

データベース作成の一事例 ～電気化学データベース～

横浜国立大学工学部 有澤 博
仁木克己

はじめに 近年の科学技術の発達は、その成果として膨大な量の学術情報の生産とつながっている。これに対して、集大成した学術情報とコンピュータの大容量記憶装置上におき、データベース管理システム (DBMS) とデータ通信システム (DCS) を用いて、データの迅速な流通をはかろうという計画が、所々で出はじめている。このようなものを一般には学術情報流通システムとよぶ。

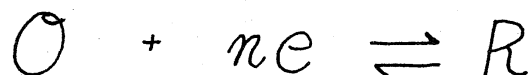
著者らは現在、文部省科学研究費「データベース作成」の助成のもとに、電気化学における電極反応パラメータのデータベース化を進めている。しかし、このような活動は、我国では先例が少なく、実際に行なってみると、多くの問題点や、新たな課題がつかえることも多かった。本稿ではこれら諸問題を整理して報告したい。

電気化学データベースの内容

電気化学データベースは次の4つの構成要素から成る。

1. 電気化学の電極反応速度パラメータ
2. 電解質溶液の電気伝導度 (伝導率) データ
3. 電解質溶液の活性度係数 (活動度) データ
4. 1~3に含まれるデータのソースレファレンス情報
(文献データファイル)

金属の電解製錬, 精製, ソーダ電解など, 産業の基幹となる電気化学工業は, 電力消費型の産業で, これらの反応装置を設計する場合, 最適条件 (省エネルギーを省めて) を求めるために, 上記のものは基礎的な必須データである。本稿でとり上げるのは, このうち1の電極反応速度パラメータであり, 電気化学データベースの中でも, 最も重要なデータである。反応速度パラメータとは電極反応



(活性物質)

において, 活性物質の種類はもちろん, 電極の種類, 支持電解質 (反応物質ではないが, 電子移動に寄与する物質), 溶媒, 添加物, 温度等, 様々な条件によって測定された平衡点に関するパラメータで, 転移係数 (Transfer Coefficient) と反応速度定数 (Rate Constant) が重要なものである。(1)

電気化学データベースのオリジナル・ソースは、世界各国で
 発行される学術雑誌に掲載された論文である。これらの文献
 点数は膨大な数にのぼり、前記の本質的諸条件の他にも、
 測定方法に関するアイデアや、表現、評価方法などによって
 非常にバラエティに富んだものになっている。

現在、国際化学会連合 (IUPAC) の正規の依頼を受け
 て、電気化学反応データ収集の国際センターとしての役割を
 担うと同時に、文献や単純な数値情報でない、複雑な学術情報
 のデータベース作成活動のあり方について、模索を続けてい
 る段階である。

データベースの作成手順

本稿では、学術情報の収集と提供に関する一般的な議論は
 他の文献⁽²⁾にゆき、電気化学データの収集とデータベース
 化の具体的な手順に焦点を絞って述べる。現在までに行なっ
 てきた、あるいは今後大膽に実行に移されることになっ
 ている手順は次の通りである。

- (1) データの解析。電極反応データのデータ構造の解析
 およびこれを利用する側の要求の解析。これには、言
 うまでもなく化学の専門家による分析が必要である。
 しかしそこから得られるものは、一般には不充分、あ

いまいなもので、データベースに熟知した者による再分析、再構成が必須である。

(2) データ収集 国内外の各文献ごとの情報の提供者を定め、それぞれの文献の別刷（またはコピー）を送ってもらうと同時に、文献中の必要なデータをサマリーしたもの（データシートとよぶ）を送ってもらう。データシートの作成は原論文から主観を交えずに情報抽出する必要があり、化学の専門的知識をもつ多くの協力者が必要である。データシート作成者をコンパイラとよんでいる。

(3) 入力変換 集められたデータシートをいかり機械可読ファイル（一定の書式をもったMTファイルなど）に変換する。データシートは化学の専門家ではあるがコンピュータの専門家では必ずしもるい人によって作成されるため、そのまま機械可読な形に変換できるとは限らない。そこで、対象分野（化学）とコンピュータの双方に、ある程度の知識を有する人による作業を伴う。このようにしてできた機械可読ファイルはデータファイル、データファイル上の書式も内部形式とよんでいる。

