社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

英文数学文書の正解付き文字・記号画像データベース

野村　明弘†　　内田　誠一††　　鈴木　昌和†††

† 九州大学大学院数理学府
†† 九州大学大学院システム情報科学研究院
††† 九州大学大学院数理学研究院
〒 812–8581 福岡市東区箱崎 6–10–1
E-mail: †nomura@math.kyushu-u.ac.jp, ††uchida@is.kyushu-u.ac.jp, †††suzuki@math.kyushu-u.ac.jp

あらまし　英文数学文書の正解付き文字・記号画像データベース (InftyCDB-1) について，その仕様と，文字および単語単位の解析結果について述べる．本データベースでは，文書中のすべての文字ならびに記号それぞれについて，文字種やフォント，異常文字正常文字の区別等の正解情報が付与されている．さらに数式中の文字・記号については，その数式を木構造表現するために必要十分な情報も付与されている．また、単語データベースや数式画像データベースとしての利用も容易となるように工夫している．本データベースは一般公開される予定である．
キーワード　数学文書, OCR, 文字画像データベース, 単語データベース、数式データベース

A Ground-Truthed Mathematical Character and Symbol Image Database

Akihiro NOMURA†, Seiichi UCHIDA††, and Masakazu SUZUKI†††

† Graduate School of Mathematics, Kyushu University
†† Faculty of Information Science and Electrical Engineering, Kyushu University
††† Faculty of Mathematics, Kyushu University
Hakozaki 6–10–1, Higashi-ku, Fukuoka-shi, 812–8581 Japan
E-mail: †nomura@math.kyushu-u.ac.jp, ††uchida@is.kyushu-u.ac.jp, †††suzuki@math.kyushu-u.ac.jp

Abstract　This paper is a specification of our ground-truthed mathematical character and symbol image database, called InftyCDB-1. The ground-truth of each character is composed of type, font, quality (touched/broken) and link (relative position), etc. The database includes all the characters and symbols of 467pages of 30articles on mathematics, and is organized so that it can be used as word image database or as mathematical formula image database. InftyCDB-1 is a public database and freely usable for research and development purposes.
Key words　mathematical documents, OCR, character image database, word database, formulae database

1. ま え が き

本論文では，英文数学文書の正解付き文字・記号画像データベースについて，その仕様と，文字および単語解析の結果について報告する．本データベースは主として以下のような研究に寄与するものと考えられる．

- 科学技術文書用文字・記号認識手法の研究開発／評価
- 数式画像の構造解析手法の研究開発／評価
- 数学文書中の単語解析

本データベースでは，文字の種類だけでなく，フォント (例えばイタリック，ボールド) も区別して正解がつけられている．従って，認識手法開発時に，イタリック文字の影響を実験的に評価する場合にも利用できる．また全ての特殊記号についても正解を付与しており，それらの頻度，認識性能等を把握するためにも利用可能である．さらに画像を利用することで標準パターンの生成も可能と考えられる．

本データベースは各文字を単位として正解が付与されている．ただし，同じ単語・数式には同じ ID を各文字・記号に付与しており，さらに単語・数式内での位置座標も収録している．また，付随する画像は単語，数式単位で収録している．このように，単文字・記号レベルだけでなく，単語および数式データベースとしての利用も容易となるように工夫をしている．従って，バイグラムやトライグラムなどへの利用，単語解析や数式構造解析にも活用できると考えられる．

著作権上の制限により，ページ画像を完全に復元できるような情報を持たせることはできないことから，本データベースは

各単語・数式画像のページ内での絶対座標は収録していない．従って，例えばページレイアウト解析には使うことはできない．

本論文は以下のように構成される．第 2 章では本データベースに使用した正解データベースの概略，およびに文字・単語・数式解析ついて述べる．第 3 章では，本データベースの仕様について述べる．

