研究報告

# OCRによるデータベース作成

轡田　收*　高瀬　誠†

## 1　本プロジェクトの目的

従来、書籍などの形で公開されてきた文献資料が、近年になって様々な形で電子化され、計算機上でも利用可能となってきている。そうした電子化された資料は、これまで手作業に頼らざるをえなかった資料研究活動には、作業効率を上げるのに極めて有益なものとなってきている。しかしその反面、提供されるデータがある特定の計算機環境に依存した形式になっているために、せっかく収集したデータも再利用が極めて困難であるなどの問題も少なくない[1]。

本プロジェクトでは、現在刊行中の Wilhelm Dilthey 著作集[2]を OCR システムを利用してテキスト型データへ変換し、テキスト型データベースにする事で、Dilthey に関する今後の基礎研究を効率的に行えるようにする事を目的としているが、同時に様々な計算機環境上でも統一的に扱えるテキスト型データベースの形式を考察し実現する事を目指すものでもある。

## 2　テキスト型データベースに必要な条件

テキスト型データベースが持つべき条件を一言で言うならば、**「現行の全ての計算機上で利用可能なデータであること」**となろう。これを具体的に見ると次のようになる:

- 特定のハードウェアやソフトウェアの利用を強要するデータ構造ではない。

  特定のワープロ・ソフトや検索ソフトの専用データは、現行の全ての計算機上で利用可能なデータにはならない。なぜならそうしたソフトウェアは特定の計算機環境を想定して設計されており、他の環境に対する配慮はまず殆んどなされていないからである。また、そうしたソフトウェアは当然の事ながら特定のハードウェアを要求する事になり、マルチ・プラットフォームでのデータ処理に全く適さない。

- 特定のハードウェアやソフトウェアに依存した文字コードを使用していない。

  いわゆる「外字」などの、特定の環境に依存した文字コードを使ってデータを作成すれば、そのデータは他の環境では正常な状態では利用できない。最悪の場合、システムを停止させることにもなりかねない。

*学習院大学非常勤講師、E-Mail: Makoto.Takase@gakushuin.ac.jp
†文学部教授、E-Mail: Osamu.Kutsuwada@gakushuin.ac.jp

[1]特に計算機上では多バイトで表現されている日本語などの文字コードと、ドイツ語などの非英語圏の文字コードとでは互いに干渉を起こしてしまい、事実上混在を不可能にしてしまっている。見かけ上混在が可能なように見えるものもない訳ではないが、それは特定の環境に依存した形で見かけだけ出来ているようにごまかしているに過ぎず、そこで作成されたデータを他の計算機環境へ移植するのはほとんど不可能な状態である。

[2]Dilthey, Wilhelm: Gesammelte Schriften, B. G. Teubner Verlagsgesellschaft, Stuttgart, Vandenhoeck & Ruprecht, Göttingen.

テキスト型データベースを様々なプラットフォームの上で利用するには、使用されている文字コード体系が共通のものである必要がある。現行の文字コード体系では対応仕切れないものがあってもそれは、複数の文字コードを組み合わせた「文字列」により表現することで対処すればよい。その際の表現方法は、国際的になるべく広く知られていて、どのプラットフォーム上でも利用可能な方式に従うのがよい[3]。

- データは統一的に記述されている。

  自明の事であるが、データベースとして有効に活用するためには、特に検索などの処理がきちんとできなければ意味がない。データ全体にわたって統一的な記述がなされていることによって、それは実現される。また、検索等の処理がどのプラットフォーム上でも統一的に行えるようにするためにも必要な事である。

  さらに現行の文字コードの体系が未来永劫にわたって不変であるとは到底考えられない。将来何らかの形で計算機で扱われる文字コードが変更される際にも、データ全体にわたって統一的な記述がなされていれば、変更は極めて容易に行える[4]。

- ネットワークを介しての利用もできる必要がある。

  ネットワークの発達に伴い、様々な地域の様々な環境を持つ計算機が、単一のデータを共有することも比較的容易になってきている。データベースは広く公開され、より多くの人達が利用できなければ、それを作成する意味はない。そしてより多くの利用者に対してデータを公開するには、ネットワークを活用するのがもっとも効果的である。