(4) Check and Authorize データファイルに作成され

に全データについて、化学の立場から見て明らかな誤りや、入力時におけるミス等を修正、除去し、データファイルとして完成させたものにする。

- (5) データブック作成 データファイルを編集し、重要な項目についてはインデックスを作成して、プリントアウトを本の形にする。これをデータブックとよぶ。データブックは図書館等で総目録として利用される他、データの提供者(外国)に配布することも考えられる。(磁気テープ媒体そのものによる提供も考えられるが、外国の中からは、必ずしもコンピュータが手近で利用できるとは限らない所もあり、書籍の形の提供も重要である)

- (6) DBMSによる運用 データファイルをもとにデータベースとDBMS(データベース管理システム)により運用する。現在商用化されているDBMSは、もともと事務計算を主体として考えられており、数値データや化学式などもとり扱うようにはできている。また検索論理も十分な能力をもちていることが多い。したがって、いくらとも学術データベースのあり方を展望するところから観点は、既存のDBMSで満足できるものはなく、システムの設計、試作も必要である。

データベース作成の現状と問題点

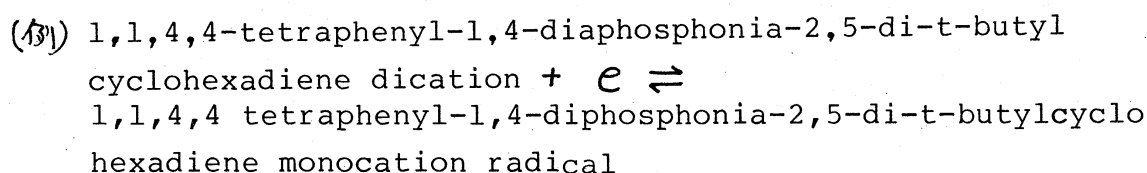
以上の考察のもとに、ここでは前項の手順の具体的な実施方法と、問題点について述べる。

(1) 入力変換に関する問題

図1にデータ・シートの例を示す。ここでも明らかなように、データ形式は、項目は分けているものの、かなり自由に記入している。本プロジェクトにおいても、データの収集段階において、各項目の記述法や欄指定を厳密に規定するべきだとの意見が初期には見られた。しかし実際にはコンパイラに必要以上の負担をかけるため、^{現在では}むしろ原論文の意図を反映すべく、できるだけ自由形式にデータシートを作成してもらい、これをもう一度コーディングシートとよばれるシートに書き写している。コーディングシートも原論文の意図を歪めることのりよう注意が払われているが、一方コンピュータ入力かきやすいように工夫されている。たとえば原論文で複数の測定（条件を変えたもの）が行なわれている場合、データシートでは1枚になるが、コーディングシートでは測定ごとに独立したシートになる。また文字種やデリミタ等もコンピュータ側で認識しやすいような形になっている。コーディングシートの例を図2に示す。これらの作成は現在は化学系学科を卒業した女性数名とお願いしている。

(2) データファイル形式の問題

電気化学データベースには有機物を含む物質名を記憶する必要があり、非常に長いものもあれば、短かくて済むものもあり、その差が激しい。長い反応式の例を下に示す。



したがって可変長のフィールドをもつことの本質的に必要である。また電極の表面処理法など、特殊な測定のみ記入され、コンシト的な役割をもつものもある。以上を勘案してデータファイルの形式(内部形式)としては、(イ)フィールド長を10バイト単位に任意に設定できること、(ロ)記述の項目は、フィールド単位にも、大項目(論理レコードタイプ)単位にも省略できること、(ハ)相対形式の記述(直前に定義されたデータに対して変更点だけを再定義するもの*)と許すこととを条件として次のように設計した(表1参照)。

*注.