以下で使用する用語について説明する．まず,「文字」は”A” などの通常文字に加え，数学記号も指す．また,「カテゴリ」は文字種別の最小単位を指し、「タイプ」はフォントなど共通の性質を持つカテゴリの集合のことを指す．例えば，“A”，“B”，“C” は同じ Roman というタイプに属する 3 つのカテゴリである．カテゴリとタイプの詳細については後述する．各文字は「テキスト領域」か「数式領域」のいずれかに属する．数式領域には，式番号がついたような数式だけでなく，行中の数式 (インライン数式) も含む．変数も数式領域とする．例えば，“Substitution x^2+y^2 for A” という表現において，“x^2+y^2” と “A” は数式領域であり，それ以外はテキスト領域である．また，この例にある “x” や “$+$” を「ベースライン文字」，“2” を「添字」と呼ぶ．分数 “$\frac{a^2}{b}$” においては，“a” と “b” をベースライン文字とし，“2” を添字とする．

2. データの収集

2.1 対象データ

本論文で用いた文字画像データベースに収録した文書は，純粋数学に関する 30 編の英語論文であり，発行年度は 1970 年代から 2000 年代である．文書は基本的にランダムに選出しているが，ほとんど数式を含まない文書は避けた．総ページ数は 467 ページ，総文字数は 688,570 文字であった．総単語数は 108,914 であり，総数式数は 21,056 であった．表 1 に本データベースにおけるタイプ毎のカテゴリ数および文字数を示す．詳細は文献 [1] を参考にされたい．ただし，本データベースでは文献 [1] の投稿後発見されたごく少数の誤りの修正を施しており，また今回新たにボールドの属性を追加したため，数値にわずかな違いはある．

このように本データベースは，数式 OCR の関する従来の検討 [2], [3] において使用されるものに比べ，相当大規模であるといえる．なお，行列・表・図形の領域については除外している．

全てのページ画像は商用スキャナ (RICOH imagio Neo 450) を使用し，自動で 2 値化したものである。その際，紙の劣化や印刷の悪い文書では接触文字や分離文字などの異常文字 [4] が散見された．

2.2 正解付け作業

全文字について，数学を専攻する学生 7 名を中心として，フォントなどの文字属性 (ground-truth) を手動で付与した．付与した主な情報は以下の通りである。

- 文字カテゴリ (“A”，“α”，“$\int$” 等) およびタイプ
- 領域 (数式領域/テキスト領域別)
- フォント (Italic・Bold)
- 正常文字と異常文字の別

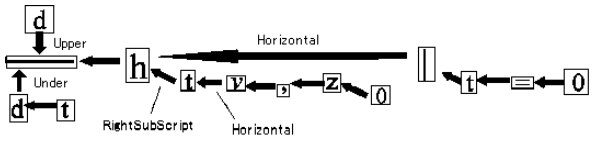

図 1　数式の構造を表現するリンク.

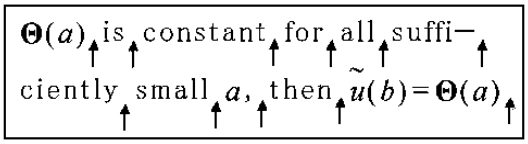

図 2　単語・数式セグメンテーションの例.

- サイズ (幅と高さ)
- 単語もしくは数式中の文字位置
- 隣接文字との位置関係を表すリンク

正解付け作業の際にあらかじめ定義たカテゴリの数は，1,561 個と非常に多くなっている．なお，文献 [1] においてはボールドと非ボールドを区別していなかったが，今回再び全文字を見直すことにより，それらを可能な限り区別した．

リンクとは，数式構造を表現するために付与された属性である．図 1 に例示したように，リンクは隣接する文字の位置関係を表すものである．このリンクにより，1 つの数式全体の構造は木構造として表現される [6], [7]．リンクの種類は水平 (`Horizontal`)，右上付き (`RightSupScript`)，右下付き (`RightSubScript`)，左上付き (`LeftSupScript`)，左下付き (`LeftSubScript`)，上付き (`Upper`)，下付き (`Under`) の 6 種である．水平以外のリンクを含むような数式は，２次元構造を持つといえる (付録 1. も参照のこと).

