数理解析研究所講究録
第461巻 1982年 353-367

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をうながしている。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，データのじん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，新たな課題にぶつかることも多かった。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（省エネルギーを含めて）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，とも主要なデータである。反応速度パラメータとは電極反応

$$O + ne \rightleftharpoons R$$

（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，電子移動に等々する物質），溶媒，添加物，温度等，様々の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）が主要なものである。(1)

電気化学データベースのオリジナル・ソースは，世界各国で発刊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，前記の本質的な諸条件の他にも，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ー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]にゆずり，電気化学データの収集とデータベース化の具体的手順に焦点をしぼって述べる。現在までに行なってきた，あるいは今後ただち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に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わゆる機械可読ファイル（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な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たちによる作業を伴う。このようにしてできた機械可読ファイルをデータファイル，データファイル上の書式を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

た全データについて，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れたものにする。

(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，外国の中には，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を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をとり扱うようにはできていない。また検索論理も十分な能力をもっていないことが多い。したがって，少なくとも学術データベースのあり方を展望するという観点からは，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1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ないよう注意が払われているが，一方コンピュータ入力がし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各形になっている。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にお願いしている。

(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，短かくてすむものもあり，その差が激しい。長い反応式の例を下に示す。

(例) 1,1,4,4-tetraphenyl-1,4-diaphosphonia-2,5-di-t-butyl cyclohexadiene dication + e ⇌ 1,1,4,4 tetraphenyl-1,4-diphosphonia-2,5-di-t-butylcyclo hexadiene monocation radical

したがって可変長のフィールドをもつことが本質的に必要である。また電極の表面処理法など，特殊な測定たのみ記入され，コメン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) 記述のない項目は，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*）を許すこ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，他の条件をすべて同じにして1つか2つのパラメ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．このような場合大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もつのである．