本データベースの特徴として、他の条件と比べて同じにして1つか2つのパラメータだけを動かした測定(たとえば温度変化など)が多く見られる。このような場合の冗長性を少なくするため、変更したところのみデータとして持つのである。

- ・ 1レコード 80バイト, 1フィールド 10バイト,
- ・ 大項目 (「反応式」, 「電極」, 「速度定数」, など) ごとに論理レコードを形成する。論理レコードはレコードをまたいで定義される。論理レコードの一部は省略可。
- ・ 1論理レコード中に, 各項目をフィールド値として記述する。フィールドは識別子を伴って与えられ, 省略可能である。また "—" 識別子による継続も可能。

レコード例を下に示す。これは図2のEAS (活性物質名とその濃度) 項目に対応している。(識別子については図3を参照)

ES	:MTCr(CN)6(3- ---)	:V12.0E-3	:**	:**	: 00000000
	----- ----- ----- ----- -----				
	1 10	1 10	1 10	1 10	1 10

(3) 表現の多様性

上例にもあるように, 測定値などのデータでは有効数字や書き方, 精度によって, いろいろの表現方法があり得る。

例. 1.2 ± 0.05 , $1.15 \sim 1.24$, $\sim 1.23 \approx 1.2$

データ作成者は測定法, 機器精度などから, とも適切な表現方法を選んでいると考えられるので, どれかに統一することは好ましくない。本データベースは上の4通りを認めることとしてある。また濃度値などでは 0.5 mol/l のような表現の他に「Saturated」(溶けるだけ溶かした) のような例もあり, 将来いっしょに INVERTED INDEX をつける場合に困難を伴うものと思われる。

(4) データ構造とデータモデルの問題

電気化学データと DBMS によって運用しようとする場合、通常のデータベース (formatted) に比べて問題点が多い。またデータ構造として、項目が多くかつ分割 (階層化) の度合いが大きい。たとえば

- ・ 1 文献は複数測定を含む
- ・ 1 測定とは同一物質 (化学種) に対して、いろいろなパラメータ (支持電解質, 溶液, 温度など) を変化させたものの 1 つである。
- ・ 測定に用いられた測定法についても、その精度や、変更点, 対象パラメータ範囲などの属性がある。

容易に分かるように、これを関係データベースのように、フラットな複数ファイルへの分散で構成しようとした場合、ファイル数が増え、しかもファイル間をまたぐ関数従属性が頻発する。

例. [測定 文献], [測定 支持電解質], [測定 溶媒], ...
[測定 測定法 変更点] ...

また検索する場合を考えると、そのキーとなる検索項目は非常に多様である。したがって、Join にあたる演算 (集合の結合) とムンバムンバに行なうことになり、データベースとしての検索効率が低くなるおそれがある。またユーザに Join 操作

を実行させるのは、誤りをおかしやすい。

さらに、このようなデータの場合、多様な属性 (attribute) が存在するが、各測定についてみると、実際に値が与えられる属性値はごくわずかである。たとえばある測定に限ったときのみ得られる測定値や、またその補正が必要な場合などである。以上から考えると、通常のデータベースのように、「ぎっしりつまったレコードの集まり」としてのファイル構造では間を合わず、新しいモデルにもとづくデータベースを考へなければならぬと思われぬ。我々は、この問題については、新しいモデル AIS によるアプローチを考えている。⁽³⁾

現状と今後の課題

昭和57年2月までに、7000データシート程度のMT化が済む見込み、今後データの収集は順調に伸びつづけゆくものと思われる。現在、これらのデータもとりあえずデータブックという形で公開することを進めており、これは、内部形式に近い形式も、できるだけユーザに分かりやすい形でフォーマット化したものであり、いくつかの項目についてはイレダックスも設ける。その形式例を図3に示す。白スキの部分に値が入れられる。またこの枠は、実際には、データ長に合わせて上下方向に伸び縮みする。この形式のデータは、ライブラリとして出

カを製本したものを公開すると同時に、最新の資料を、N-1ネットワーク(大学間学術情報ネットワーク)も含めて、オンライン的にアクセスできるようにする予定である。

前に述べたように、この種のデータベースの作成は我国では先例が少なく、また情報の内容が高度に専門的で、利用者のレベルも高いだけに、多くの未解決の問題がある。最終的には、DBMSまでを含めて、まったく新しい思想でシステムを設計しなければならぬと思われる。今後、学術情報システムに広く利用できるDBMSの試作を進めるとともに、電気化学データベースの高度な運用を試行してゆきたいと思っている。

