フランス語テクストの保存に関する覚書

重　見　晋　也

文学研究へのコンピュータの活用を推進する上で書かせないのは電子化されたテクストであろう。1次資料のデジタル化は、大きなプロジェクトを計画しなくても、マン・パワーの問題を別にすれば、研究者個人で行うに十分な技術的な発展がなされてきた。

その結果として、中世文学のテクストから現代作家のテクストまで数多くの文学テクストがデジタル化され、有償無償を問わず様々な形で配布されるようになってきている。

このような現状は歓迎すべきであるが、しかしそこには問題があることもまた事実である。すなわち、配布するテクストがデジタル化を行った研究者のコンピュータ環境に依存している点である。次のような例を考えてみればよい。

研究者Aがある文学テクストをデジタル化したとする、AはMacintosh上で市販のソフトウェアxを使ってファイルを保存している。Aが別のBという研究者にファイルを渡すことになった。しかし、BはWindows98を使っており、Windowsでは、xというソフトは動かないため、Aが作ったファイルはBのコンピュータ上では利用することができない。

こうした場合には、確かに、しばしばファイルをテキスト形式という汎用形式に変換すれば問題が解決する。しかし、それはファイルの中身が英語の文章である場合であり、フランス語のファイルではこの方法を使うことはできない。

現在行われている個人レベルでのテクストのデジタル化がはらむファイルの共有に関する問題は、以上のようにはっきりしている。しかし、それに対する解決策は論議されていない。

本稿は、プラットフォームに依存せずにフランス語のテクストファイルを共有するにはどのような方法があるか、その方法を考察するものである。ところで、プラットフォームに依存しないテクストファイルの共有に関しては、Standard Generalized Markup Language（以下SGMLと略す）という国際規格が存在している。それ故、本稿では主としてSGMLを対象として考察を行う。まず最初に、SGMLの基本概念についてふれ、次にSGMLを使ってどのようにフランス語を表示するかについて述べる。そして最後にSGMLを使ったファイル共有の問題点について考察する。

I. SGML

i) SGMLの基本概念

1978年、ANSI（American National Standards Institute）に一つの作業グループが作られ、将来のいかなる処理も可能となるような十分な機能を持ち、曖昧さのないテクスト交換形式とマークアップ言語の開発が始められた。この作業グループはISOのSC18の作業グループへと引き継がれたそうして誕生したのがSGMLである。すなわち、SGMLは電子的な文章を扱うための言語として企画・開発され、1986年に制定、国際規格ISO8879として承認され、1993年にはJISX4151として承認された。

SGMLの基本概念とは非常に明確である。

前述したように、テクスト処理に際してコンピュータで使用されるフォーマットには、OSや作成アプリケーションの違いから互換性がないことが多い。SGMLの基本にある考え方は、こうした非互換性を排除し、デジタル化されたテキストデータを再入力することなく共有し、活用しようというものである。

SGMLは本来、デジタル化された文書を共有するための汎用フォーマットを提供するための規格であったが、Webを記述するための言語としてその機能を限定した形で用いられ、現在ではHTML（Hyper Text Markup Language）の方が人口に膾炙している。しかし、SGMLが目指したデジタル文書の共有という目標は、WWWによってまさしく実現されているともいえる。

ところで、前述した目的を実現するために、SGMLにおいては：

1) 文書の論理構造を表示方法（物理構造ともいう）から分離して記述する。
2) 論理構造記述をDTD（Data Type Definition）により汎用化する。

という2つの考え方に基づいて規格が策定されている。

論理構造については後述するが、ここでいう物理構造とは、ユーザが実際に目にすることになる文書であり、それは印刷されたものやコンピュータのディスプレイに表示されたものなどである。

SGMLでは、印刷や表示の結果というユーザの目に触れる際の形式については、基本的に指定することはできない。すなわち、一般に紙媒体による出版・編集の作業には欠かすことのできないレイアウトという概念は、基本的にSGMLには欠如していることになる。

