Research Reports on Information Science and
Electrical Engineering of Kyushu University
Vol.12, No.1, March 2007

Databases of Mathematical Documents

Masakazu SUZUKI*, Christopher MALON* and Seiichi UCHIDA **

(Received December 12, 2006)

Abstract: This paper describes the specifications for three ground-truthed mathematical character and symbol image databases, called InftyCDB-1, InftyCDB-2, and InftyCDB-3. In the former two databases, the ground-truth of each character is composed of type, font, quality (touching/broken) and link (relative position), etc. InftyCDB-1 includes all the characters and symbols of 30 articles on mathematics, and is organized so that it can be used as word image database or as mathematical formula image database. InftyCDB-2, which is a continuation of InftyCDB-1, includes 37 articles including French and German articles and is organized like InftyCDB-1. InftyCDB-3 is a single character database for training and evaluating single-character recognition engines.

Keywords: Database, Mathematical document, Mathematical expressions, OCR

1. Introduction

In this paper, we describe the specifications for three ground-truthed mathematical character and symbol image databases, called InftyCDB-1, InftyCDB-2, and InftyCDB-3. Those databases contain not only ordinary characters (e.g., "A", "B", "3") but also special characters and symbols which only appear in mathematical articles (e.g., "α", "$\oint$", "$\mathbb{R}$"). Thus, they can be used for the development and evaluation of single-character recognition engines for math-OCRs.

Furthermore, in InftyCDB-1 and InftyCDB-2, attributes describing mathematical structural context (e.g., superscript and subscript) are attached to each character. This will be useful for the development and evaluation of mathematical structure analysis modules, which are essential for math-OCR. In those two databases, the image data are stored separated into word or math formula units. Thus, those databases can be used as word image databases or as mathematical formula image databases. The image data are arranged in the alphabetic order of the words, regardless of the order in which the words appear in the papers. No whole pages are included in the database, to avoid copyright problems.

Hereafter, the term *character* means not only ordinary letters (e.g., "A"), but also math symbols (e.g., "+"), unless otherwise noted. Characters are the samples in our databases. The term *category* means the finest level of character classification, and the term *type* is applied to certain sets of categories having some visually discernible property. In each database, we have divided the categories into disjoint types. For example, "A", "B" and "C" are three categories belonging to the same type (Roman). In contrast, "A", "A", "$\mathcal{A}$", "$\mathbb{A}$", and "$\mathfrak{A}$" are five categories belonging to different types (Roman, italic, calligraphic, blackboard bold, and German). The term *style* means the font style: italic or upright, and bold or non-bold. Each character belongs to either the *text region* or the *math region*. The math region includes not only numbered equations but also in-line math expressions. Note that many in-line math expressions are composed of a single character, such as "x" in the phrase "The variable x denotes ...".

2. InftyCDB-1

2.1 Data collection

The documents contained in this database are 30 English articles on pure mathematics, published between 1970 and 2000. The articles are listed in Appendix A1. The database comprises 688,750 characters, representing 466 pages. This database is larger than other databases used in past research on math-OCR (e.g., about 15,000 characters of 1), about 10,000 characters of 2), and 350 math expressions of 3)). Note that matrices, tables, and figures are excluded from the database.

All pages were scanned in 600 dpi and binarized automatically by the same commercial scanner (a RICOH Imagio Neo 450). The quality of the resulting page images varies with the quality of paper, and other factors. Several page images are noisy

* Faculty of Mathematics
** Department of Intelligent Systems

Table 1 Statistics of InftyCDB-1.

type	style	category examples	#def cat.	text region #cat.	text region #char (%)	math region #cat.	math region #char (%)	total #cat.	total #char (%)
accent		ˆ ˜ ¯ ⃗ ˘	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		∑ ∫ ∏	18	0	0 (0.00)	11	2,458 (1.57)	11	2,458 (0.36)
blkb. 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	$\mathfrak{ABCabc}$	52	0	0 (0.00)	25	1,041 (0.66)	25	1,041 (0.15)
	bold		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		11	0	0 (0.00)	1	3 (<0.01)	1	3 (<0.01)
	italic bold		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	§@©∞∂	42	10	2,903 (0.55)	15	1,797 (1.14)	20	4,700 (0.68)
	bold		16	3	42 (0.01)	0	0 (0.00)	3	42 (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	, . ‘ ’	17	11	21,599 (4.06)	11	8,443 (5.41)	14	30,042 (4.36)
	bold	**, . ‘ ’**	17	6	469 (0.09)	0	0 (0.00)	6	469 (0.07)
Roman	upright	ABCabc	61	57	414,825 (78.05)	55	8,259 (5.26)	57	423,084 (61.44)
	italic	*ABCabc*	61	55	63,590 (11.96)	53	49,072 (31.24)	56	112,662 (16.36)
	bold	**ABCabc**	61	56	6,178 (1.16)	13	538 (0.34)	56	6,716 (0.98)
	italic bold	***ABCabc***	61	0	0 (0.00)	19	1,508 (0.96)	19	1,508 (0.22)
script			52	0	0 (0.00)	7	176 (0.11)	7	176 (0.03)
total			1,571	294	531,512 (100.00)	373	157,058 (100.00)	487	688,570 (100.00)

Notes: (1) Each “Roman” and “italic” type includes nine double letters (i.e., ligatures), such as “fi”. (2) “blkb. bold”= “blackboard bold”. (3) “Script” is a calligraphic-like typeface.

and include many abnormal characters, particularly touching characters and broken characters.