参考文献

- (1) 玉虫・伊豆津他：電極反応の基礎，共立出版
- (2) 有澤：大学図書館におけるデータベース・システムの役割，大学図書館研究(掲載予定)，(1982)
- (3) 有澤：Entity-Associationモデルによるデータベース設計
情報処理学会「アドバンス・データベース・システム」シンポジウム，(1981)

謝辞 データベース設計、データ作成にご協力いただいている三國房子氏
他データベース作成室の方々、プリントアウト・ツールと設計・製作していただいている
森雅一氏他の方々に謝意を表します。

表1. 内部データ形式

項番	先頭からの バイト数	大きさ	タイプ	内容	
1.	0	2	A	論理レコードタイプ名	
2.	2	3	A	補助記述子 相対/絶対形式 種別	
3.	5	1		"：" (デリミタ)	
4.	6	2	A	フィールド記述子	
5.	8	10	A	フィールド値(データ)	
6.	18	1		"：" (デリミタ)	
7. ? 9.	19			4~6 のくり返し	
10. ? 12.	32			4~6 のくり返し	} 4~8で合計5桁の フィールド値を記述 できる。
13. ? 15.	45			4~6 のくり返し	
16. ? 18.	58			4~6 のくり返し	
19.	71	1	~	(空白)	
20.	72	8		シーケンスフィールド	

注意 ・ フィールド記述子 "****" は Filler 用として予約されている。

・ 論理レコードタイプ名およびフィールド記述子における"ー"は継続指定。

例 1. データシート例

Type of Electrode Reaction		$\text{Cr}(\text{CN})_6^{3-} + e \rightleftharpoons \text{Cr}(\text{CN})_6^{4-}$	
Electrode System	Working electrode	Hg (DME)	
	Electroactive species	$\text{Cr}(\text{CN})_6^{3-}; 2.0 \times 10^{-3} \text{ M}$	
Medium (supporting electrolyte, solvent, surface-active substance, etc.)		1.0 M KCN	
Temperature		25°C	
Electrode potential	$E^\circ, E^\circ, E^{1/2}$ etc.		
	Range of the measurement		
Other conditions			
Transfer Coefficient	anod.	meas.	
		cor.	
	cathod.	meas.	$\alpha = 0.59$
		cor.	
Rate Constant	meas.	$k_{10} = 0.24 \text{ cm sec}^{-1}$	
	cor.		
Activation energy (including Gibbs energy, enthalpy, and entropy of activation)			
Diffusion Coefficient			
Experimental Conditions	Method	AC polarographic measurements	
	Time-domain of measurement	d.c scan rate 25 mV/min.	
	WE (geometry, etc.)	Hg (DME)	
	Counter electrode		
	Reference electrode		
	Hydrodynamic condition		
Remarks			
Reference	Authors	Eric R. Brown, Haying L. Hung, Thomas G. McLeod,	
	Journal	Donald E. Smith and Glenn L. Boorman Analytical Chemistry 40, (1968), 1424	

COMPILER: F. Mikuni

例 2. コーディングシート例 ECDATA Coding Sheet

03302

YNU Electrochemical Data Center

SNO			177-φφ-φφ		
OXST			Cr(CN) ₆ (3-)		
ORXN			Cr(CN) ₆ (3-) + e ⇌ Cr(CN) ₆ (4-)		
NOSN			1		
WE			; Hg; DME		
EAS			Z, φE-3; Cr(CN) ₆ (3-)		
MEDS	A		H ₂ O		
MEDE			1φ; KCN		
MEDA					
PHIS					
TEMP			25		
ELPO					
OCON			dc scan rate = 25 mV/min.		
ATCM					
ATCF					
CTCM	A		φ, 59		
CTCF					
ATSL					
CTSL					
RCME	φA		φ, 24		
RCFR					
ACEN					
DOX					
DRED					
METD			ACP } fundamental harmonic analysis		
RANG					
STRT					
REFE			SCE		
HYCD					
RXMN					
AUTR			Brown, E.R.; Hung, H.L.; McCord, T.G.; Smith, D.E.; Braman, G.L.		
JOUR			Anal. Chem., 1968, 40, 1424 (Eng)		

