CTSデータとデータベースデータとをつなぐコラボレーション
―「古典人名・人物年表データベース」プロジェクトから―

国文学研究資料館
相田 満

The collaboration what connects CTS and database data.
― From "a Classic Person's Name and a Character Chronological Table Database" Project ―

AIDA,MITSURU
National Institute of Japanese Literature (NIJL)

概 要

組み版出力を行うことを前提に開発されたCTS（Computerized Typesetting System）システムを利用して入力作業を進めることには、(1)安定した文字セットを使用して、データ入力が行いやすいこと、(2)組み版機能を併用することによるマーキングの効率化が図れること、(3)データ処理が行いやすいなどの点で、大量のデータ、構造の複雑なデータを数段階に分けて入力する際に、さまざまな利点がある。

Outline

CTS（Computerized Typesetting System） was developed for composition printing.We receive profits at various points by use of this system.For example, the following occurs.

(1) We can use the stable character set and perform data input.

(2) We can use a composition function together. By that, marking work can be done efficiently.

(3) We can process data easily...etc.

Thus, the outstanding performance is demonstrated when treating a lot of data and complicated structure.

1.印刷業者とのコラボレーション

大量データの電子化・データベース化を行うためには、データの入力作業を専門業者に委託する方が効率的である場合が多い。

入力委託が可能なデータには、文字データのほか画像・音声など、さまざまなものがある。これらの内、文字データの入力を専門に扱う業者としては、データ入力だけを専門に扱う入力専門業者と、本来は印刷を行うことを目的としていたシステムを使用して、電子化入力作業も行うようになった印刷業者との二種類がある。

電子出版などのデジタルコンテンツの比重が高くなってきている昨今、両業者の職掌と境界は、次第に曖昧になってきてはいる。しかし、デジタルコンテンツの作成システムや方法の面から両業者を比較すると、前者の入力専門業者による入力作業は、汎用の機器と規格に則った入力作業が中心に行われるため、比較的単純な仕様のデータを入力する場面に向いており、小ロットのデータ入力では、印刷業者よりも経済的に上がる場合が多いという利点がある。

一方、後者の印刷業者による場合では、組み版出力を行うことを前提に開発されたCTS（Computerized Typesetting System）システムを使用しての入力作業が行われることが可能と

なる。そのため、複雑な仕様で、多段階にわたってデータ形成を行うことが可能で、しかも、大量のデータを安定的に運用・管理することができる。

もっとも、印刷業者のシステムと運用ノウハウを利用して、印刷物ではなく、データベースのデータ自体を入力・形成しようとする事例は、一般にはそれほど多くはない。しかし、報告者が構築を進めているデータベースプロジェクトでは、印刷業者の CTS システムを利用した入力作業が行われており、効率的なデータ形成・管理のための工夫・開発も進められている。

こうした工夫は、必ずしも一朝一夕に出来上がるものではない。しかしながら、データ形成のために互いに可能なことを模索する中で、ノウハウを積み上げてきた事柄は、まさに印刷業者とのコラボレーションによるたまものといっても過言ではあるまい。

2.CTS（Computerized Typesetting System）システムとは

CTS（Computerized Typesetting System）は、もともと日本語組み版を前提として生み出されたシステムである。その発達をうながしたのは、大部数の即時作成を要求される新聞制作であった。CTS システムの発達により、コンピュータを利用しての、集配信から、組み版、紙面出力までを行う一連のシステムが構築され、出版というマスメディアにおいても、放送メディアにも遜色のない即時性と大量配信を可能とた。また、頻繁に生じる情報の加除・修正作業の効率化と管理の必要性からは、バッチ処理による自動化、さらにはデータベース・パブリッシングへの道を拓くことにもなった。

電算写植システムとも呼ばれるこのシステムは、日本では 1970 年に初めて開発された。しかも、厳格で複雑な日本語組み版の決まり事を十分に反映させるため、欧文環境で発達した DTP システムとは異質で独自の機能を盛り込みながら、現在に至っている。