2.3 単語単位へのセグメンテーション

文書中の単語・数式には固有の番号 (以下、単語 ID と呼ぶ．) がふられている．この単語・数式のセグメンテーションには基本的にスペースを用いた．数式に関しては 1 数式を 1 単語とみなす．ただし，単語および数式が 2 行にまたがる場合は，行の終わりで分けている．また，数式領域とテキスト領域の境界でも分けている．点類 (“,” や “.” など) は直前の単語・数式に含めた．括弧類は開き括弧類は後ろの単語・数式に，閉じ括弧は前の単語・数式に含めた．引用符に関しても括弧類と同様とした．例えば，図 2 では矢印の場所で分割され，“$\mathbf{\Theta}(a)$”，“is”，“constant”，“for”，“all”，“suffi-”，“ciently”，“small”，“a,”，“then”，“$\tilde{u}(b)=\mathbf{\Theta}(a)$” の 11 単語となる．

3. 本データベースの仕様

3.1 概　　要

今回提供する数学文書の文字記号データベース，InftyCDB-1 は以上の正解付け結果を基に作成したものである．本データベースは (i) 正解情報からなるテキストデータと (ii) 画像データの 2 つからなる (図 3).

テキストデータは Microsoft Access もしくは CSV 形式である．その具体的な項目は 3.2 節で示す．画像データは PNG 形

表 1　データベース中の文字の内訳.

type	font	category examples	#predefined categories	text region #cat.	text region #char (%)	math region #cat.	math region #char (%)	total #cat.	total #char (%)
accent		ˆ ˜ ¯ ¨ ˋ ˇ _	13	1	2 (<0.01)	7	2,700 (1.72)	7	2,702 (0.39)
arrow		←↔←↖	16	1	3 (<0.01)	7	1,103 (0.70)	7	1,106 (0.16)
big symbol		$\sum\int\prod$	18	0	0 (0.00)	11	2,458 (1.57)	11	2,458 (0.36)
blackboard bold		$\mathbb{ABCDEF}$	26	0	0 (0.00)	9	427 (0.27)	9	427 (0.06)
calligraphic		$\mathcal{ABCDEF}$	26	0	0 (0.00)	19	592 (0.38)	19	592 (0.09)
German	Upright	$\mathfrak{A}\,\mathfrak{B}\,\mathfrak{C}\,\mathfrak{a}\,\mathfrak{b}\,\mathfrak{c}$	52	0	0 (0.00)	25	1,041 (0.66)	25	1,041 (0.15)
	Bold	$\boldsymbol{\mathfrak{A}\mathfrak{B}\mathfrak{C}\mathfrak{a}\mathfrak{b}\mathfrak{c}}$	52	0	0 (0.00)	0	0 (0.00)	0	0 (0.00)
Greek	Upright	ΓΔΘ	11	0	0 (0.00)	10	2,148 (1.37)	10	2,148 (0.31)
	Italic	$\alpha\beta\gamma$	29	5	19 (<0.01)	23	10,618 (6.76)	23	10,637 (1.54)
	Bold	**ΓΔΘ**	11	0	0 (0.00)	1	3 (<0.01)	1	3 (<0.01)
	Italic Bold	$\boldsymbol{\alpha\beta\gamma}$	29	0	0 (0.00)	5	31 (0.02)	5	31 (<0.01)
extended Latin	Upright	Ä Æ è	182	30	392 (0.07)	2	3 (<0.01)	30	395 (0.06)
	Italic	*Ä Æ è*	182	9	55 (0.01)	2	10 (0.01)	10	65 (0.01)
	Bold	**Ä Æ è**	182	4	6 (<0.01)	0	0 (0.00)	4	6 (<0.01)
	Italic Bold	***Ä Æ è***	182	0	0 (0.00)	0	0 (0.00)	0	0 (0.00)
numeric	Upright	0 1 2	10	10	12,018 (2.26)	10	15,294 (9.74)	10	27,312 (3.97)
	Italic	*0 1 2*	10	10	140 (0.03)	4	118 (0.08)	10	258 (0.04)
	Bold	**0 1 2**	10	10	923 (0.17)	4	26 (0.02)	10	949 (0.14)
	Italic Bold	***0 1 2***	10	0	0 (0)	0	0 (0.00)	0	0 (0.00)
operator		+ − ×/ < &	92	6	154 (0.03)	49	20,359 (12.96)	50	20,513 (2.98)
others	Upright	§ ; : ∞ ∀ ∃ †	44	12	3,532 (0.66)	17	2,567 (1.63)	22	6,099 (0.89)
	Bold	**§ ; :**	8	4	63 (0.02)	0	0 (0.00)	4	63 (0.01)
parenthesis	Upright	( ) { } [ ]	20	7	8,082 (1.52)	12	30,334 (19.31)	12	38,416 (5.58)
	Bold	**( ) { } [ ]**	20	2	112 (0.02)	0	0 (0.00)	2	112 (0.02)
point	Upright	, . ‘ ’	15	9	20,970 (3.95)	9	7,673 (4.89)	12	28,643 (4.16)
	Bold	**, . ‘ ’**	15	5	448 (0.08)	0	0 (0.00)	5	448 (0.06)
Roman	Upright	A B C a b c	61	57	414,825 (78.05)	55	8,259 (5.26)	57	423,084 (61.44)
	Italic	*A B C a b c*	61	55	63,590 (11.96)	53	49,072 (31.24)	56	112,662 (16.36)
	Bold	**A B C a b c**	61	56	6,178 (1.16)	13	538 (0.34)	56	6,716 (0.98)
	Italic Bold	***A B C a b c***	61	0	0 (0.00)	19	1,508 (0.96)	19	1,508 (0.22)
script		$\mathcal{A}\,\mathcal{B}\,\mathcal{C}$	52	0	0 (0.00)	7	176 (0.11)	7	176 (0.03)
total			1,561	294	531,512 (100.00)	373	157,058 (100.00)	487	688,570 (100.00)