  しかしテキスト型データベースは、その元となる資料の性質上、あらかじめ著作権者の承認を必要とする。公開に先立って著作権者との交渉は不可欠な要素となる。

- 検索などにより抽出したデータは容易に再利用できる。

  この種のテキスト型データベースを利用する主な目的は、得られたデータを利用して、学術的な調査・研究を行うことである。得られたデータは当然そうした調査・研究を発表する際にも利用される。その際、データの構造や記述方法に問題があると、発表の際の利用を妨げる事になり、研究作業効率を落してしまう。そのような事を起こさないためにも、データの構造や記述には、移植性の高い方式を導入すべきである。

## 3 作成するテキスト型データベースの基本仕様

2 章で述べた条件に基づいて、本プロジェクトでは W. Dilthey 著作集のデータを以下に述べるような仕様のものにした。

### 3.1 データ

W. Dilthey 著作集中のテキストのうち Dilthey 自身の手によると思われるテキスト・註などをデータとして収録する。データはテキスト型データとし、検索用のツールには取り敢えず awk を

[3] 例えば TEX の記述方法などがこれに該当する。

[4] 現在、国際標準化機構 (ISO) を中心に、世界各国で用いられている計算機用の文字コードを見直し、国や地域に依存せずに計算機上で多言語を扱えるようにしようという国際規格の制定作業が進められている。一般に ISO10646 と呼ばれているこの規格案には、様々な問題が指摘されてはいるものの、これに合わせる形で、現在 JIS でも日本語文字コード規格の改訂作業が進んでいる。これについては [13]、[17] を参照。

利用することにする。これは以下の理由による:

- awk で利用できるのは、ごく普通のテキスト型データであるので、そのデータは awk 以外のツールでも容易に取り扱うことができるため。
- awk — とりわけ gnu 版 awk — は現在ではさまざまなプラットフォーム上に移植されており、どのプラットフォームでも共通のオペレーションが行えるため。
- awk のもつ正規表現による検索機能は、テキスト型データに対して極めて柔軟に、かつ有効に機能すると考えられるため。

## 3.2 データ構造

### 著者名・巻数

オリジナルデータの 1 巻分を 1 つのファイルに収め、ファイル名+拡張子 (DOS の場合、8+3 文字) によって原著者名、巻数を表すことにする。つまり今回の場合は `dilthey.1` などのようになる。

### データ内部

データはごく普通のテキスト型データであるが、取り敢えずは awk での処理を前提として、以下に述べる複行レコード形式[5]をとる事とする。

### 前提となる基本構造 (複行レコード形式):

- データは複数のレコードで構成される。
- 各レコードの区切り子は、二つの連続する改行で表す。
- 1 レコード内の各フィールドの区切り子は、改行一つで表す。

### レコードの仕様

- 1 レコードは複数フィールド+レコード区切り子で構成される。
- 1 レコードをオリジナルテキストの 1 ページと対応させる。
- 1 レコードの構成要素は
  1. 著作集内のページ番号
  2. 当該ページのテキスト (オリジナルの 1 行 = 1 フィールド)
  3. レコード区切り子 (連続する改行 2 つ)

  とする。

レコードの概念図を付録 A に示しておく。

[5]複行レコード形式に関しては、[5], [18], [22] を参照。また正規表現に関してはこの他に [15] も参照。

## フィールドの仕様

1. 1 フィールドは文字列+フィールド区切り子で構成される。
2. 1 フィールドをオリジナルテキストの 1 行と対応させる。
3. ただし、第一フィールドは著作集内での実際のページ番号を入れるものとする。
4. テキスト中の段落の最初は、3 バイトのスペースを入れる。
5. テキスト中の章などのタイトルは、行頭にタブコードを入れて他と区別する。
6. 欧文特種文字は原則として TEX の記述方式に従うものとする。ただしドイツ語に関しては、ドイツ語用スタイルファイル german.sty による記述方式をとるものとする。(付録 B および [4] 参照。これによりデータは現行のあらゆる環境下で利用可能となる。)
   例:
   オリジナル: Dieser Tag, der dem großen Leibniz gewidmet ist, legt die Erwägng
   ⇓
   データ: `Dieser Tag, der dem gro"sen Leibniz gewidmet ist, legt die Erw"agung`
7. ダッシュは "-" 3 バイトで、範囲指定のハイフンは "-" 2 バイトで表わす。
   例:

   | オリジナル: | | データ: |
   |---|---|---|
   | ··· zu können — weil ··· | ⟹ | ··· `zu k"onnen --- weil` ··· |
   | 1545-1563 | | `1545--1563` |