2.2 Ground truth

The ground truth for each character was attached *manually* by seven university mathematics students. The ground truth of each character consists of the following attributes:

- type, style, and category
- region (text or math)
- quality (normal or abnormal)
- size (height and width)
- location on page
- link to “parent character,” as defined in 4)
- sub/super-script level.

The numbers of types and categories pre-defined on attaching the ground truth were 15 and 1,571, respectively. All 15 types are listed in **Table 1**, and all the pre-defined categories of types “big symbol”, “extend Latin”, “operator”, and “others” are listed in Appendix A2.

Similarly-shaped categories are sometimes defined in different types. For example, Σ (capital sigma / Greek) is similar to ∑ (sum / big symbol), and ∪ (cup / operator) is similar to ⋃ (bigcup / big symbol). These similarly-shaped categories were distinguished manually according to their context and/or usage. Similarly, the single category Ê (extended Latin) was distinguished from the pair of categories Ê (E /Roman, plus ˆ / accent).

Classification of a character implicitly involves selection of a *style*, which is constant within any category. The difference between styles is often very subtle. For example, a bold character in one article may resemble a non-bold character in another article. Therefore, it was important to investigate style variations within a single document carefully, before assigning styles to its characters.

The third attribute represents whether the character is normal or abnormal. As shown in **Fig. 1**, we defined five kinds of abnormal characters: touching, self-touching, broken, touching and broken, and overlaid characters. Overlaid characters are distinguished from touching characters, because they are caused by typographical errors.

The sixth attribute, the link, represents a character’s positional relation to the preceding character (as defined in 4)). The link relations encode the structure of each mathematical expression as a tree. There are six kinds of links: horizontal,

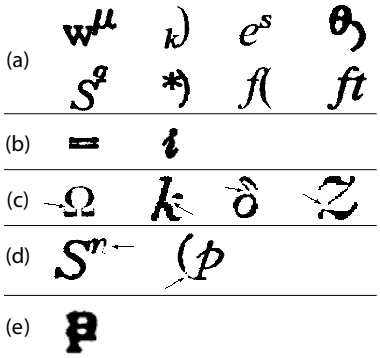

Fig. 1 Abnormal characters. (a) Touching characters, (b) self-touching characters, (c) broken characters, (d) touching and broken characters, and (e) overlaid characters. Broken points are indicated by arrows.

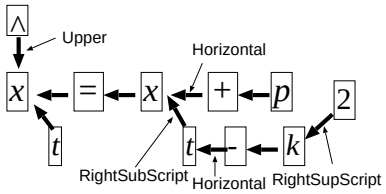

Fig. 2 Links representing the structure of the math expression "$\hat{x}_t = x_{t-k^2} + p$".

right-superscript, right-subscript, left-superscript, left-subscript, upper, and under. A math expression that includes one or more links other than a Horizontal one has a two-dimensional (2D) structure. **Fig. 2** shows a math expression whose structure is represented by 10 links including four non-Horizontal links. Thus, this math expression has a 2D structure.

The seventh attribute, the sub/super-script level, describes the depth of sub/super-scripts. For example, in the math expression in **Fig. 2**, "t", "-", and "k" are first-level subscripts, and "2" is a second-level subscript. Note that baselines are assumed on both the numerator and the denominator of a fraction.

2.3 Statistics

The contents of the database are summarized in **Table 1**. The text and math regions characters appearing in the database comprise 294 and 373 categories, respectively (487 total). Thus, math documents are composed of about 60% more categories than non-math documents. A detailed statistical analysis of InftyCDB-1 is provided in 5),6). Especially in 5), abnormal characters, character size variation, and complexity of math expressions are analyzed carefully.

3. InftyCDB-2

3.1 Data collection and ground truth

InftyCDB-2 is a continuation of InftyCDB-1. It contains 26, 4, and 7 documents in English, French, and German, respectively, none of which appear in InftyCDB-1. See Appendix B for the list of the documents. There are 662,141 characters from English articles, 37,439 from French articles, and 77,808 from German articles.

