

Finnegans Wake のテキスト生成過程研究のための TEI 準拠電子アーカイブの試作

扶瀬 幹生 北本 朝展
聖心女子大学 文学部 国立情報学研究所

本研究の目的は James Joyce, *Finnegans Wake* (1939) のテキスト生成過程の緻密な分析に用途を特化した実用的な電子アーカイブ作成にむけて、その基礎を固めることにある。文書記述のために TEI の規格を採用したが、折しも今年に入ってマニスクリプト関連の文書記述ルールの充実を図った TEI の新しいガイドラインに準拠して、実際に本研究の目的にかなった電子アーカイブの制作がどのように、どの程度可能なかを検証することが試作のポイントとなる。

Groundwork for a TEI-Based Digital Genetic Archive of James Joyce's *Finnegans Wake*

Fuse Mikio
Faculty of Liberal Arts
University of the Sacred Heart, Tokyo

Kitamoto Asanobu
National Institute of Informatics

The project aims at consolidating the basis for a digital archive of James Joyce's *Finnegans Wake* specifically dedicated to close scholarly analysis of its textual genetics. We have followed the latest TEI guidelines to see how far their updated rules for encoding manuscript sources work for one of the most complicated examples of literary creation. So far the TEI rules have been found compatible with out project, with a couple of useful analysis tools created out of the documents as HTML views.

1. 目的

過剰に多義的な文学テキストとして知られる *Finnegans Wake* (以下 *FW*) の電子アーカイブ化を試みることの意義は大きい。特にこの多義的なテキストを作り上げる過程で使用された 50 冊以上の「ノートブック」やその典拠となる「ソース」、最終的に書物として印字されるに至るまでの数多くの「草稿」類まで視野にいれて *FW* のテキスト生成過程を記述・分析しようとする場合、電子アーカイブの構築は「ジョイスの創作原理そのものに要請される形で」必須の作業であると言ってよい[1]。

本年作者の没後 70 年を迎えて、長年 *FW* のテキスト生成過程研究を進めてきた Danis Rose が電子版 *FW* 刊行の意思を表明したほか、昨年公開された *Samuel Beckett Digital Manuscript Project* の主幹の一人 Dirk van Hulle が所属する Antwerp James Joyce Centre にも *FW* の電子アーカイブを制作する意思があることが確認できているが、いずれについても現時点ではいつどのような形で世に現れるのかは確定していない。

このような状況のもと本研究の目指す電子アーカイブは、テキスト生成過程研究のプロセス自体を再検証可能な形で可視化することによって、既存データの精練や新データの発見を促す推進力を持つものとして、将来的に上記その他の電子アーカイブと有意義な共存関係を保ちながら独自の立ち位置を確保できると考えている。

2. *Finnegans Wake* のテキスト生成過程

2.1 各種文書の種類と関係

まず *FW* のテキスト生成過程にかかわる文書の種類とその関係を説明しておく。

ジョイスは *FW* の執筆の全過程にわたって B シリーズと呼ばれる多くの自筆ノートブックを作成して創作に活用した(図 1 の NB-B)。そこには当時手に入る書物や新聞雑誌の記事等のソース(SO)から種々雑多なテキストの断片が書き写されている。ジョイスはこのノートブックを見ながらその時執筆中の草稿(MS)の文章に使用できる項目(ユニット)をみつけて本文に組み入れ、使用済ユニットを色鉛筆で打ち消すことによって重複使用を避けるようにした。これが創作の一つの主要経路であるが、ジョイスはさらに NB-B で使用しなかったユニットをリサイクルするために、パリで秘書として雇った

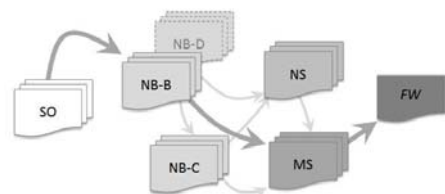


図 1 *FW* のテキスト生成過程

Raphael 夫人に NB-B のページで色鉛筆で打ち消していないユニットを新しい C シリーズのノートブック(NB-C)に転写してもらい、そこからさらに MS に組み込める語句を見つけていった。また 1930 年代に入ってからノートシート(NS)といわれる NB と MS の中間的性格を持つ文書に NB-B や NB-C で未使用のユニットを利用して短い断片的文章のスケッチを作り、それを MS に移して使用した。

以上が FW の執筆過程の概要である。テキスト生成過程研究の仕事は NB-B の SO の発見、NB-B、NS 等の手稿の判読、および MS での使用箇所の特定を行いながら、これら諸文書間のリンクを明示していくことにある。(注: 自筆ノートブックには A と分類される、NB-B とは性格を異にする一冊もあるが、ここでは触れない。)

2.2 ノートブック VI.C.2 の重要性

実際のアーカイヴ構築を始めるにあたって、NB-C のうち特に VI.C.2(以下 C02)に照準をあわせ、そこから関連文書のデータと分析機能を拡充していくことにした。