8. 分綴により二行にまたがっているものは、分綴を解消し、その語が始まっている行に入れるものとする。
   例:

   | オリジナル: | | データ: |
   |---|---|---|
   | ··· gewidmet ist, legt die Er-↵ wägung nahe, ··· | ⟹ | ··· `gewidmet ist, legt die Erw"agung`↵ `nahe,` ··· |

9. 脚注は以下のように扱う:
   (a) 本文中の註を指示する箇所には、`\footnotemark[1]` で註がある事を示し、レコードの最後に `\footnotetext[1]{}` で註のテキストを示す。
   (b) 註のテキストは本文の場合と同様に改行して表すものとする。
10. gesperrt で表記されている箇所は当面 {\em } で表す。二行にまたがっているものは各行で {\em } の記述を行う。
    例:
    オリジナル: Soll sich diese Körperschaft als l e b e n d i g e E i n h e i t ···
    ⇓
    データ: `Soll sich diese K"orperschaft als {\em lebendige Einheit}` ···

11. itaric で表記されている箇所は当面 {\it　　} で表す。二行にまたがっているものは各行で {\it　　} の記述を行う。
    例:
    オリジナル: Soll sich diese Körperschaft als *lebendige Einheit* ···
    ⇓
    データ: `Soll sich diese K"orperschaft als {\it lebendige Einheit}` ···

12. テキスト中にギリシア語が出て来た場合、とりあえずその位置には $$ を入力しておく[6]。

具体例はこの他に付録 C にも掲載しておく。

# 4　1994 年度までに作成したデータについて

## 4.1　利用した計算機環境

本プロジェクトで使用した計算機環境は以下の通りである:

- コンピュータ本体: Dell 466/L (含: 80486DX2-66MHz, 24MB RAM, 230MB HDD 他)
- イメージ・スキャナ: Hewlett Packard HP-ScanJetIIp
- OS: IBM DOSJ Ver.5.02D/V + 英語版 Microsoft Windows Ver. 3.10
- 欧文 OCR ソフトウェア: OMNIPAGE Professional
- テキスト・エディタ: MIFES Ver. 5.5
- その他: フリーソフトウェア多数 (jgawk, Microspell, ASCII 日本語 pTEX 他)

## 4.2　作業手順

1. まず Windows 環境下で OCR ソフトを起動し、対象となっているオリジナルテキストを読み取ってテキストファイルへ出力させる。
2. awk script 等を利用してデータをある程度整形する。
3. OCR システムの読み取りエラーのチェック並びにデータの整形作業を行う[7]。
4. スペルチェッカー (Microspell) を用いて再度エラーの有無をチェックしデータ形式の最終的整形を行う。

## 4.3　1994 年度までに作成したデータ

上記の作業手順に基づいて 1994 年度までに作成したデータは、Dilthey 著作集 20 巻のうち第 1 巻〜第 5 巻および第 19 巻の合計 6 巻分である。なおデータ作成にあたり、法政大学大学院哲学専攻科に所属の伊藤　直樹 氏より、Dilthey 著作集のデータ化についての先行するプロジェクト等に関する情報の提供を頂いた。また伊藤氏の仲介で、豊田短期大学日本文学科の森本 司 氏

[6] これは、入力・チェック等の負担軽減のための暫定的措置で、ギリシア文字の扱いについては今後さらに検討する。
[7] この段階で多くの学生諸君の協力を得られたことに深く感謝する。

から第 19 巻の入力済みデータを提供して頂いた。また、東京大学教養学部の麻生 健 氏より第 5 巻の入力済みデータ並びに OCR 作業用資料等を提供して頂いた。三氏からはこの他にもアドヴァイスを頂いており、データ作成に大変役立った。ここに深く感謝したい。

## 5 今後の課題

これまでの作業の結果、今後次に挙げる事を行う必要が出てきている:

- 既刊の Dilthey 著作集の残りのすべての巻についてのデータ化
- テキスト中のギリシア語の表記の確定と既存データへの追加

  現在、LaTeX の 次期 Version である LaTeX $2_\varepsilon$ の開発が、Mainz 大学の Frank Mittelbach を中心として進められている。この LaTeX $2_\varepsilon$ によるギリシア語の記述方法を検討し、それに合せていく形で作業を進めていきたい。
- 検索用スクリプトの改良