注: (1) Roman には，それぞれ “fi” のような合字が 9 カテゴリ含まれている．

式であり，ファイル名の具体的な命名規則およびファイルの保存場所は 3.3 節で示す．以上のテキストデータと画像データは相互に関連付けられている．具体的には，テキストデータの各文字毎にその文字が該当する画像データ名，パス，および画像に対する文字矩形情報が収録されている．

3.2　テキストデータ

各文字についてのテキストデータは表 2 の全 29 項目を準備した．

3.2.1　各文字毎の情報 (項目 (1)〜(15))

項目 (1)〜(3) は各文字を区別するための固有の番号が収録されている．項目 (4) のタイプは表 1 の type(15 種) のいずれかである。項目 (5) のカテゴリ (OCR Code) は鈴木研究室が開発している InftyReader/InftyEditor [5] で定義しているコードでありカテゴリの区別に利用する．項目 (6) は文字の読み方である．例えば，“$\int$” は “int”，“Ω” は “Omega” と表記される．項目 (7) はテキスト領域であれば “text”，数式領域であれば “math” が収録されている．項目 (8) はベースライン上の文字であれば “True”，添字であれば “False” が収録されている．項目 (9) では斜体であれば “True”，そうでなければ “False” が，項目 (10) ではボールド文字であれば “True”，そうでなければ “False” が収録されている．項目 (11) では，正常文字であれば “normal”，接触文字であれば “touched”，分離文字であれば “separate”，接触かつ分離文字は “touch_ and_ sep” が収録されている．項目 (12)，(13) は各文字画像の幅および高さである．項目 (14)，(15) で，リンク先の文字と，その文字との相対位置関係がわかる．この (12)〜(15) を用いることで，単語及び数式の木構造が復元できる．

3.2.2　画像ファイルとの関連情報 (項目 (16)〜(20))

3.3.1 節及び 3.3.2 節で命名する画像ファイル名とその画像の保存ディレクトリをデータベースのテキストデータの項目 (16) に収録している．また，項目 (17)〜(20) に各文字の作成したファイル内での文字位置情報を収録している．従って，これらの (17)〜(20) は単語・数式内部のレイアウト情報を担っている．

3.2.3　単語毎の情報 (項目 (21)〜(29))

項目 (21) は各文字が含まれる単語・数式の固有の番号である．項目 (22) は同じく各文字が含まれる単語を MathML 形式で表現したときの単語文字列である．項目 (23) は TeX 形式で表現したときの単語文字列である．項目 (24) は鈴木研究室が開発している InftyEditor で定義している単語文字列表現である．項目 (25)〜(28) では，画像ファイルでの単語位置情報である．項目 (29) は改行により音節で分けられた後の文字であるか否かの

表 2　データベース中のデータ項目.