・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1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いるため，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   :MICr(CN)6(3-:--)    :V12.0E-3    :**          :**          : 00000000
     1---10---1---10---1---10---1---10---1---10---1---10---1
```

(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，いろいろな表現方法があり得る。

例. 1.2 ±0.05 , 1.15～1.24 , ～1.23 ≒1.2

データ作成者は測定法，機器精度などからもっ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通りを認めることにしてある。また濃度値などでは「0.5mol/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，将来いわゆる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

電気化学データをDBMSによって運用しようとする場合，通常なデータベース（formatted）に比べて問題点が多い。まずデータ構造として，項目が多くかつ分割（階層化）の度合が大きい。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，ファイル数がふえ，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）をひんぱん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な，「き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に合わず，新しいモデルにもとづくデータベースを考えなければならないと思われる。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 化がすみ，今後データの収集は順調に伸びつづけてゆくものと思われる。現在，これらのデータを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いる。これは，内部形式に近い形を，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ンデックスも設ける。その形式例を図3に示す。白ヌ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，ラインプリンタ出

力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）を介して，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なくなると思われる。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「アドバンスト・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国房子氏他データベース作成室の方々，プリントアウト・ルーチンを設計・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	4〜18で合計5個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"--"は継続指定。

図1. データシート 例

Type of Electrode Reaction			$Cr(CN)_6^{3-} + e \rightleftharpoons Cr(CN)_6^{4-}$
Electrode System	Working electrode		Hg (DME)
	Electroactive species		$Cr(CN)_6^{3-}$; 2.0×10^{-3} M
Medium (supporting electrolyte, solvent, surface-active substance, etc.)			1.0M KCN
Temperature			25°C
Electrode potential	E°, $E^{\circ\prime}$, $E^{\prime}_{1/2}$ etc.		
	Range of the measurement		
Other conditions			
Transfer Coefficient	anod.	meas.	
		cor.	
	cathod.	meas.	$\alpha = 0.59$
		cor.	
Rate Constant		meas.	$k_s = 0.2$ 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 measurem.		dc scan rate 25 mV/min.
	WE (geometry, etc.)		Hg (DME)
	Counter electrode		
	Reference electrode		
	Hydrodynamic condition		
Remarks			
Reference	Authors		Eric R. Brown, Hoying L. Hung, Thomas G McLord, Donald E. Smith and Glenn L. Booman