当初、CTS の多くは閉じられた、専用データ環境ではあった。ところが、ここ数年、相当のオープン化が進み、パーソナルコンピュータのビジネス系ソフトのデータも、問題なく受け取れるようになっている。しかも、JepaX(ジェパ・エックス)[*1] のように、W3C 規格として勧告された XML1.0 との互換性に配慮された文書型定義ファイル(DTD)も策定されている（1999.9)。

このように、CTS は次第に開かれた存在となりつつはあるものの、日本語組み版特有の規則を処理するために拡張された機能と仕様は、XML などの標準規格となったものに対して、今なおオーバースペックの状態にある。

しかしながら、入力作業すなわち電子化コンテンツの生産を効率的に進めることを助ける技術の歩みはゆるやかで、XML1.0 に準拠した入力ツールさえも現在では十分に整っていないのが現状である。

3.CTSシステム利用による印刷業者とのコラボレーションの利点

国文学研究資料館で進められている大規模データベースコンテンツのひとつに「古典人

---

*1 日本電子出版協会（JEPA）と JEPA 出版データフォーマット標準化研究委員会が策定した、出版データフォーマット標準化交換フォーマット（http://www.jepax.org/）

名・人物年表データベース」がある。

これは、近代以前の日本の人物に関するさまざまな基礎史料をデータベース化するもので、『平安人物志』や『皇代記』のような小規模なものから、『公卿補任』や『尊卑分脈』のような大部なものまで、原本資料ことごとくデータベース化し、さらにそこに記載される人物の詳細情報を、別に蓄積される人物画像とリンクさせようとするものである。

現在、30 種類をこえる各資料の電子化を終了しているが、テキスト情報のソースとなる個々の史料の特性・形態はさまざまである。なかには、『尊卑分脈』や『本朝皇胤紹運録』のような系図史料のように、特別なデータベースシステムを使用しなければ、その資料の特性を十分に発揮できないものも存在する（相田満「日本古典系図データベースの構築」2001-CH-51：系図データベースは国文学研究資料館内公開中）。

報告者が印刷業者とのコラボレーションによって、CTS システムの潜在力を引き出しながら進めているデータベース・プロジェクトは、このような資料を対象としている。

では、CTS システムを擁する印刷業者に資料の入力委託を行うことにどのような利点があるだろうか。それにはおよそ次の 3 点が考えられる。

(1)安定した文字セットを使用して、データ入力が行いやすいこと。

(2)組み版機能を併用することによるマーキングの効率化

(3)データ処理

以下、各項について、データベースプロジェクトにおける実例と照らし合わせながら詳述したい。

3.1.安定した文字セットと文字管理

JIS 制定以前に印刷システムの電算化を果たした大手印刷業者では、独自の文字セットを運用している場合も少なくない。JIS 漢字コードとの情報整合性は、あくまで各ユーザ毎に定義されるコンバート表によって確保している。

印刷業者で使用される文字セットは、あくまでユーザ本意に規定されている。そのため、基本的には異体字を体系化する文字シソーラスは作成されず、あくまでユーザの要望によって外字が天井知らずに作成される。

しかしながら、周知のように JIS 漢字は、必ずしも安定したものではない。規格改訂により、時には字形変化を生じてしまうことも少なくないのである。たとえば、JIS 漢字の情報交換用符号文字は、JIS X 208 系の文字（とくに漢字）に対する改変を見ても、いわゆる 78JIS から 83JIS、さらには 90 年に若干の追加を経て、97 年に包摂概念などの整備によりやっと字形概念に安定を見たかとも思えたのも束の間、常用漢字外の文字は「いわゆる康熙字典体」の字体を情報処理機器にても使用される「印刷標準字体」として採用するべきとの 2000 年の国語審議会答申を反映させることを余儀なくされた。そのため、従来の「日本における一般的・日常的に頻用される最小限の文字集合」としての JIS X 208 という文字体系は、構成文字の字形・文字種の意義などの根幹的な面で、少なくとも 20 年間の間に 3 度の大きな変更を受けている。そのため、厳密な意味では活字情報の忠実な電子的保全には向いているとは言いがたいのである。それ程のタイムスパンで比較するならば、印刷業者の CTS で管理される文字セットの方に委ねて情報を保存した方が、原データの字形を安定的に保全することになるというのも皮肉なことである。