	項目	説明
(1)	CharID	各文字ごとにふられている番号
(2)	JornalID	各論文にふられている固有の番号
(3)	SheetID	ページ番号
(4)	Type	タイプの種別
(5)	Code	カテゴリ (OCR Code)
(6)	Entity	カテゴリ (読み方)
(7)	Region	テキスト領域 (text)・数式領域 (math) の別
(8)	Baseline	ベースライン (true) と添字 (false) の別
(9)	ItalicFlag	Upright(false) と Italic(true) の別
(10)	BoldFlag	Upright(false) と Bold(true) の別
(11)	Quality	正常文字 (normal) と異常文字 (touched/separate/touch_ and_ sep) の別
(12)	Width	幅
(13)	Height	高さ
(14)	PareCharID	リンク先の親の CharID
(15)	Link	隣接文字との位置関係を表すリンク
(16)	ImageName	その文字が含まれている画像ファイルの名前とディレクトリ
(17)(18)(19)(20)	Rect	文字が含まれている画像ファイルでの矩形 (Left,Top,Right,Bottom)
(21)	WordID	文書中の単語・数式の固有の番号
(22)	WordMathML	MathML 形式での単語文字列
(23)	WordTeX	TeX 形式での単語文字列
(24)	WordIML	IML 形式での単語文字列
(25)(26)(27)(28)	WordRect	画像ファイルでの単語の矩形 (Left,Top,Right,Bottom)
(29)	SyllableAfter	改行によって単語が分割されたかのフラグ

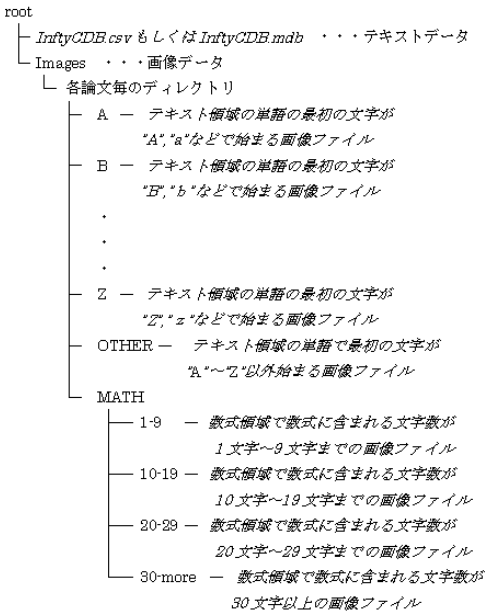

図 3　本データベースのディレクトリ/ファイル構成.
立体：ディレクトリ, 斜体：ファイル

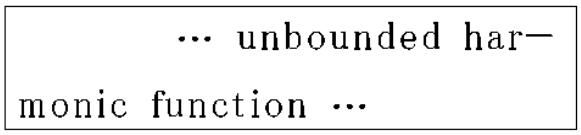

図 4　改行による単語分割の例.

フラグである．図 4 の “harmonic” ではハイフンの後の改行の “monic” に SyllabelAfter にフラグがたっている．同じ単語内に含まれる文字について，以上の項目 (21)～(29) は同一である．従って，データベースとしては冗長な構成ではあるが，使用する利便を考慮し各項目に収録した．

テキストデータの例として数式 “$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{h}_{t\nu,z_0}|_{t=0}$” と単語 “and” は表 3 のようになる．

3.3　画像データ

画像ファイル数の爆発を避けるため，同一論文内の同じ単語は，なるべく 1 つの画像ファイルにまとめた．ただし，イタリック，太文字，大文字，小文字は別カテゴリとしたので，例えば “And” と “and” は別ファイルとなる．数式は多くの場合 1 つの数式で 1 つの画像ファイルとなるが，同一の数式が出現する場合は単語と同様に同じファイルにまとめている．

3.3.1　画像ファイルの命名規則

ファイル名は以下のような定義でつけている．

- 単語 · · · “文字列”(_ FontFlag)_ “番号”.png
- 数式 · · · “文字列長” _ “文字列のはじめの 3 文字” (_ FontFlag)_ “番号”.png