The same ground-truth attributes listed in Section 2.2 were attached to each character manually, just as in InftyCDB-1. The types defined in InftyCDB-2 are slightly different from those of InftyCDB-2. This difference, however, is less important; in fact, the number of the predefined categories are almost the same. Note that "reject" is a type newly introduced in InftyCDB-2 and includes characters not belonging to any predefined categories. Also note that the type "extended Latin" of InftyCDB-1 is separated into two types "extended Latin" and "Latin" in InftyCDB-2.

3.2 Statistics

Tables 2-4 show the statistics of InftyCDB-2. The number of categories (487) appearing in the 26 English articles (**Table 2**) is the same as that of InftyCDB-1, unexpectedly. This fact implies the category variation in InftyCDB-2 is very similar to that in InftyCDB-1. French and German articles include more (extended) Latin characters in their text parts as shown in **Tables 3** and **4**.

4. InftyCDB-3

4.1 Data collection and ground truth

InftyCDB-3 is a database of *single* alphanumeric characters and mathematical symbols, divided into two data sets, InftyCDB-3-A and InftyCDB-3-B. Unlike InftyCDB-1 and InftyCDB-2, word and mathematical expression structure is not included. The images are of individual characters only. To make it easy to use when experimenting with single-character recognition engines, symbols whose form is identical (for example, the summation symbol ("$\sum$") and the Greek capital sigma ("Σ")) are assigned the same character code. In InftyCDB-3-A, there are 188,752 characters; in InftyCDB-3-B, there are 70,637 characters. We recommend using InftyCDB-3-A as a training dataset and InftyCDB-3-B as a testing dataset.

Table 2 Statistics of InftyCDB-2. (26 English articles)

type	style	#def cat.	text region #cat.	text region #char (%)	math region #cat.	math region #char (%)	total #cat.	total #char (%)
accent		14	0	0 (0)	8	2267 (1.63)	8	2267 (0.34)
arrow		16	2	13 (0)	7	864 (0.62)	7	877 (0.13)
big symbol		18	1	1 (0)	11	2002 (1.44)	11	2003 (0.3)
blackboard bold		52	1	1 (0)	9	655 (0.47)	9	656 (0.1)
calligraphic		26	0	0 (0)	7	75 (0.05)	7	75 (0.01)
German	upright	52	1	1 (0)	14	243 (0.18)	14	244 (0.04)
	bold	52	0	0 (0)	0	0 (0)	0	0 (0)
Greek	upright	11	4	11 (0)	9	1356 (0.98)	10	1367 (0.21)
	bold	11	0	0 (0)	0	0 (0)	0	0 (0)
	italic	29	6	19 (0)	25	9411 (6.79)	25	9430 (1.42)
	italic bold	29	0	0 (0)	1	1 (0)	1	1 (0)
extended Latin	upright	127	15	96 (0.02)	0	0 (0)	15	96 (0.01)
	bold	127	0	0 (0)	0	0 (0)	0	0 (0)
	italic	127	4	7 (0)	0	0 (0)	4	7 (0)
	italic bold	127	0	0 (0)	0	0 (0)	0	0 (0)
Latin	upright	55	19	279 (0.05)	1	2 (0)	20	281 (0.04)
	bold	55	2	4 (0)	0	0 (0)	2	4 (0)
	italic	55	9	58 (0.01)	3	7 (0.01)	12	65 (0.01)
	italic bold	55	0	0 (0)	0	0 (0)	0	0 (0)
ligature	bold	9	1	16 (0)	0	0 (0)	1	16 (0)
	italic	9	4	217 (0.04)	0	0 (0)	4	217 (0.03)
	italic bold	9	0	0 (0)	0	0 (0)	0	0 (0)
	upright	9	5	1066 (0.2)	0	0 (0)	5	1066 (0.16)
numeric	upright	10	10	12484 (2.38)	10	15678 (11.3)	10	28162 (4.25)
	bold	10	10	1097 (0.21)	5	27 (0.02)	10	1124 (0.17)
	italic bold	10	0	0 (0)	0	0 (0)	0	0 (0)
	italic	10	10	152 (0.03)	7	23 (0.02)	10	175 (0.03)
	type writer	10	0	0 (0)	1	4 (0)	1	4 (0)
operator (binary)		23	5	141 (0.03)	14	8447 (6.09)	15	8588 (1.3)
operator (relational)		69	1	7 (0)	32	9823 (7.08)	32	9830 (1.48)
others	upright	42	11	2326 (0.44)	13	763 (0.55)	21	3089 (0.47)
	bold	16	2	34 (0.01)	0	0 (0)	2	34 (0.01)
parenthesis	upright	20	6	8160 (1.56)	12	24290 (17.51)	12	32450 (4.9)
	bold	20	4	56 (0.01)	2	6 (0)	4	62 (0.01)
point	upright	17	12	21893 (4.18)	10	9353 (6.74)	14	31246 (4.72)
	bold	17	6	583 (0.11)	1	1 (0)	6	584 (0.09)
Roman	upright	52	52	407374 (77.82)	46	4662 (3.36)	52	412036 (62.23)
	bold	52	52	5855 (1.12)	16	830 (0.6)	52	6685 (1.01)
	italic	52	52	61495 (11.75)	52	46894 (33.81)	52	108389 (16.37)
	italic bold	52	0	0 (0)	15	26 (0.02)	15	26 (0)
script		52	0	0 (0)	20	862 (0.62)	20	862 (0.13)
reject		21	4	8 (0)	3	116 (0.08)	4	124 (0.02)
total		1629	311	523454 (100)	354	138688 (100)	487	662142 (100)