報告者のプロジェクトでは、印刷所の保有するフォントセットを使用して、原本の字形になるべく忠実に入力され、紙面に出力されたデータに対して校正作業を進めている。なるべく原本の出力に近似した文字セットを印刷することにより、校正者の能力と判断のばらつきをなるべく回避するためである。

3.1.1CTS用データと情報用交換符号文字（JIS漢字）とのコンバート処理

CTS 用の文字体系で打ち込まれたデータと情報用交換符号文字（JIS 漢字）とのコンバートは、以下の要領で行われる。

①典拠外文字データの報告（受注者側）［図①］

まず、入力依頼を行ったデータ中に含まれる情報用交換符号文字への変換を指定するべく、業者より［図①］の報告書が提出される。

国文学研究資料館「ユニコード変換不可リスト」. 公卿補任　2002年 5月10日　1頁

メンテキィ	[3]外字	[5]凸版コード	[7]ユニコード	[4]外字を含む項目データ
000010	[3]靜	[5]000005F5	[7]	[4]法名靜覺＠號枇杷左大臣
000020	[3]暲	[5]00001F8A	[7]	[4]〈保延6年／11月／4日／1140／敍從五位上（自小六條行幸上御門皇后日。暲子内親王職事）。〉

図①　変換文字指定用リスト（受注者側）

この文字リストは、基本的に CTS システム内にて使用される文字セットと、UCS2.0 日本語面に対する典拠外字との変換不可リストとなっている。

たとえば、［図①］の１字目は、「静」の旧字「靜」の偏が「青」ではなく、「青」の字形のため、情報用交換符号文字には含まれない異体字に扱われている。２字目は、旁の「章」の字形が異なるため、リストに挙がっている。

②典拠外字変換指示リストの作成［図②］

情報用交換符号文字への変換は、［図①］のリストに対して、発注者（ここでは報告者）の判断により指示が行われる。現在の所、作成済みの変換指示リストは約 1,000 字である。[*1]

データ名	種別	I.D.	凸版コード	文字	UCSコード	代替文字	代替UCSコード	縮約文字	縮約UCSコード	備考	文字鏡備考
系・地		000010	000001C2	¢	FFE0						文字鏡番号は62081
和			000003D7	醴	7067	醴	7054				
系			000004D7	楳	6843						
系・地		000020	000005F1	廣	5EE3			広	5E83		
系			000005F4	葛	845B						
系			000006E8	迩	8FE9						
系			000006ED	珎	73CE			珍	73CD		
地			00000876	□	25A1						
地			00000877	□	25A1					丸みの付いた口	
系			00000B8A	幷	5E77	并	5E76				
系			00000E4A	:	003A					半角文字	
系			00000F4E	.	002E					半角文字	
系			0000103A	㈡		（う）	FF08／3046／FF09			3文字化	
系			0000103B	㈢		（え）	FF08／3048／FF09				
系			0000118E	䆉	7AC9						
地			000011B5	郤	90E4						