この点において、現在インターネットを中心に標準的なファイル形式として用いられ始めているAdobe社が提供するPDF（Portable Document Format）形式と比較すると、SGMLは正反対の特徴を持っているといえる。PDFはフォントやレイアウトといった表示方法を含めて、文

書の汎用性を高め、異なったプラットフォームで同一のレイアウトを持つ文書を閲覧することを可能にする。それに対して、SGMLにはレイアウトの概念がないのであり、そこで問題となるのは、文書の内容とその論理構造だけなのである。

以上のように、SGMLによって記述されるのは、デジタル化される文書の論理構造だけである。

ii) SGMLの特徴

SGMLの最大の特徴は、テクストを表示媒体と独立させた形態において格納することを可能にする点にある。

伝統的なテクストの保存形態は、紙を媒体として行われてきた。紙媒体の有用性は、それが常に特別な機器を必要とせず人間が閲覧することができる、という点にあると思われる。

デジタル・テクストの登場は、こうした紙媒体の利点に加えて、大量の文書を対象とした検索が可能である、という固有の特性から、印刷されたテクスト取って代わるだろうと考えられた。しかし実際には、コンピュータが普及してなお紙媒体への依存度は非常に強いと言わざるを得ない。

このような現象の背景として一つ指摘できるのは、デジタル化されたテクストの閲覧には、単にコンピュータという特別な機器が必要というばかりではなく、当該のテクストがデジタル化された環境が違えばテクストを閲覧することが非常に困難になるという問題点を指摘することができる。すなわち使用ソフトやデジタル化に際してのフォーマットの問題や文字コードの問題により、あるコンピュータによってデジタル化され保存されたテクストは、別のコンピュータでは閲覧できないという問題が起こってくるのである。この点は、本稿の冒頭で述べたとおりである。

しかし、この点が克服されたならば、テクストのデジタル化によって、大量の情報が集中管理されるようになるばかりではなく、それらのテクストを対象とした検索や解析が可能になるのであり、電子媒体が紙媒体に取って代わるとはいわないまでも、それに匹敵するメディアとなる可能性を持つのは間違いない。

iii) SGMLとHTML

SGMLはその名前が示すとおりマークアップ言語であり、前述したようにWWWに用いられるHTMLはSGMLから派生した言語である。

マークアップは特定マークアップと一般化マークアップとに分類する事ができる。

特定マークアップの例としては、WYSIWYG (What You See Is What You Get) を実現する現在のワードプロセッシング・ソフトウェアをあげることができる。WYSIWYG型のワードプロセッサは、ユーザがモニタ上に映し出された映像を見ながら文章を編集し、それをモニタのイメージのままプリント結果に反映させることを可能にする。このため、WYSIWIG型ワード

プロセッサによって作成された文書は、必然的にフォントや文字サイズといった情報を含むことになり、それが前述のような複数のプラットフォーム間での互換性の問題を生むことになるのである。

しかし、SGMLがその理論的基盤をおいているのは、特定マークアップではなく、一般化マークアップである。

一般化マークアップが特定マークアップと大きく異なる点は、テクストを章や段落などの構成要素や論理要素によって成り立つもの、という前提にたっているという点である。

SGML自体は、タグと呼ばれるマクロによってテクストをマークアップする、という方法でテクストを記述する。この際に、タグが示しているのは、マークアップされた文字列が、テクスト全体の中でどのような価値を持つかということだけである。たとえば、題名は本文とは異なる価値をテクストにおいて占めているが、SGMLでは、題名を本文とは異なるものとしてマークアップすることにより、題名というテクストの要素を記述するのである。

また、SGMLは文書の論理要素として木構造を仮定している。すなわち、ある文書は、複数の章からなっており、章は複数の節から構成される。そして節は複数の段落によって構成され、段落は複数の単語からできており、単語は文字データによって構成されている。