文字列 (および，はじめの 3 文字) には，その単語・数式中の文字列をそのまま表記する．ただし，ファイル名として使えない文字 (“ : ” や “<” など) や表記できない文字 (“∞” や “α”) が存在するため，アルファベットと数字以外は “-読み方-” としている．例えば，単語 “(and)” と数式 “$\alpha \leqq 1$” の画像ファイル名は，それぞれ “LeftPar-and-RightPar_0.png”，“3_alpha-le-1_0.png” である．また，拡張ラテン文字 (“ĉ”) は Roman 文字 (“c”) で表現している．“_ FontFlag” に関して，文字画像に Italic を含んでいれば “_I”，Bold を含んでいれば “_B”，と記述する．Upright の場合は省略する．番号に関して，Windows ではファイル名に大文字と小文字の区別がないため，単語 “And” を集めた画像に “And.png” ファイル名をつけた場合と，単語 “and” を集めた画

表 3　数式 “$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{h}_{t\nu,z_0}|_{t=0}$” と単語 “and” の本データベースにおける表現．各項目は表 2 参照．

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
14	5	10	BigSymbol	33D1	fractionalLine	math	TRUE	FALSE	FALSE	normal	82	12
15	5	10	Roman	0164	d	math	TRUE	FALSE	FALSE	normal	38	60
16	5	10	Roman	0164	d	math	TRUE	FALSE	FALSE	normal	38	59
17	5	10	Roman	0174	t	math	TRUE	FALSE	FALSE	normal	24	53
18	5	10	Roman	0168	h	math	TRUE	FALSE	FALSE	normal	39	60
19	5	10	Roman	0174	t	math	FALSE	FALSE	FALSE	normal	27	52
20	5	10	Greek	426D	nu	math	FALSE	TRUE	FALSE	normal	42	37
21	5	10	Point	142C	comma	math	FALSE	FALSE	FALSE	normal	17	26
22	5	10	Roman	017A	z	math	FALSE	FALSE	FALSE	normal	38	39
23	5	10	Numeric	0130	zero	math	FALSE	FALSE	FALSE	normal	34	49
24	5	10	Parenthesis	197C	vert	math	TRUE	FALSE	FALSE	normal	11	187
25	5	10	Roman	0174	t	math	FALSE	FALSE	FALSE	normal	26	52
26	5	10	Operator	1D3D	equal	math	FALSE	FALSE	FALSE	touched	43	22
27	5	10	Numeric	0130	zero	math	FALSE	FALSE	FALSE	normal	35	48
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
804	2	13	Roman	0161	a	TRUE	text	FALSE	FALSE	normal	37	38
805	2	13	Roman	0164	n	TRUE	text	FALSE	FALSE	separate	42	36
806	2	13	Roman	0164	d	TRUE	text	FALSE	FALSE	normal	42	56

・・・

(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)
-1	TOP	※ 1	0	73	82	85	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
14	UNDER	※ 1	12	101	50	161	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
14	UPPER	※ 1	24	0	62	59	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
15	HORIZONTAL	※ 1	54	109	78	162	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
14	HORIZONTAL	※ 1	95	35	134	95	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
18	RSUB	※ 1	135	78	162	130	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
19	HORIZONTAL	※ 1	173	96	215	133	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
20	HORIZONTAL	※ 1	234	119	251	145	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
21	HORIZONTAL	※ 1	262	92	300	131	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
22	RSUB	※ 1	306	116	340	165	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
18	HORIZONTAL	※ 1	363	18	374	205	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
24	RSUB	※ 1	386	130	412	182	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
25	HORIZONTAL	※ 1	423	154	466	176	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
26	HORIZONTAL	※ 1	487	135	522	183	28005695	※ 2	※ 3	※ 4	0	0	522	205	FALSE
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-1	TOP	※ 5	0	5850	37	5888	1000299	and	and	and	0	5831	129	5888	FALSE
804	HORIZONTAL	※ 5	40	5851	82	5887	1000299	and	and	and	0	5831	129	5888	FALSE
805	HORIZONTAL	※ 5	87	5831	129	5887	1000299	and	and	and	0	5831	129	5888	FALSE

注:　(※ 1) AnnMS_1971_157_173¥MATH¥10-19¥14_-fractionalLine-dd_0.png

(※ 2) $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{h}_{t\nu,z_0}|_{t=0}$ の MathML の表記が挿入 (略).

(※ 3) $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{h}_{t\nu,z_0}|_{t=0}$ の LaTeX の表記が挿入 (略).