図②　典拠外字変換指示リスト（発注者側）

*1『岩波古典文学大事典』（要語のみ）、『尊卑分脈』、『群書類従系図部』、『古事類苑』（索引）、『地下家伝』などに出現する典拠外字数。

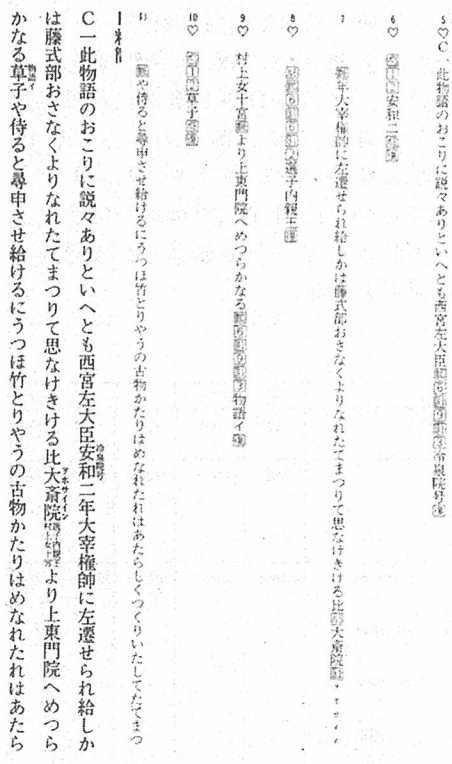
③変換不可字の管理［図③］

図③ 作成された外字

資料の性質上、どうしても典拠不明の字が発生する。しかも、人物・人名に使用される文字の場合には、他字を以て代替し難いものも少なからず出現する。そこで、本プロジェクトでは、CTS システムによる文字セットにおいては、テキスト情報を電子的に完結させるべく、外字（図③）を作成している。このことにより、全てのデータに対して、コーパスの作成など、さまざまなテキスト処理が可能となるからである。

作成される前項の典拠外字指示リスト所収字および、作成される外字には、今昔文字鏡やＧＴ書体などの大規模文字セットにも未収録のものも少なくない。

3.2.組み版機能の利用



図④ 組み版原稿とソース（『河海抄』）

CTS 組み版ならではの機能も利用すると、傍線・ルビ・割り注、さらには表組みなどのレイアウト上の特殊情報を、見たままに出力させることも可能となる。

たとえば、［図④］は、ルビ・傍記・割り注などについて、原資料の体裁に準じた形で入力したものである。本データについては、この後、年月日、人物などの情報についてはさらにマーキングを施すことを予定しているが、その場合、校正者へは、種類の異なる傍線を関係する箇所に付記してもらい、入力者が、その指示通りに傍線を入力することとなる。

入力作業は、現在では WYSIWYG（ういしうぃぐ）［What You See Is What You Get.］のモニタにより、原稿に見たままの通りの入力が可能となっており、かつてのようにすべてソース（図④右）を相手にしながら制御コードを入力するというものではなくなっているようである。

しかし、それでもソースに直接ふれなくてはならない場合もあり、ごく稀にデータ修正時にコーディングミスによる文字化けの発生もある。

その後、それぞれの制御コードおよびタグは、予定されたデータベースのタグやデリミタなどの情報に置換される。

このような手法の採用は、データ校正者への負担軽減にも大いに寄与しており、情報の正確さを確保する面で、効果を上げている。

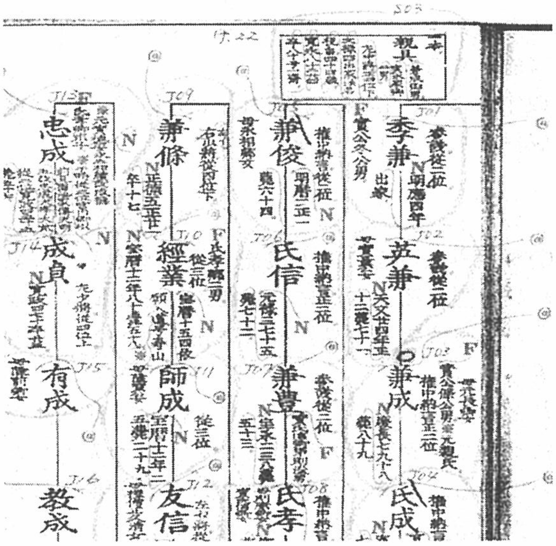
3.3.データ処理

CTS システムにおいては、前述の組み版制御のかかったデータに対してもプログラム処理が可能である。たとえば、文字置換はおろか、抽出、ソーティングなど、テキストだけで数百メガバイトという大規模データに対しても安定したデータ処理が可能となる。

