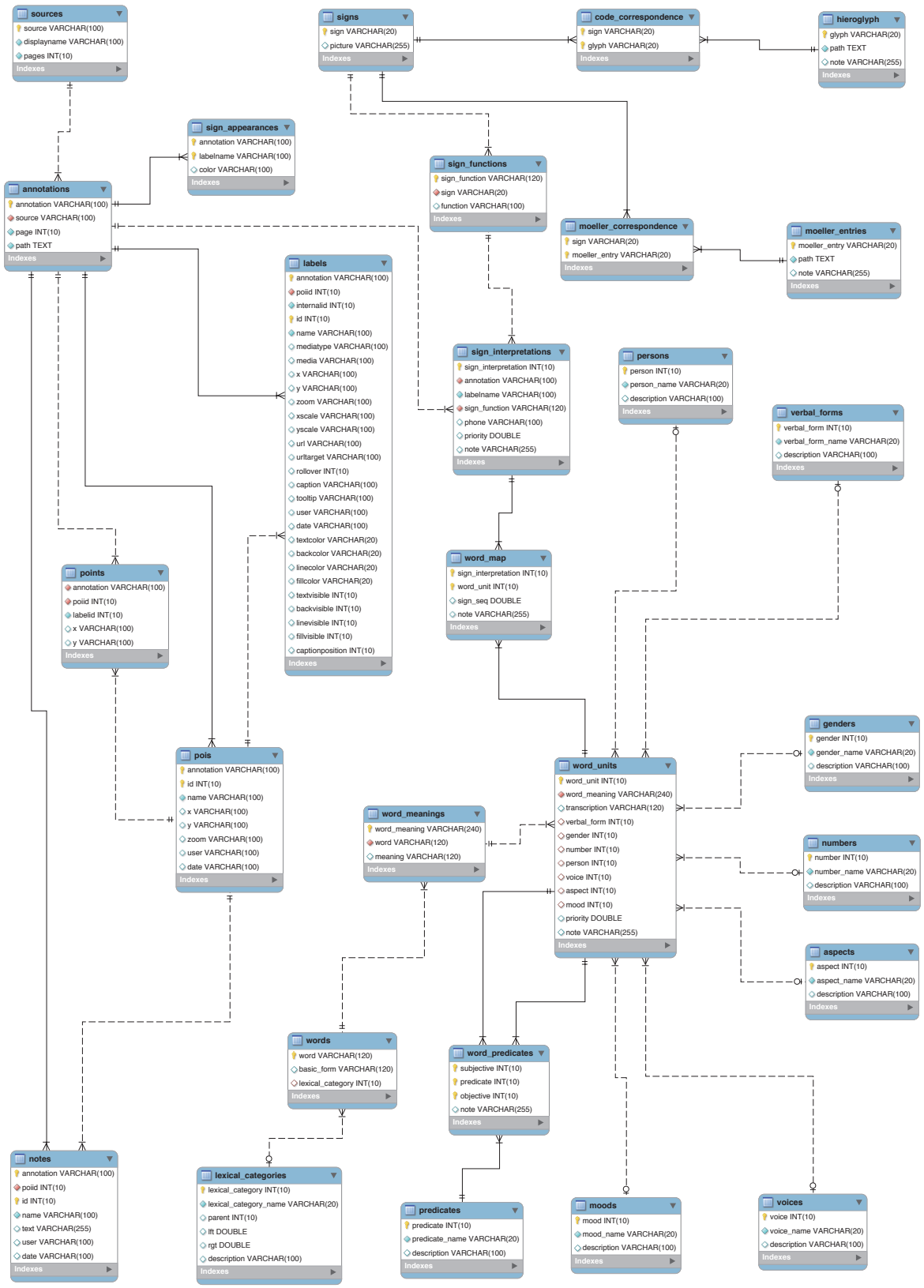

図 8　ER 図

以下，重要と思われるポイントにしぼって説明する．

ひとつの字形に対する複数の文字機能の対応：

signs / sign_functions

古代エジプト語の場合，ある字形が表音文字として使われる場合と，限定詞（Determiner）として使われる場合がある．こういった状況に対応しつつ，同一の文字であることをデータ上保証するため，文字と文字機能を別テーブルに保存している．

実際の資料上の字形要素に対する文字機能の解釈：

sign_interpretations / sign_functions

すでに述べたとおり，古代エジプト語の場合，文字の判読自体が自明なことではなく，研究者によってその解釈が揺れる可能性がある．こうした複数の解釈可能性を排除することなく保存できるように，資料上の字形要素に対して単一の文字機能を結びつけるのでなく，複数の文字機能を「解釈」として保存するようにしている．またその際，おのおのの解釈について優先度を付与できるようにしてある．

ひとつの単語に対する複数の意味の対応：

words / word_meanings

どのような言語であっても，当然のことながら，ひとつの語に対して複数の語義が対応しうる．そのような状況に対応できるよう，単語と語義を別のテーブルに保存している．

実際の資料上の字形要素のまとまりに対する単語の意味の解釈：

word_units / word_meanings

上述の文字機能の解釈の場合と同様に，単語についても，その語の解釈が揺れる可能性がある．そのような状況に対応できるよう，字形要素のまとまりと語義との対応表を用意している．また，字形要素のまとまりに対する複数の解釈について，解釈の優先度を付与できるようにしてある．