  作業の途中で、jgawk によるごく簡単な検索スクリプトを作成し、検索のテストを行ってみた。awk の正規表現を用いた文字列検索を用いたこのテスト用スクリプトでは、当該検索文字列がヒットした著作集の巻数、ページ数、行数などが表示されるようになっているが、ほぼ予想通りの結果を得る事ができた。しかしながらなお改善の余地があるものであるので、今後改善した後公表していきたい。
- 作成されたデータに基づくインデクスの作成

以上をクリアした段階で、出版元とのデータ等公開のための交渉に入りたいと考えている。

## 謝辞

最後に、本プロジェクトは学習院大学計算機センター特別研究費によって運営されてきたものであることをここに記しておく。

## 参考文献

[1] Knuth, Donald Ervin: The TeXbook (COMPUTERS & TYPESETTING Volume A), Addison Wesley Publishing Company, 1986.

[2] Knuth, Donald Ervin: The METAFONTbook (COMPUTERS & TYPESETTING Volume C) Addison Wesley Publishing Company, 1986.

[3] Lamport, Leslie: LaTeX: A Document Preparation System User's Guide & Reference Manual Addison Wesley Publishing Company, 1986.

[4] Raichle, Bernd: Kurzbeschreibung — german.sty (Version 2.4a) 1992, in: emTeX Distribution von Version 25. Sep. 1990. Mit Änderungen vom 23. Jun. 1992.

[5] Aho, A. V./Kernighan, B. W./Weinberger, P.J./足立　高徳　訳: 「プログラミング言語 AWK」　株式会社　トッパン, 1989.

[6] Knuth, D. E. 著／鷺谷　好輝　訳／斉藤　信男　監修:「TEX ブック」　アスキー出版局, 1989.

[7] Knuth, D. E. 著／鷺谷　好輝　訳:「METAFONT ブック」　アスキー出版局, 1994.

[8] Leslie Lamport 著／ Edgar Cooke・倉沢良一　監訳／大野俊治・小暮博道・藤浦はる美　訳: 「文書処理システム　LATEX」　アスキー出版局, 1990.

[9] アスキー出版技術部責任編集: 「日本語 TEX テクニカルブック I」　(株) アスキー出版局, 1990.

[10] 阿瀬はる美　著: 「てくてく TEX 〈上〉〈下〉」　アスキー出版局, 1994.

[11] 磯崎　秀樹　著: 「LATEX 自由自在」　サイエンス社, 1992.

[12] 伊藤　和人　著: 「LATEX トータルガイド」　秀和システムトレーティング株式会社, 1991.

[13] 太田　昌孝　著: 「文字コードと国際化」　**bit**, Vol.27, No.6, 共立出版, 1995.

[14] 大野　義夫　編: 「TEX 入門」　共立出版, 1989.

[15] 小山　裕司／斎藤　靖／佐々木　宏／中込　知之　著: 「Linux 入門 — PC 互換機の最新 UNIX 環境」　アジソン・ウェスレイ・パブリッシャーズ・ジャパン (株), 1995.

[16] 坂本　文　著: 「たのしい UNIX — UNIX への招待 —」　アスキー出版局, 1991.

[17] 「しにか　特集 漢字コードと国際標準化」　Vol.4, No.2, 大修館書店, 1993.

[18] 志村　拓／鷲北　賢／西村　克信　共著: 「AWK を 256 倍使うための本」　アスキー出版局, 1993.

[19] すずきひろのぶ　著: 「やさしい LATEX のはじめかた」　オーム社, 1991.