たとえば、系図データベースを構築する際、初期原稿の体裁［図⑤］が複雑なため、データ入力・形成作業を、数段階に分けて行う必要がある。

その際、行われた処理には、たとえば次のようなものがある。

①句読点をデリミタとしてのレコード分割［図⑥］



図⑤ 系図データベース用入力原稿

②年月日情報の切り出しと西暦付加［図⑦］

③レコード分割の際の管理番号付加［図⑦］

などである。

あるいはまた、テーブルを利用しての読みの付加作業なども行われる。

これらの処理は、PC が高性能化したとはいえ、なお高負荷が要求される処理である。ましてや、前述のように、JIS では収まりきれない文字を多数含むデータに対して、汎用 PC で処理を行うことは、まださまざまな困難がつきまとうと言わざるをえない。

H	I	K	J	L	M	N	O	B	D	A	C	E
データID	嫡庶記号	右脇書	名	名ルビ	分書1	分書2	参考頭注	叢書名	編次	編者名	部立	資料名
え 07. 03	○○○		中臣 @常 盤大 連公		FX黒田大連公一男。 BS始賜中臣連姓仍為祖。			群書類従	巻第六十二	検校塙保己 一集	系譜部三	中臣 氏系 図
え 07. 14	○	一	@頼 基		KX祭主。 KX正四位下。 KX＼少祐。 KX＼権大祐。 KX＼権少祐。 KX＼大副。 NE（天暦10年／／／／）天 暦十年#卒。七十三。		NK（延喜22年／／／／）延喜廿二年#正 六位上。任少祐。 NK（延長5年／1月／／／）延長五年正月 #転権大祐。 NK（承平3年／1月／／／）承平三年正月 #権少祐。 NK（天慶3年／10月／7日／／）天慶三 年十月七日#補祭主。	群書類従	巻第六十二	検校塙保己 一集	系譜部三	中臣 氏系 図

図⑥　系図データベースの形成(1)

H	I	K	J	L	B	D	A	C	E
データID	嫡庶記号	右傍書	名	名ルビ	叢書名	編次	編著名	部立	資料名
え 07. 03	○○○		中臣@常盤大連公		群書類従	巻第六十二	新校群書類従一集	系譜部三	中臣氏系図
え 07. 14	○	一	@頼基		群書類従	巻第六十二	新校群書類従一集	系譜部三	中臣氏系図

①	②	う	え	お	か	あ	い	き	く	け	こ
第1階層	第2階層	嫡親 (1人前) ID	子 (一人後) ID	兄姉 (1人前) ID	弟妹 (1人後) ID	原夫婦ID	行夫婦ID	性別	姓見出	名見出	氏系図
		え 07. 03	え 07. 04						中臣	常盤連大公	中臣
		え 0713	え 07. 15						中臣	頼基	中臣

H
データID
え 07. 03
え 07. 14

	H	I	K	J	L					M
連番	データID	和暦年	閏	月	日	西暦	小連番	位置	種別	分書1
10	え 07. 03								FX XX	黒田大連公一男。
20	え 07. 03								BS XX	始賜中臣連姓初祖。
30	え 07. 14え						10		KX XX	祭主。
40	07. 14え 0						20		KX XX	正四位下。
50	7. 14え 07.						30		KX XX	\少祐。
60	14え 07. 1						40		KX XX	\権大祐。
70	4え 07. 14						50		KX XX	\権少祐。
80	え 07. 14え						60		KX XX	\大副。
90	07. 14え 0	延喜12年				0922	70	頭書	NK XX	延喜廿二年#正六位上、任少祐。
100	7. 14え 07.	延長5年		1月		0927	80	頭書	NK XX	延長五年正月#転権大祐。
110	14え 07. 1	承平3年		1月		0933	90	頭書	NK XX	承平三年正月#権少祐。
120	4	天慶3年		10月	7日	0940	100	頭書	NK XX	天慶三年十月七日#補祭主。
130		天暦10年				0956	110	分書	NE XX	天暦十年#卒、七十三。