実際の資料上の文字機能の解釈と字形要素のまとまりとの対応関係：

word_map / sign_interpretations / word_units

複数の文字（正確には文字の機能）がひとつのまとまりを構成する状況を保存するためのテーブルを用意してある．文字あるいは語の解釈の揺れにより，単語としての区切り位置もずれる可能性があるが，そういった状況にも対応できるような設計としてある．

現実の資料を対象とし，発展途上にある研究の状況を考慮した場合，「複数の解釈可能性」をそのままに保持できるということは極めて重要である．そのことを考えたとき，RDB が持つ「集合と集合の関係」を保持するという特性は，このような研究に対して非常に高い親和性を持っていると言うことができる．

なお，本システムでは，上で述べた「入れ子区間モデル」によるデータ定義は，今のところ，品詞テーブルについてのみ実装している（利便性を向上させるため，隣接リストモデルによる階層定義も併用）．以下実データを示す．

品詞 ID	品詞名	parent	lft	rgt
1	ROOT	NULL	1	80
2	adjective	1	2	3
3	adverb	1	4	5
4	article	1	6	13
5	definite	4	7	8
6	indefinite	4	9	10
7	possessive	4	11	12
8	auxiliary	1	14	15
9	converter	1	16	25
10	adnominal	9	17	18
11	adverbial	9	19	20
12	focalization	9	21	22
13	preterit	9	23	24
14	demonstrative	1	26	27
15	interrogative	1	28	29
16	negator	1	30	31
17	noun	1	32	39
18	common	17	33	34
19	personal	17	35	36
20	place	17	37	38
21	number	1	40	45
22	cardinal	21	41	42
23	ordinal	21	43	44
24	particle	1	46	51
25	non-enclitic	24	47	48
26	enclitic	24	49	50
27	preposition	1	52	53
28	pronoun	1	54	61
29	independent	28	55	56
30	dependent	28	57	58
31	suffix	28	59	60
32	verb	1	62	71
33	strong	32	63	64
34	weak	32	65	66
35	double	32	67	68
36	irregular	32	69	70
37	cartouche	1	72	73
38	punctuation	1	74	79
39	dot	38	75	76
40	verse point	38	77	78
41	genitive	2	2.3333	2.6667
42	preformative	28	58.3333	58.6667

表 3　品詞テーブル

語形態素同士の階層性については，現在のところ未実装であるが，今後，関係するテーブルの定義を修正して取り入れていくことを計画している．これにより，「語が語を構成する」ということが階層数の制限なくデータ的に表現できるようになる．

4.3 XML との連携

現在人文情報学において重要なデータ保持・交換の手段となっている XML についてもここで触れておく．基本的なデータを RDB で保持しておき，必要に応じて RDB から XML を出力することは難しいことではない．むしろ，そのような設計としておいた方が，データ構造の設計をより柔軟に行えるし，最終的に出力する XML の書式の変更にも対応しやすいだろう．本システムにおいても，RDB のデータ構造を拡張した上で，最終的には TEI 準拠の XML を出力する機能を実装する予定である．

4.4 他資料への展開の可能性

ここまで古代エジプト語の画像データベースシステムについて述べてきたが，最後に今後の計画として，他の言語・非言語資料への展開可能性について触れておきたい．本システムのデータベースは，古代エジプト語の画像資料を対象に設計を始めたものではあるが，より長期的には，日本語やその他の言語資料にも対応できるようなデータ構造に拡張していく予定である．またさらには，言語以外の資料，例えば図版や，さらには建築物等の立体構造物の画像を扱うようなシステムへと発展させていくことも計画している．今年度採択された科研費（基盤研究（C））「アノテーション付与型画像データベースシステムのための汎用プラットフォーム構築」（研究代表者：和氣愛仁）では，このことを目的として，まず今年度，明治時代の日本の国文典（日本語文法書）の画像データベースシステムを構築する予定である．言語学的な研究に適したデータ構造と，例えば建造物を扱うデータベースのデータ構造とでは，設計を変えなければならない点も生じるだろうが，できるだけ汎用的なデータ構造を，できるだけ少ない数設計することを目標に据えている．またユーザインターフェイスの点では，前述の Drupal および Zoomify を利用することで，基本的に同じ操作性を持つシステムを構築できるだろうと考えている．RDB をデータ保存の基礎とし，ユーザインターフェイスを CMS やその他の既成ソフトウェアに任せることで，こうしたデータベースシステムの構築をかなり高速に行うことが可能になるだろう．

5. おわりに

以上，本稿では，特に言語学的なデータを扱う画像データベースシステムにおいて，RDB を利用することのメリット，および，システム開発における CMS 等の既成ソフトウェアを利用することの有効性について述べた．ここで得た知見を活用し，今後のプロジェクトにおいて有意義な成果を得たいと考えている．

参考文献

[1] 永井正勝・和氣愛仁「古代エジプト神官文字写本を対象とした言語情報表示システムの試作」『つながるデジタルアーカイブ―分野・組織・地域を越えて』人文科学とコンピュータシンポジウム（じんもんこん 2012）論文集，2012(7)，pp.225-230，情報処理学会，2013
[2] http://hdb.jinsha.tsukuba.ac.jp/
[3] Möller, G: Hieratisch Paläographie, 3 Vols., Leipzig, 1909-1912.
[4] British Museum: Selected Papyri in the Hieratic Character from the Collections of the British Museum, Vol.2, Manchester, 1930.
[5] http://drupal.org/
[6] http://www.zoomify.com/
[7] J.セルコ（秋田昌幸訳）『プログラマのための SQL』第 2 版，ピアソン・エデュケーション，2001
[8] http://gihyo.jp/dev/serial/01/sql_academy2/000601