このようにSGMLによって記述されるのは、純粋に文章の論理構造だけであり、その文書が実際に表示される場合に、どのような体裁になるかという点は記述されない。

II. SGMLの実践

i) 実践の概要

SGMLは上述のような基礎概念の上に成り立っているのであるが、それでは実際にはどのように用いられるのか。まず、文書を用意しSGMLによってマークアップする。SGML化された文書自体は文書実現値と呼ばれる。SGML化に際して、前述した木構造におけるそれぞれの要素はDTDによって、各文書ごとにユーザによって定義される。そのようにして定義されたSGML文書とDTDの整合性をチェックするのがSGMLパーサと呼ばれるプログラムである。

それ故SGML化された文書は次の3つの要素から成り立っている

1) SGML宣言：DTD及びテクスト文書で用いられる文字と使用するSGMLの機能を定義する
2) DTD：文書構造とマークアップのための規則を定義する
3) 文書実現値：マークアップされた文書。DTDへの参照と文書テキストからなる。

これら3つのSGMLで文書を記述するのに必要な要素は、それぞれがテキスト形式のファイルによって記述されており、そのファイル自体は基本的にはプラットフォームに依存することはな

い。すなわち、WindowsからMacintoshにこれらのファイルをコピーしたとしても、ファイルの中身を確認することは常に可能である。

ii) フランス語文書の記述

では、実際にフランス語の文書をどのように記述するのか、SGML宣言、DTDそして文書実現値のそれぞれについて、次のような電子メールの場合を例にとって見てみる。

De: Alain
À: Paule
Sujet: Invitation

Veuillez noter que ... le 27 septembre 1996.

SGML宣言は、次のように記述することができる。

```
<!SGML                  "ISO 8879:1986"
CHARSET
BASESET
"ISO 646-1983//CHARSET International Reference Version (IRV) //ESC 2/5 4/0"
    DESCSET         0               9               UNUSED
                    9               2               9
                    11              2               UNUSED
                    13              1               13
                    14              18              UNUSED
                    32              95              32
                    127             1               UNUSED
    CAPACITY        PUBLIC          "ISO 8879:1986//CAPACITY Reference//EN"
    SCOPE           DOCUMENT
SYNTAX              SHUNCHAR    CONTROLS    0 1 2 3 4 5 6 7 8 9
                    10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
                    27 28 29 30 31 127 255
    BASESET
"ISO 646-1983//CHARSET International Reference Version (IRV) //ESC 2/5 4/0"
```

```
    DESCSET         0               128             0
    FUNCTION        RE              13
                    RS              10
                    SPACE           32
                    TAB             SETPCHAR        9
    NAMING          LCNMSTRT        ""              UCNMSTRT        ""
                    LCNMCHAR        "-."            UCNMCHAR        "-."
                    NAMECASE GENERAL YES            ENTITY NO
                    DELIM GENERAL SGMLREF
                    SHORTREF SGMLREF
    NAMES           SGMLREF
    QUANTITY        SGMLREF
FEATURES
    MINIMIZE        DATATAG NO      OMITTAG YES     RANK NO
    SHORTTAG YES
    LINK            SIMPLE NO       IMPLICIT NO     EXPLICIT NO
    OTHER           CONCUR NO       SUBDOC NO       FORMAL NO
APPINFO                     NONE
>
```

以上がSGML宣言の例である。SGML宣言は、文字コードについての定義を付加することからも分かるように、使用するプラットフォームを考慮して記述しなければならない。言い換えれば、あるシステムでSGMLによって文書をやりとりする場合、SGML宣言は同じものを使用することが望まれる。

次にDTDであるが、DTDは文書構造を定義づけるものであり、文書の種類によって異なってくることは言うまでもない。ここで取り上げた例は、電子メールであるから、文書の要素としては、「発信者 [auteur]」、「日付 [date]」、「サブジェクト [sujet]」、「受信者 [dest]」、「同報 [cc]」そして「本文 [corps]」があることになる。これらの要素はDTDでは次のように定義することができる。

```
<!DOCTYPE mail [
<!ELEMENT mail - -
      ((auteur & (date?) & sujet & dest & (cc?)), corps)>
<!ELEMENT (dest | cc) - - (nom+)>
```

```
<!ELEMENT corps - - (par*)>
<!ELEMENT (auteur ¦ date ¦ sujet ¦ nom ¦ par) - - (#PCDATA)>
]>
```