4.1.1 Data collection of InftyCDB-3-A

InftyCDB-3-A is the training set used to produce recent versions of InftyReader 7) (Versions 2.0 - 2.5.0). Taking data from more than 300 sources, we have tried to cover as many varieties of characters and symbols as possible. The data was extracted from the following three kinds of sources:

- The books and journals from various publishers listed in Appendix C.
- Fonts used internally on Windows and Macintosh computers, and LaTeX fonts. These are entered into the database by means of scanned printouts.
- Mathematics articles in the Kyushu University Library were individually sought out to provide specimens of infrequently appearing symbols and fonts. To our regret, in these cases there is no record of the document source. However, any two symbols that came from the same book are always assigned the same article ID number.

4.1.2 Data collection of InftyCDB-3-B

InftyCDB-3-B is an extract of InftyCDB-1, which includes data from 20 of its articles. To reduce the number of samples with the same character code, size, and shape, clustering was applied to the data from these 20 articles, reducing the number of data samples to about 70,000. The data are written in the same format as in InftyCDB-3-A.

4.2 Statistics

The statistics of InftyCDB-3-A and InftyCDB-3-B are summarized in **Table 5**. The category variation of InftyCDB-3-A is wider than that of InftyCDB-3-B because of the policy of data collection. For example, the former includes all of the 26

Table 3 Statistics of InftyCDB-2. (4 French articles)

type	style	#def cat.	text region #cat.	text region #char (%)	math region #cat.	math region #char (%)	total #cat.	total #char (%)
accent		14	0	0 (0)	3	145 (2.36)	3	145 (0.39)
arrow		16	0	0 (0)	1	30 (0.49)	1	30 (0.08)
big symbol		18	0	0 (0)	7	137 (2.23)	7	137 (0.37)
blackboard bold		52	0	0 (0)	1	17 (0.28)	1	17 (0.05)
calligraphic		26	0	0 (0)	0	0 (0)	0	0 (0)
German	upright	52	0	0 (0)	0	0 (0)	0	0 (0)
	bold	52	0	0 (0)	0	0 (0)	0	0 (0)
Greek	upright	11	0	0 (0)	4	116 (1.89)	4	116 (0.31)
	bold	11	0	0 (0)	0	0 (0)	0	0 (0)
	italic	29	1	2 (0.01)	18	432 (7.03)	18	434 (1.16)
	italic bold	29	0	0 (0)	0	0 (0)	0	0 (0)
extended Latin	upright	127	0	0 (0)	0	0 (0)	0	0 (0)
	bold	127	0	0 (0)	0	0 (0)	0	0 (0)
	italic	127	0	0 (0)	0	0 (0)	0	0 (0)
	italic bold	127	0	0 (0)	0	0 (0)	0	0 (0)
Latin	upright	55	15	702 (2.24)	1	1 (0.02)	15	703 (1.88)
	bold	55	4	21 (0.07)	0	0 (0)	4	21 (0.06)
	italic	55	10	85 (0.27)	0	0 (0)	10	85 (0.23)
	italic bold	55	1	2 (0.01)	0	0 (0)	1	2 (0.01)
ligature	bold	9	1	1 (0)	0	0 (0)	1	1 (0)
	italic	9	1	4 (0.01)	0	0 (0)	1	4 (0.01)
	italic bold	9	0	0 (0)	0	0 (0)	0	0 (0)
	upright	9	3	39 (0.12)	0	0 (0)	3	39 (0.1)
numeric	upright	10	10	835 (2.67)	5	503 (8.18)	10	1338 (3.57)
	bold	10	10	206 (0.66)	0	0 (0)	10	206 (0.55)
	italic bold	10	0	0 (0)	0	0 (0)	0	0 (0)
	italic	10	0	0 (0)	1	2 (0.03)	1	2 (0.01)
	type writer	10	0	0 (0)	0	0 (0)	0	0 (0)
operator (binary)		23	4	25 (0.08)	11	372 (6.05)	11	397 (1.06)
operator (relational)		69	2	4 (0.01)	13	395 (6.42)	15	399 (1.07)
others	upright	42	5	149 (0.48)	5	94 (1.53)	9	243 (0.65)
	bold	16	1	4 (0.01)	0	0 (0)	1	4 (0.01)
parenthesis	upright	20	4	411 (1.31)	9	967 (15.73)	9	1378 (3.68)
	bold	20	0	0 (0)	0	0 (0)	0	0 (0)
point	upright	17	10	1521 (4.86)	9	363 (5.9)	12	1884 (5.03)
	bold	17	3	107 (0.34)	0	0 (0)	3	107 (0.29)
Roman	upright	52	52	22113 (70.67)	30	662 (10.77)	52	22775 (60.83)
	bold	52	38	619 (1.98)	2	8 (0.13)	38	627 (1.67)
	italic	52	50	4429 (14.15)	45	1860 (30.25)	51	6289 (16.8)
	italic bold	52	10	12 (0.04)	4	30 (0.49)	13	42 (0.11)
script		52	0	0 (0)	2	14 (0.23)	2	14 (0.04)
reject		21	0	0 (0)	0	0 (0)	0	0 (0)
total		1629	235	31291 (100)	171	6148 (100)	306	37439 (100)