図⑦　系図データベースの形成(2)

4.まとめにかえて

大量のデータを入力するプロジェクトは、巨大艦船の操縦にもたとえられよう。そこにでは意志の疎通を明確にし、確実な操作が求められる。

データ入力に取りかかる際には、データ構造を適切に解析することと、データを効率的にかつ確実に入力・形成するめの工夫が不可欠である。さらに、それがデータベースデータである場合には、入力されるデータの正確さはもちろんであるが、データ構造の適切な解析と構造化ということも、同様に重要である。とくに多人数を動員する場合、入力規則をより正確にかつ平準化して伝達するとともに、各作業工程の判断分岐をできるだけ最小限にとどめたものを、多数回繰り返すことにより、各工程のチェックも容易にせしめる工夫が必要である。

しかしながら、このような切実な要求があるにもかかわらず、多くの大規模データベースが実際に作成される現場では、多量の付加情報が付加された、あたかもプログラム言語ごときソースデータを相手にしなければならない場面が多い。

これは、データシート作成やデータベースデータ校正の点で大きな障害となる。

報告者が進めるデータベース構築は、毎年 200 万字から 300 万字のデータを常に発生させている。その際、入力用データシート作成時から、直接データタグを入力シートに記述した原本、あるいは原本の構造を解釈した仕様書を作成し、入力者（業者）は、そのシートと仕様に従った情報仕様・タグを付加して入力する手法を採用している。入力用データシートに付加されるタグには、あらかじめ構造化を念頭に置いた体系に従ったフラグが示

されており、それを一括処理作業をかけながら、規則的なデータベースデータを構築するものである。

しかし、多くのプロジェクトがそうであるように、データベース作成作業の過半は、データシート作りと、データの検証作業に費やされる。そして、その作業の従事者の多くは、限りなくプレーンな状態に近いテキストに付与される様々な付加情報(タグ・マーキング)と格闘することとなる。

構造化言語 SGML は、文書の本文情報と体裁情報を分離することから発想され、その制定当初の 1980 年代には、印刷・組み版用の言語として注目を浴びた。

その後、90 年代に入ると、HTML、XML、さらにはそこから派生して、携帯電話などに特化した WML、3D グラフィックスのようなマルチメディアを志向する VRML(Virtual Reality Modeling Language)など、データベースやブラウザとの親和性の考慮された構造化言語は、データ交換プロトコルとして長足の進歩を遂げているといえる。

それにくらべて、入力作業すなわち電子化コンテンツの生産を効率的に進めることを助ける技術の歩みはゆるやかである。

歴史の長い CTS システムは、必ずしもデータベース化を念頭に作られているとはいいがたい。これは、おもに「表現」の工夫を追求することに特化したシステムであるためである。また、文字コードについても、独自の文字セットが採用されており、必ずしも JIS 規格文字に準拠した入力が行われるわけではない。

ダウサイジングが既定の潮流となった今、重厚長大な観のある CTS システムは、専門企業・業者にのみ採用されるもので、一般には DTP がその役割にとって代わるものであるかもしれない。その意味で、本プロジェクトのような入力・データ構築手法は、異質であるかもしれない。

しかし、本報告で紹介したことがらは、いずれも印刷業者との対話の中から引き出されたものであり、ＣＴＳシステムの潜在力を引き出しながら次第にノウハウを蓄積してきたものである。

本データベースプロジェクトを通じての諸事例が、類似のデータベース形成プロジェクトにおいて、何らかの一助になれば望外の幸せである。