そして、SGML宣言文とDTDによって以上のように文書を定義づけた場合、文書実現値は以下のようにマークアップされるだろう。

```
<mail>
<auteur> Alain </auteur>
<dest>
<nom> Paule </nom>
</dest>
<sujet> Invitation </sujet>
<corps>
<par> Veuillez noter que ... le 27 octobre 1999. </par>
</corps>
</mail>
```

マークアップされた文書を見てみると、最初の電子メールでは本文とサブジェクトの間にあった2行の改行に関しては全く考慮されず「<par>...<\par>」とだけなっていることが分かる。このことからも、SGMLが文書の論理構造のみを記述していることが分かる。また、元の文書では«De :» や«À :» といった表示により、「発信者」と「受信者」を示していたが、SGML化された文書では、それらの要素はタグによって置き換えられていることがわかる。

III. SGMLのメリットとデメリット

i) 文字コードの問題

コンピュータにより文書を作成し、保存・共有する際の問題点の一つに文字コードがある。特に、フランス語と日本語を混在させた場合には、フランス語の一部が文字化けするという問題が起こる。このもっともわかりやすい例は、日仏混合文をWebブラウザによって表示させた際に顕著である。

この現象は、日本語を表示する際の文字コードとフランス語を表示する際の文字コードとの間に、相互に衝突する文字コードがあるために起こる。すなわち、同じコードにより、二つの別の文字が指示されることが原因である。

SGMLでは、SGML宣言文の中で"BASESET"において"CHARSET"を定義することができ、

文書において基準となる文字コードを設定することができる。また、ASCIIコードだけを用いて次のようにアクサン記号のついた文字を代替させることも可能である。

«é» = > «é» / «á» = > «à» / «ï» = > «ï»
«û» = > «û» / «Ç» = > «Ç»

こうした表記法は、実はそのままHTMLにもそのまま受け継がれており、特殊記号などについてもHTMLでの表記法をほとんどそのまま用いることができる。

ii) SGMLとTeX

SGMLが文書の論理構造を記述する言語だったのに対して、TeXは論理構造と物理構造（表示方法）を一度に記述するための組版システムである。その機能は市販のワードプロセッシングソフトウェア（以下ワープロソフトと略す）とあまり変わらない。但し、TeXがワープロソフトと異なるのは、特定のコンピュータのアーキテクチャやオペレーションシステムに依存することなく、文書を整形することができるという点である。

それ故SGMLとTeXとは全く別の目的を持ったソフトウェアである。しかし、TeXの、文章（SGMLの文章実現値）にタグに似たものを書き加えることにより組版を実現するという手法は、SGMLと類似していると言えなくもない。SGMLにおけるタグに相当するものは、TeXではコマンドといわれる。TeXのこれらのコマンドは、SGMLがタグをDTDの定義によって拡張できのと同様に、マクロを作ることによってユーザが拡張することが可能である。
フランス語をTeXで整形する場合に関してのみいえば、TeXには"french"という名前の拡張パッケージが存在しており、このパッケージを用いることで、フランス語特有の組版規則を用いることができる。

但し、フランス語パッケージで用いられている文字が別の言語のパッケージではコントロールシーケンスとして用いられている場合がある。すなわち、PC上でフランス語のファイルをやりとりする場合の文字コードの問題は、TeXにおいても存在しているのであり、結局根本的な解決策とはならない。

iii) SGMLのメリットとデメリット

SGMLは、文書をプラットフォームに依存することなく保存・共有するための言語であった。そして、SGMLで文書を保存するためには、SGML宣言とDTDそしてマークアップされた文書（文書実現値）の3つが必要であった。