categories of blackboard bold, but the latter does not.

5. Conclusion

Three ground-truthed mathematical character and symbol image databases, called InftyCDB-1, InftyCDB-2, and InftyCDB-3, have been described. Ground truth information, including category, font style, and size, was attached manually to each character and symbol in the databases. Furthermore, in InftyCDB-1 and InftyCDB-2, data representing mathematical structure was also attached. This data will be useful for the development and evaluation of mathematical structure analysis modules. The databases are large enough to develop and evaluate OCR engines for mathematical documents.

The databases are publicly availably, subject to the conditions of use in Appendix D.

References

1) H.-J. Lee and J.-S. Wang, "Design of a mathematical expression understanding system," *Pattern Recognition Letters*, **18(3)**, 289–298, 1997.
2) M. Okamoto, H. Imai, and K. Takagi, "Performance evaluation of a robust method for mathematical expression recognition," *Proc. Int. Conf. Document Analysis and Recognition*, 121–128, 2001.
3) J. Mitra, U. Garain, B. B. Chaudhuri, K. Swamy, and T. Pal, "Automatic understanding of structures in printed mathematical expressions," *Proc. Int. Conf. Document Analysis and Recognition*, 540–544, 2003.
4) Y. Eto and M. Suzuki, "Mathematical formula recognition using virtual link network," *Proc. Int. Conf. Document Analysis and Recognition*, 762–767, 2001.
5) S. Uchida, A. Nomura, and M. Suzuki, "Quantitative analysis of mathematical documents," *Int. J Document Analysis and Recognition*, **7(4)**, 211-218, 2005.
6) Masakazu Suzuki, Seiichi Uchida, and Akihiro Nomura,

Table 4 Statistics of InftyCDB-2. (7 German articles)