(※ 4) $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{h}_{t\nu,z_0}|_{t=0}$ の IML の表記が挿入 (略).

(※ 5) ActaM_1970_37_63¥A¥and_1.png

像に “and.png” ファイル名をつけた場合とでは，ファイル名が同じとなる．それを本データベースでは別の画像ファイルにしたいため，任意の番号を最後につけた．

3.3.2　画像ファイルの保存場所

図 3 に示すように，Images ディレクトリに，各論文毎にディレクトリを作成し，各論文ディレクトリに対して，“A”〜“Z” と “OTHER” と “MATH” ディレクトリを作った．MATH ディレクトリの中には，“1-9”，“10-19”，“20-29”，“30-more” の 4 つのディレクトリがある．各画像ファイルはその単語がテキスト領域の単語の場合，最初の文字のディレクトリ場所にある．数字やアルファベットの場合は OTHER ディレクトリにある．MATH ディレクトリの “1-9” などの各ディレクトリ名は数式に含まれる文字数を表す．例えば，$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{h}_{t\nu,z_0}|_{t=0}$ の画像は文字数が 14 個であるため，MATH ディレクトリの 10-19 のディレクトリにある．上記のようにディレクトリを分類したのは，ファイルアクセス速度向上のためである．図 5 は上が “*approaches*”(ファイル名:“Images¥ActaM_1970_37_63¥A¥approaches_I_0.png”)，下が “$D_{\Omega}[u] < \infty$”(ファイル名:“Images¥ActaM_1970_37_63¥MATH¥1-9¥7_D-Omega- -BigLeftPar-_0.png”) の画像である．

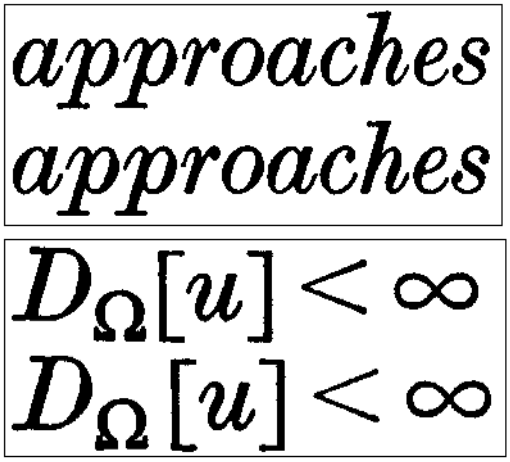

図 5　画像ファイルの例

4. 配布手段ならびに使用条件

本データベースは OCR ソフトウェアの数式を含む科学技術文書への対応の研究・開発に寄与することを目的として作成・公開するものである．研究・開発目的であれば，ユーザー登録手続きをとることにより，誰でも無償で利用できる．データベースの再配布は有償・無償を問わず，認めない．また，本データベースを用いて，製品開発や評価結果発表などを行う場合は，データベース名を記すことが使用条件に含まれる．

配布手段は CDROM/ DVDROM から提供する予定である．テキストデータについては前述のように CSV もしくは Microsoft Access フォーマットで提供する．テキストデータのサイズは CSV が 16.8MB(Zip 圧縮後)，Microsoft Access が 26.7MB(Zip 圧縮後) である．画像ファイルは合計 202MB である．

5. ま と め

本論文では，公開する文字画像データベース (InftyCDB-1) の仕様と，文字および単語・数式の解析について述べた．本データベースは正解情報からなるテキストデータと画像データの 2 つからなる．テキストデータは各文字毎を単位とし，文字属性 (ground-truth) を付与した．さらに数式中の文字・記号については，その数式を木構造表現するために必要十分な情報も付与されている．また，各文字に単語 ID を付与し，各単語毎に画像データを作成し，その単語内での文字情報をテキストデータに収録することで，単語データベースや数式データベースとしての利用も容易になるように工夫した．

文　献

[1] S. Uchida, A. Nomura, M. Suzuki “Quantitative Analysis of Mathematical Documents,” *Int. J. Doc. Anal. Recog.*, to appear.

[2] H.-J. Lee and J.-S. Wang, “Design of a mathematical expression understanding system,” *Pattern Recognition Letters*, 18(3):289–298, 1997.