研究者が文書をSGMLで共有しようとする場合には、DTDによって定義されたタグを利用して、ユーザ自身が文書をマークアップする必要がある。そしてそのためには、どのようなタグが

使えるかについて文書を共有する研究者間での合意に基づいてDTDの定義を作成する必要がある。さらに、そうして定義したDTDを研究者に対して公開することも必要である。

このように、タグによって文書をマークアップするだけで、プラットフォームに依存しない形式で文書を保存し、共有するための準備が整うという点では、SGMLは非常に有効である。

しかし、フランス語で記述された様々なジャンルの文書の論理構造を規定し、それをタグとして定義するという作業は、個人の研究者が行う作業の範囲を超えているように思える。また、実際にフランス語テクスト用のDTDを策定し、それが実際に運用されるまでにはある程度の年月が必要になると思われる。この点は、個人レベルで文書のSGML化を上での障害となるであろう。

また、SGML宣言はプラットフォームごとに定義しておく必要があったが、異なるプラットフォーム間で文書を共有する場合には、各プラットフォーム上で用いられている文字コードなどを考慮に入れた上で定義する必要がある。こうしたプラットフォームに依存する問題を解決する方法の一つとして、SGMLが開発されたことを考えると、SGMLを使用する上でのデメリットであると考えられる。確かに、異なるプラットフォーム間での文書の取扱方の違いを解消する何らかのシステムは必要であり、それがSGML宣言として実現されているわけであるが、大規模なプロジェクトによって文書を保存・共有するのではなく、個人レベルでの文書保存を考える場合には、SGML宣言を記述することは、研究者の負担となると言わざるを得ない。

おわりに

以上のように、SGMLにしろTeXにしろ、ファイルをやりとりする上での文字コードの問題は避けて通れない。あるファイルを複数のプラットフォームで共有するためには、常にこの文字コードの問題がつきまとうことになる。それ故、結論としては、文字コードの概念を超越するためには、文字コードの原点に戻って、ASCIIコードにより文書を保存するという方法が、もっとも汎用性の高い文書の保存形態ということになる。

但し、この場合にも、イタリックやボールドといった、いわゆる文字のスタイル情報は、表記することができない。さらに、同じASCIIコードを使ったとしても、プラットフォームによって、改行コードなどの制御文字が異なるため、文書の保存・共有の際には注意が必要である。また、フランス語などで用いられているアクセントつきの文字は、ASCIIコードで用いられている文字セットだけで表記しなければならない。たとえば"é"を表記する代替として"e+"を用いるといった具合である。

どのようなデジタル・テクストを保存し共有する場合でも、簡便で自明な保存法というものは存在しない。文字コードの問題をどのように回避するにせよ、そこには文書を共有するユーザ間の同意、すなわち標準化の作業が必要になってくるのである。そして、文書の保存に関する標準化の手法が定義づけこそが、デジタル・テクストの共有と活用を促す上での大きな鍵であること

は間違いないだろう。そして、その際には、SGMLやTeXといった特別な知識を必要とするような方法によって標準化するのか、それともASCIIコードによる場合のように簡便な方法を選ぶのか、選択をする必要があるだろう。

参考文献：

Eric van Herwijnen,『実践SGML』, SGML懇談会実用化WG監訳, 日本規格協会, 1992年.
吉岡　誠　編著,『SGMLのススメ』, オーム社, 1993年.
吉岡　誠　編著,『SGMLを使いこなす』, オーム社, 1996年.
Michel Goossens, Frank Mittelbach, Alexander Samarin共著,『The LATeXコンパニオン』, アスキー書籍編集部監訳, アスキー出版局, 1998年.
長尾　真, 黒橋　禎夫, 佐藤　理史, 池原　悟, 中野　洋　共著,『言語情報処理』, 岩波講座　言語の科学9, 岩波書店, 1998年.
田中穂積　監修,『自然言語処理—基礎と応用—』, 社団法人電子情報通信学会, 1999年