	Journal		Analytical Chemistry 40, (1968), 1024

COMPILER: F. Mikuni

図2. コーディングシート例

ECDATA Coding Sheet

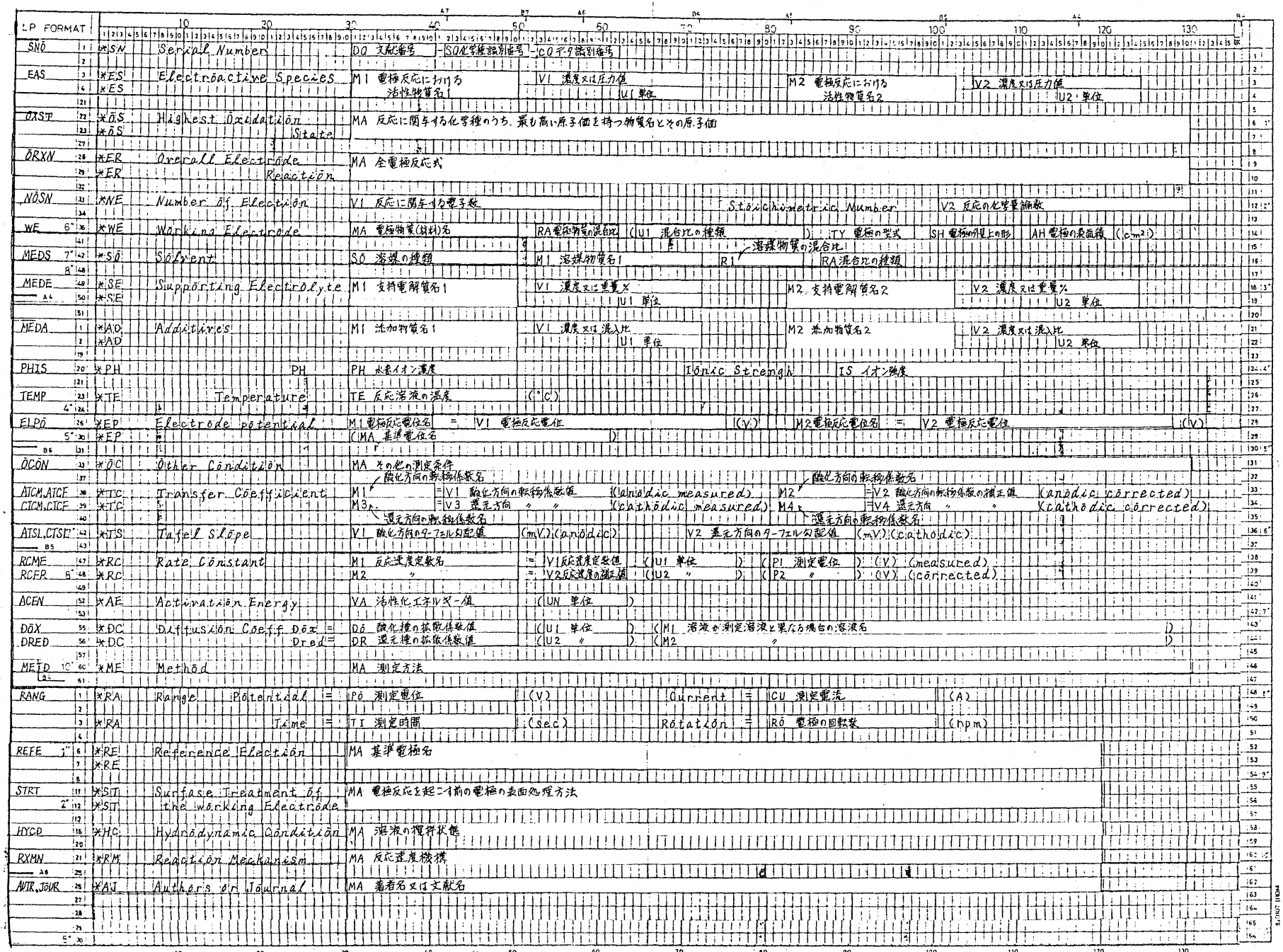
03302

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			2, φE−3 ; Cr(CN)6(3−)	
MEDS		A	H2O	
MEDE			1, φ ; KCN	
MEDA				
PHIS				
TEMP			25	
ELPO				
OCON			dc scan rate = 25 mV/min.	
ATCM				
ATCF				
CTCM		A	φ, 59	
CTCF				
ATSL				
CTSL				
RCME	φ	A	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.;	Bowman, G.L.
JOUR			Anal. Chem., 1968, 40, 1424 (Eng)	

1 2 5 6 7 8 72 73 79 80

図3. 出力形式設計図


LP FORMAT
SNO
*SN Serial Number
DO 文献番号 - SO化学種識別番号 - CO データ識別番号
EAS
*ES *ES Electroactive Species
M1 電極反応における活性物質名1
V1 濃度又は圧力値
U1 単位
M2 電極反応における活性物質名2
V2 濃度又は圧力値
U2 単位
OXST
*OS *OS Highest Oxidation State
MA 反応に関与する化学種のうち、最も高い原子価を持つ物質名とその原子価
ORXN
*ER *ER Overall Electrode Reaction
MA 全電極反応式
NOSN
*NE Number of Election
V1 反応に関与する電子数
Stoichiometric Number
V2 反応の化学量論数
WE
*WE Working Electrode
MA 電極物質(材料)名
RA電極物質の混合比
(U1 混合比の種類)
TY 電極の型式
SH 電極の形上の形
AH 電極の表面積 (cm²)
MEDS
*SO *SO Solvent
SO 溶媒の種類
M1 溶媒物質名1
R1 溶媒物質の混合比
RA混合比の種類
MEDE
*SE *SE Supporting Electrolyte
M1 支持電解質名1
V1 濃度又は重量%
U1 単位
M2 支持電解質名2
V2 濃度又は重量%
U2 単位
MEDA
*AD *AD Additives
M1 添加物質名1
V1 濃度又は混入比
U1 単位
M2 添加物質名2
V2 濃度又は混入比
U2 単位
PHIS
*PH PH
PH 水素イオン濃度
Ionic Strength
IS イオン強度
TEMP
*TE Temperature
TE 反応溶液の温度
(°C)
ELPO
*EP *EP Electrode potential
M1 電極反応電位名 = V1 電極反応電位 (V)
M2 電極反応電位名 = V2 電極反応電位 (V)
(MA 基準電位名)
OCON
*OC Other Condition
MA その他の測定条件
ATCM,ATCF
CTCM,CTCF
*TC *TC Transfer Coefficient
M1 酸化方向の転移係数名 = V1 酸化方向の転移係数値 (anodic measured)
M2 酸化方向の転移係数の補正値 (anodic corrected)
M3 還元方向の転移係数名 = V3 還元方向 〃 〃 (cathodic measured)
M4 還元方向の転移係数名 = V4 還元方向 〃 〃 (cathodic corrected)
ATSL,CTSL
*TS *TS Tafel Slope
V1 酸化方向のターフェル勾配値 (mV)(anodic)
V2 還元方向のターフェル勾配値 (mV)(cathodic)
RCME
RCFR
*RC *RC Rate Constant
M1 反応速度定数名
M2 〃
= V1反応速度定数値 (U1 単位) (P1 測定電位) (V) (measured)
= V2反応速度の補正値 (U2 〃) (P2 〃) (V) (corrected)
ACEN
*AE *AE Activation Energy
VA 活性化エネルギー値
(UN 単位)
DOX
DRED
*DC *DC Diffusion Coeff Dox = Dred =
DO 酸化種の拡散係数値
DR 還元種の拡散係数値
(U1 単位) (M1 溶液が測定溶液と異なる場合の溶液名)
(U2 〃) (M2 〃)
METD
*ME *ME Method
MA 測定方法
RANG
*RA *RA Range Potential = Time =
PO 測定電位 (V)
TI 測定時間 (sec)
Current = IO 測定電流 (A)
Rotation = RO 電極の回転数 (rpm)
REFE
*RE *RE Reference Election
MA 基準電極名
STRT
*ST *ST Surface Treatment of the working Electrode
MA 電極反応を起こす前の電極の表面処理方法
HYCD
*HC *HC Hydrodynamic Condition
MA 溶液の撹拌状態
RXMN
*RM *RM Reaction Mechanism
MA 反応速度機構
AUTR,JOUR
*AU *AU Author's or Journal
MA 著者名又は文献名