[3] M. Okamoto, H. Imai, and K. Takagi, “Performance evaluation of a robust method for mathematical expression recognition,” *Proc. ICDAR*, 121-128, 2001.

[4] A. Nomura, K. Michishita, S. Uchida, and M. Suzuki, “Detection and segmentation of touching characters in mathematical expressions,” *Proc. ICDAR*, 1:126-130, 2003.

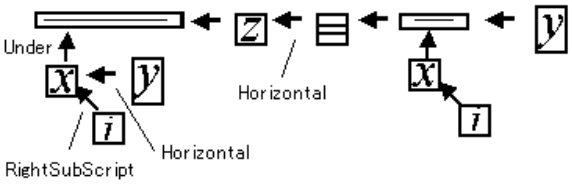

図 A· 1　Accent を含む数式の構造を表現するリンク.

[5] M. Suzuki, F. Tamari, R. Fukuda, S. Uchida, T. Kanahori “Infty- an integrated OCR system for mathematical documents,” *ACM Symposium on Document Engineering*, 95-104, 2003

[6] J. Ha, R. M. Haralick, and I. T. Phillips, “Understanding mathematical expressions from document images,” *Proc. ICDAR*, 956-959, 1995.

[7] Y. Eto and M. Suzuki, “Mathematical formula recognition using virtual link network,” *Proc. ICDAR*, 762–767, 2001.

付　　録

1. アクセント記号のリンクの取り扱い方

本データベースではリンク付けの際，数式の木構造はリンクを用いて一意的に表現できるように，アクセント記号を親とする (図 A· 1)．アクセント記号を親とすることは直感と少し異なるかもしれない．“$\overline{x_i}y$” において，“$\overline{}$” が “x” の上付き添字にするのが直感かもしれない．しかし，上記のようにすると，“$\overline{}$” がどこまでの文字を含んでいるのかという目印を作らなくてはならない．“$\overline{}$” がどこまで文字を含んでいるかを矩形において判断しようとすると，あいまい性が残る．

本手法ではそのような不具合を回避するため,「直感」とは少し異なるが，アクセント記号を親とした。それにより，全ての数式の木構造をリンクのみで表現できる．

2. データベース中の文献リスト

本論文で対象とした数学文書データベースに含まれる英語論文 30 編のリストを以下に示す．いずれも (純粋) 数学に関するものである．

● Acta Math., 124(1-2), 37-63, 1970. ● *ibid.*, 181(2), 283-305, 1998. ● Ann. Sci. Ecole Norm. Sup., 4d sér, t.3, 273-284, 1970. ● *ibid.*, t.30, 367-384, 1997. ● Ann. Inst. Fourier, 20(1), 493-498, 1970. ● *ibid.*, 49(2), 375-404, 1999. ● Ann. Math., 91, 550-569, 1970. ● Ann. Math. Studies, 66, 157–173, 1971. ● Arkiv für Matematik, 9(1), 141-163 1971. ● *ibid.*, 35(1), 185-199, 1997. ● Bull. Amer. Math. Soc., 77(1), 157-159 1971. ● *ibid.*, 77(1), 160-163 1971. ● *ibid.*, 80(6), 1219-1222, 1974. ● *ibid.*, 35(2), 123-143, 1998. ● Bull. Soc. Math. France, 98, 165-192, 1970. ● *ibid.*, 126, 245-271, 1998. ● Invent. Math., 9, 121-134, 1970. ● *ibid.*, 138, 163-181, 1999. ● J. Math. Soc. Japan, 27(2), 281-288, 1975. ● *ibid.*, 27(2), 289-293, 1975. ● *ibid.*, 27(2), 497-506, 1975. ● J. Math. Kyoto Univ., 11(1), 181-194, 1971. ● *ibid.*, 11(1), 373-375, 1971. ● *ibid.*, 11(2), 377-379, 1971. ● Kyushu J. Math., 53, 17-36, 1999. ● Math. Ann., 225(3), 275-292, 1977. ● *ibid.*, 315, 175-196, 1999. ● Tohoku Math. J., 25, 317-331, 1973. ● *ibid.*, 25, 333-338, 1973. ● *ibid.*, 42, 163-193, 1990.