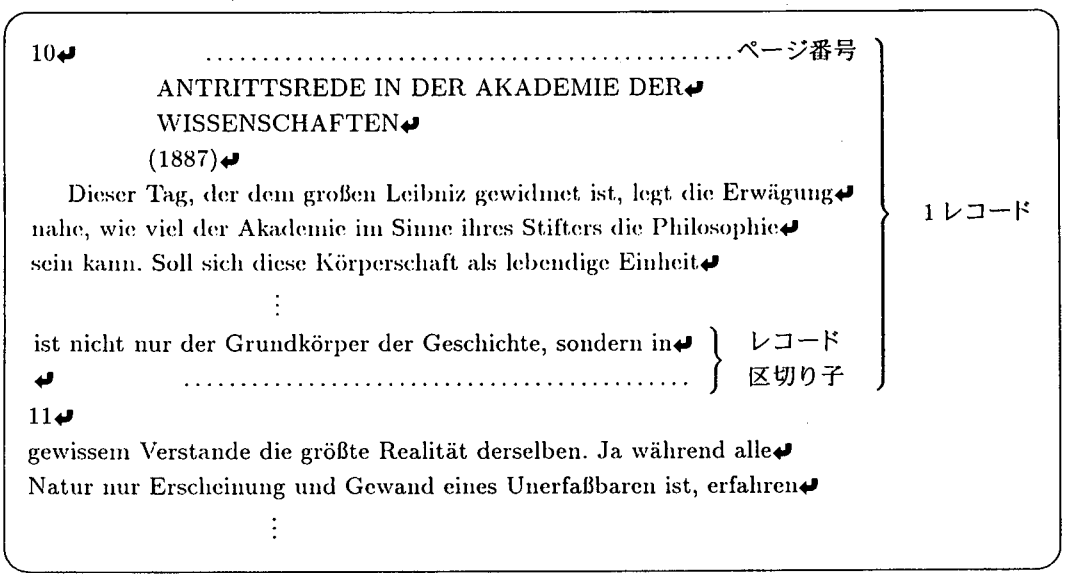
[20] 高瀬　誠: 「計算機による欧文テキスト処理について— 機種依存しない統一的処理についての一考察 —」学習院大学大学院ドイツ文学語学研究 第 18 号, 1994.

[21] 野寺　隆志　著: 「楽々LATEX・第 2 版」　共立出版, 1994.

[22] 松岡　裕典／松井　浩／網本　淳／原　信一郎　著: 「MS-DOS テキストデータ料理学」翔泳社, 1992.

# 付録

## A レコードの概念図




## B 欧文特種文字等の読み替え一覧

| オリジナル | データ | オリジナル | データ |
|---|---|---|---|
| ä | "a | Ä | "A |
| ö | "o | ö | "O |
| ü | "u | ü | "U |
| ï | \"{\i} | „···“ | "‘···"’ |
| ß | "s | ‚···‘ | {\glq}···{\grq} |
| á | \’a | é | \’e |
| à | \‘a | è | \‘e |
| â | \^a | ç | \c{c} |
| | | § | \S |

この他のものについては、[1], [3], [4], [14], [21] 等を参照。

# C　レコード内データの例:

(　　) 内の記号は本文中の「フィールドの仕様」の項の項目番号に対応している。

```
134↵
in den dialektischen Widerstreit zwischen Dogmatism (Metaphysik)↵
und Skepticism verwickelt, die Aufl"osung dieses Widerstreits durch↵
Erkenntnistheorie ist aber der Kriticism.\footnotemark[1]↵ ..... (9a)
   Sowohl diese Theorie von Kant als die von Comte enthalten eine↵
einseitige Auffassung des Tatbestandes. Comte hat die historischen↵

                                    ⋮

   nur den geschichtlichen Tatbestand; an sp"aterer Stelle kann↵
ihm das Ergebnis aus der Analysis des Bewu"stseins zur Best"atigung↵
dienen.↵
                DRITTES KAPITEL↵ ..... タイトルはTABでおくる (5)
        DAS RELIGI"OSE LEBEN ALS UNTERLAGE DER METAPHYSIK.↵
        DER ZEITRAUM DES MYTHISCHEN VORSTELLENS↵
␣␣␣Niemand kann bezweifeln, da"s der Entstehung der Wissenschaften↵ ..... 段落のはじめは3バイト空ける (4)
in Europa eine Zeit vorausgegangen ist, in welcher d.e intellektuelle↵
Entwicklung sich in der Sprache, Dichtung und im mythischen vorstellen↵
sowie im Fortschritt der Erfahrungen des praktischen Lebens↵
vollzog, dagegen eine Metaphysik oder Wissenschaft noch nicht bestand.\footnotemark[2]↵. (9a)
--- Wir treffen die europ"aische Menschheit, ungesondert von↵
den kleinasiatischen Griechen, in intimer Wechselwirkung mit den↵
\footnotetext[1]{↵ ..... (9b)
␣␣␣Kant 2, 241ff.↵
}↵
\footnotetext[2]{↵
␣␣␣Turgot hat zuerst versucht, das Gesetzma"sige in der Entwicklung der Intelligenz↵ ..... (4)
zu entwickeln, da Vicos scienza nuova (1725) sich auf die Entwicklung der Nationen bezieht.↵

                                    ⋮

}↵
↵
135↵
umgebenden Kulturl"andern, sechs Jahrhunderte v. Chr. im "Ubergang↵
zu dem Stadium der Wissenschaft vom Kosmos sowie der Metaphysik↵

                                    ⋮
```