type	style	#def cat.	text region #cat.	text region #char (%)	math region #cat.	math region #char (%)	total #cat.	total #char (%)
accent		14	0	0 (0)	4	182 (1.96)	4	182 (0.23)
arrow		16	0	0 (0)	2	127 (1.37)	2	127 (0.16)
big symbol		18	0	0 (0)	6	55 (0.59)	6	55 (0.07)
blackboard bold		52	0	0 (0)	0	0 (0)	0	0 (0)
calligraphic		26	0	0 (0)	0	0 (0)	0	0 (0)
German	upright	52	0	0 (0)	16	374 (4.03)	16	374 (0.48)
	bold	52	0	0 (0)	0	0 (0)	0	0 (0)
Greek	upright	11	0	0 (0)	7	72 (0.78)	7	72 (0.09)
	bold	11	0	0 (0)	0	0 (0)	0	0 (0)
	italic	29	2	21 (0.03)	19	355 (3.82)	19	376 (0.48)
	italic bold	29	0	0 (0)	1	1 (0.01)	1	1 (0)
extended Latin	upright	127	0	0 (0)	0	0 (0)	0	0 (0)
	bold	127	0	0 (0)	0	0 (0)	0	0 (0)
	italic	127	1	1 (0)	0	0 (0)	1	1 (0)
	italic bold	127	0	0 (0)	0	0 (0)	0	0 (0)
Latin	upright	55	9	860 (1.26)	0	0 (0)	9	860 (1.11)
	bold	55	4	16 (0.02)	0	0 (0)	4	16 (0.02)
	italic	55	8	119 (0.17)	0	0 (0)	8	119 (0.15)
	italic bold	55	0	0 (0)	0	0 (0)	0	0 (0)
ligature	bold	9	0	0 (0)	0	0 (0)	0	0 (0)
	italic	9	4	16 (0.02)	0	0 (0)	4	16 (0.02)
	italic bold	9	0	0 (0)	0	0 (0)	0	0 (0)
	upright	9	4	58 (0.08)	0	0 (0)	4	58 (0.07)
numeric	upright	10	10	1705 (2.49)	9	776 (8.35)	10	2481 (3.19)
	bold	10	10	132 (0.19)	0	0 (0)	10	132 (0.17)
	italic bold	10	0	0 (0)	0	0 (0)	0	0 (0)
	italic	10	10	55 (0.08)	2	8 (0.09)	10	63 (0.08)
	type writer	10	0	0 (0)	0	0 (0)	0	0 (0)
operator (binary)		23	4	20 (0.03)	10	547 (5.89)	11	567 (0.73)
operator (relational)		69	1	3 (0)	15	580 (6.24)	15	583 (0.75)
others	upright	42	7	449 (0.66)	5	75 (0.81)	11	524 (0.67)
	bold	16	2	7 (0.01)	0	0 (0)	2	7 (0.01)
parenthesis	upright	20	4	934 (1.36)	7	1709 (18.4)	7	2643 (3.4)
	bold	20	0	0 (0)	2	2 (0.02)	2	2 (0)
point	upright	17	8	2613 (3.81)	9	792 (8.53)	12	3405 (4.38)
	bold	17	2	30 (0.04)	0	0 (0)	2	30 (0.04)
Roman	upright	52	52	51873 (75.7)	31	327 (3.52)	52	52200 (67.08)
	bold	52	46	858 (1.25)	7	45 (0.48)	46	903 (1.16)
	italic	52	50	8735 (12.75)	48	3210 (34.55)	52	11945 (15.35)
	italic bold	52	10	14 (0.02)	4	32 (0.34)	14	46 (0.06)
script		52	0	0 (0)	2	19 (0.2)	2	19 (0.02)
reject		21	1	3 (0)	1	2 (0.02)	1	5 (0.01)
total		1629	249	68522 (100)	207	9290 (100)	344	77812 (100)

"A ground-truthed mathematical character and symbol image database," *Proc. Int. Conf. Document Analysis and Recognition* **2**, 675–679, 2005.

7) `http://www.inftyproject.org`

Appendix

A Miscellaneous of InftyCDB-1

A1 List of articles

The documents contained in InftyCDB-1 comprise 30 English language articles on pure mathematics:

• 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.

Table 5 Statistics of InftyCDB-3.

type	style	#def cat.	InftyCDB-3-A		InftyCDB-3-B	
			#cat.	#char (%)	#cat	#char (%)
accent		14	5	1721 (0.91)	2	58 (0.08)
arrow		16	15	3731 (1.98)	5	285 (0.4)
big symbol		18	4	1482 (0.79)	2	65 (0.09)
blackboard bold		52	26	1291 (0.68)	6	77 (0.11)
calligraphic		26	26	1084 (0.57)	17	103 (0.15)
German	upright	52	0	0 (0)	0	0 (0)
	bold	52	0	0 (0)	0	0 (0)
Greek	upright	11	10	7245 (3.84)	9	329 (0.47)
	bold	11	0	0 (0)	0	0 (0)
	italic	29	24	23321 (12.36)	23	1580 (2.24)
	italic bold	29	0	0 (0)	0	0 (0)
extended Latin	upright	127	0	0 (0)	0	0 (0)
	bold	127	0	0 (0)	0	0 (0)
	italic	127	0	0 (0)	0	0 (0)
	italic bold	127	0	0 (0)	0	0 (0)
Latin	upright	55	7	33 (0.02)	7	102 (0.14)
	bold	55	0	0 (0)	0	0 (0)
	italic	55	1	1 (0)	4	22 (0.03)
	italic bold	55	0	0 (0)	0	0 (0)
ligature	upright	9	7	895 (0.47)	0	0 (0)
	bold	9	0	0 (0)	0	0 (0)
	italic	9	7	349 (0.18)	0	0 (0)
	italic bold	9	0	0 (0)	0	0 (0)
numeric	upright	10	10	14689 (7.78)	10	6618 (9.37)
	bold	10	0	0 (0)	0	0 (0)
	italic	10	10	465 (0.25)	10	122 (0.17)
	italic bold	10	0	0 (0)	0	0 (0)
	type writer	10	0	0 (0)	0	0 (0)
operator (binary)		23	20	13164 (6.97)	15	3246 (4.6)
operator (relational)		69	56	13781 (7.3)	28	1720 (2.43)
others	upright	42	29	6454 (3.42)	20	501 (0.71)
	bold	16	0	0 (0)	0	0 (0)
parenthesis	upright	20	17	10754 (5.7)	8	2805 (3.97)
	bold	20	0	0 (0)	0	0 (0)
point	upright	17	6	5813 (3.08)	5	746 (1.06)
	bold	17	0	0 (0)	0	0 (0)
Roman	upright	52	52	52268 (27.69)	52	38393 (54.35)
	bold	52	0	0 (0)	0	0 (0)
	italic	52	52	30211 (16.01)	52	13865 (19.63)
	italic bold	52	0	0 (0)	0	0 (0)
script		52	0	0 (0)	0	0 (0)
reject		21	0	0 (0)	0	0 (0)
total		1629	384	188752 (100)	275	70637 (100)

A2 Detail of Categories

A2.1 Categories of "big symbol" Type

The "big symbol" type consists of the following 18 pre-defined categories: $\sqrt{\ }$, $\sum$, $\prod$, $\coprod$, $\bigcup$, $\bigcap$, $\bigvee$, $\bigwedge$, $\bigoplus$, $\bigotimes$, $\int$, $\oint$, $\iint$, $\iiint$, $\iiiint$, $\int\cdots\int$, ——(fraction bar), and $\lrcorner$ (continued fraction).

A2.2 Categories of "extended Latin" Type

The "extended Latin" type in InftyCDB-1 consists of 364 pre-defined categories. Among them, of 110 rather common categories are the following 55 and their italic versions: À, Á, Â, Ã, Ä, Å, Ç, È, É, Ê, Ë, Ì, Í, Î, Ï, Ñ, Ò, Ó, Ô, Õ, Ö, Ø, Ù, Ú, Û, Ü, Ý, ß, à, á, â, ã, ä, å, ą, è, é, ê, ë, ì, í, î, ï, ñ, ò, ó, ô, õ, ö, ø, ù, ú, û, ü, and ý. Among the 466 extended Latin characters in InftyCDB-1, 451 come from the 110 categories noted above. The remaining 15 come from the other 254 pre-defined categories, that are composed of 127 very rare categories and their italic versions. The 15 characters are: Š (8 characters), Č (3), Ş (1), ţ (1), ş (1), and č (1). The 5 most frequent categories of extended Latin are: é (128 characters), É (75), *é* (43), ü (25), and ö (23).

A2.3 Categories of "operator" Type

The "operator" type consists of 92 pre-defined categories, which are divided into 69 relational and 23 binary operators. The relational operators are: $<$, $=$, $>$, $\in$, $\ni$, $\subset$, $\supset$, $\Subset$, $\Supset$, $\leq$, $\geq$, $\leqq$, $\geqq$, $\ll$, $\gg$, $\equiv$, $\div$, $\neq$, $\not\subset$, $\not\supset$, $\not\subseteq$, $\not\supseteq$, $\sim$, $\approx$, $\simeq$, $\cong$, $\prec$, $\succ$, $\propto$, $\parallel$, $\perp$, $\vdash$, $\dashv$, $\lhd$, $\rhd$, $\doteqdot$, $\fallingdotseq$, $\subseteq$, $\supseteq$, $\subseteqq$, $\supseteqq$, $\subsetneq$, $\supsetneq$, $\subsetneqq$, $\supsetneqq$, $\preceq$, $\succeq$, $\models$, $\not\equiv$, $\not\leq$, $\not\geq$, $\not\leqq$, $\not\geqq$, $\not\equiv$, $\lesseqgtr$, $\gtreqless$, $\not\ni$, $\pitchfork$, $\lessapprox$, $\gtrapprox$, $\smile$, $\frown$, $\asymp$, $\barwedge$, $\subsetneqq$, $\supsetneqq$, $\nparallel$, $\not\prec$, and $\not\succ$. The binary operators are: &, $*$, $+$, $-$, $/$, $\setminus$, $\pm$, $\mp$, $\times$, $\div$, $\oplus$, $\ominus$, $\otimes$, $\cup$, $\cap$, $\vee$, $\wedge$, $\dotplus$, $\ltimes$, $\rtimes$, $\odot$, $\hat{}$, and $\circ$.

A2.4 Categories of "others" Type

The "others" type consists of the following 42 pre-defined categories: !, #, $, %, : (colon), ; (semi-

colon), ?, @, ¥, _ (under score), ∞, ∂, ∇, ℓ, $\hbar$, $\Re$, $\Im$, $\aleph$, $\emptyset$, $\forall$, $\exists$, $\neg$, $\angle$, $\triangle$, $\square$, $\measuredangle$, ■, §, †, ‡, ¶, ♭, ♮, ©, $\wp$, $\star$ (star), – (hyphen), — (long hyphen), $\hbar$, "differential", "ImaginaryNumber", and "NapierNumber". The 5 most frequent categories of this type are: – (hyphen), : (colon), ∂, ∞, and ; (semicolon).

The following 16 categoies have bold style: !, #, $, %, ?, @, ¥, §, †, ‡, ♭, ♮, ©, $\star$, –, and —.

B List of articles in InftyCDB-2

The documents contained in InftyCDB-2 comprise 26 English, 4 French, and 7 German language articles on pure mathematics:

[**English articles**]

• Collected Works, **1**, 921-945. • Can. J. Math., **48(6)**, 1286-1295,1996. • Canad. Math. Bull., **38(4)** 390-395 ,1995. • Compositio Mathematica, **38(3)**, 253-276 ,1979. • jahresbericht der Deutshen mathematiker Vereinigung, **80**, 111-128, 1978. • International Mathematics Reserch Notices, **14**, 699-703, 1996. • Inventationes mathematicae, **18**, 119-141, 1972. • *ibid.*, **89**, 225-246 ,1987. • *ibid.*, **95**, 31-62, 1989. • J. Algebraic Geometry, **6**, 671-195, 1997. • manuscripta mathematica, **89**, 439-459, 1996. • Math. USSR Izvestiya **35(1)**, 61-81, 1990. • *ibid.*, **38(2)**, 435-437 ,1992. • Math. USSR Sbornik, **55(1)**, 55-70, 1986. • Mathematische Zeitschrift, 449-456 ,1985. • Proc. Am. Math. soc., 562-568 ,1969. • *ibid.*, 183-192 ,1978. • Soviet Math. Dokl, **20**, 1262-1266, 1979. • Soviet Math. Dokl, **42(2)**, 636-640, 1991. • The journal of the London Math. Soc., 445-451 ,1968. • American Journal Mathematics, 132-138, 140-141, 143-148, 1979. • J. American Mathematical Society, **1(4)**, 699-727, 730-735, 752-757, 778, 1988. • Math. Ann. 34-38, 40-46 ,1964. • Pacific journal of mathematics, **127(1)**, 141-142, 144-154, 1987. • Trans. American Mathematical Society, 214-215, 217-227, 1964. • Topology, **11**, 151-153, 156-158, 1972.

[**French articles**]

• C. R. Acad. Sci. Paris, t., **304**, I, 191-194, 1987. • C. R. Acad. Sci. Paris, t., **306**, I, 535-538 ,1988. • Duke Mathematical Journal, **52(1)** 157, 178, 182-183, 188, 190, 196-197 ,1985. • J. Math. Soc. Japan, **26(2)**, 241-244, 256-257, 1974.

[**German articles**]

• Annals of Mathematics, **68(2)**, 393, 404-405, 410, 413, 443, 1958. • Commentarii Mathematici Helvetici, **46(1)**, 48, 55-56, 63-64, 1971. • jahresbericht der Deutshen mathematiker Vereinigung, **71**, 48-50, 52, 54, 1969. • *ibid.*, **89**, 81, 83, 85-87, 1987. • Math. Annalen, **135**, 219, 225, 227, 233-234, 1958. • Mathematische Zeitschrift, **65**, 175, 182-188, 1956. • Proc. Int. Congr. Mathematicians, 86-88, 91-92, 101, 1962.

C List of books and journals in InftyCDB-3

A part of data of InftyCDB-3-A were extracted from the following books and journals from various publishers.

• Journal of Approximation Theory (Academic Press) • Journal of Differential Equations (Academic Press) • Journal of Functional Analysis (Academic Press) • Introduction to Algebraic Curves (American Mathematical Society) • 30 Lectures in Mathematics Series: 30 Lectures on Complex Numbers (Japanese, Asakura Shoten) • Complexity Theory of Real Functions (Birkhauser) • Journal of Fluid Mechanics (Cambridge University Press) • Image Processing: Mathematical Methods and Applications (Clarendon Press) • Automorphisms of Affine Spaces (Kluwer Academic Publishers) • Typesetting samples (Kotobuki Printing Co.) • Hyperbolic Manifolds and Holomorphic Mappings (Marcel Dekker, Inc.) • Complex Analysis (Springer-Verlag) • Introduction to Complex Hyperbolic Spaces (Springer-Verlag) • Knot Theory (Japanese, Springer-Verlag) • Rational Points on Elliptic Curves (Springer) • Communications on Pure and Applied Mathematics (John Wiley & Sons) • Explanation of Braille Codes for Mathematics (Japanese, Japanese Braille Association) • Value Distribution Theory (D. Van Nostrand Co.) • An Introduction to Differentials and Integrals (Yuuseisha)

D Conditions of use

Each of the databases may be used free of charge in the research, development, and testing of OCR systems for scientific documents. Please refrain from other forms of use.

You are permitted to use the databases for commercial OCR engines. However, if you do use it for such a purpose, you must contact Masakazu Suzuki (`suzuki@math.kyushu-u.ac.jp`).

You may not sell this database, or any part of it. If you incorporate it into a new distribution, please include the database in its entirety, without transformation or modification.

We hope this database will contribute to your development or research. If it does, please include the Infty Project URL 7) with your publication or release.