その理由は、一つには自筆の NB-B よりも NB-C のほうがより複雑な生成過程の一端を担っており、それだけ文書・メタ文書の記述に工夫を要するからである。NB-B から直接 MS に行くのではなく途中で NB-C を経由することは、実は単に一段階経路が増えるという問題にとどまらない。なぜなら後でも触れるように Raphael 夫人はジョイスから渡された NB-B を転写する際にしばしば写し間違いを犯すので、この部分の生成過程が特異な形で複雑になっているからである。また C02 には直接 MS に使用されるだけでなく、一旦 NS に移されてから MS に使用されたユニットも実際多く含まれている。C02 をアーカイヴに組み込もうとすると、これらの複雑な生成過程の記述を行わなければならないのである。

さらに C02 の役割を一層複雑にするのは、このノートブックが散逸してしまった D シリーズと呼ばれる自筆ノートブック(以下 NB-D)の一冊である VI.D.1(以下 D01)のうち色鉛筆で打ち消されなかったユニットを書き写している貴重な文書であるという点である。

図 2 に示したように、失われた NB-D の痕跡を残すのは、使用されたユニットについては NS/MS、使用されなかったユニットについては NB-C ということになるが、NB-C に(しばしば不正確に)転写されたユニットという形で痕跡を残している NB-D ユニットの SO の特定、および NS/MS で使用された語句(NB-D 由来のもの含む)のもとになっている諸 NB についての情報分析を緻密に行うことによって、失われた NB-D を復元することがある程度可能になる。

C02 をアーカイヴに組み込むということには、失われたノートブック D01 を仮想空間において復元するという課題、またそのために必要な徹底した SO 特定の方法論や NB ユニットの MS での使用箇所に関する総合的なデータ分析機能を有するアーカイヴ構築が要請されるわけである。

以上の理由から、C02 をアーカイヴに組み込む方法を模索することは、FW のテキスト生成過程のもっとも複雑な(しかし決して例外ではない)パスを通すということを意味し、結果的にそれほど複雑でない生成過程の記述や分析のしかたも自ずと拓かれていくことになる。

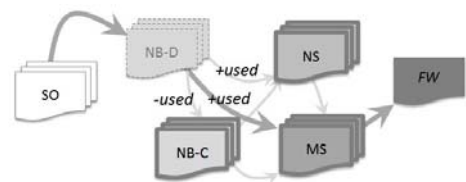


図2 NB-D の痕跡を残す NB-C と NS/MS

3. 文書作成の方法

本研究に着手するまで、SO 以外の文書は Excel (3.4 参照)、SO 文書は pdf で蓄積・保存し、これにメタ文書検索をかけてデータ分析を行う方法で研究を進めてきた。本研究では TEI に準拠した形で各種の文書を電子アーカイヴに統合することで、FW のテキスト生成過程を文字通り一体として記述し、個々の文書や文書間の関連をより柔軟かつ緻密に分析・明示することを可能とした。

3.1 TEI 新ガイドラインへの準拠

そのためには、まず諸文書のデータをアーカイヴに収納するための適当なフォーマットを選定しなければならない。文学テキストの電子化については TEI(Text Encoding Initiative)の提唱する XML 文書としてのタグ付けのルールが汎用性の高い基準として存在するが、特に今年に入ってそれまで Manuscript SIG (Special Interest Group)が提唱・開発を進めていた Genetic Editions のためのタグ付けモデルが公式のガイドライン(P5 2.1.0)に盛り込まれた形となった[2]。本研究では TEI が今年 6 月に新しく P5 2.1.0 に準拠したものとして公表したスキーマを採用して諸文書を TEI 文書として電子化していく方針である。

TEI はそのガイドラインで文書記述のための膨大な種類のタグとその使用のルールを規定しているが、実際の適用に際しては文書データの使途目的をはっきりさせて、必要な情報に対して効率よくタグ付けできるように適用方法を慎重に吟味する必要がある。

TEI では文書本体とヘッダー部分でそれぞれ文書の本文内容と文書に関連する諸情報を記述するので、以下それぞれにおける記述方法を説明する。

まず本文について。FW のテキスト生成過程に関係する文書は次の二種類に大別される:

1. 印刷物として刊行されたもの: FW および SO 文書
2. 創作過程で使われた手稿類: NB、NS、および MS 文書

TEI はこの二種類の文書に対して本文部分に異なる階層構造を用意する。ページレベルまでの文書構造を XPath の記法を用いて示すとそれぞれ次のようになる:

1. /TEI/text/body/pb
2. /TEI/sourceDoc/surfaceGrp/surface

3.2 意味的・論理的構成要素の前景化

問題はここから先、ページ内にどのような構造を見るかであり、それによってデータ運用上の利便が大きく左右されるので、TEI の用意する選択肢を慎重に吟味しなければならない。本研究では結論として行(l/line)、センテンス(s)等の形式的・物理的なページ内テキストの構造を背景化ないし無視することによって、生成過程分析の機軸となる NB、NS のユニット(seg[@type= 'unit'])やそれに対する色鉛筆の打ち消し(del[@type= 'used']), MS の加筆箇所(add)、および SO、FW の関連箇所(seg)を意味的・論理的構成要素としてつぎのように前景化することにした:

1. SO および FW 文書
/TEI/text/body/pb|lb|seg
- 2a. NB および NS 文書
/TEI/sourceDoc/surfaceGrp/surface/zone/pb|lb|del[@type= 'used']/seg[@type= 'unit']
- 2b. MS 文書
/TEI/sourceDoc/surfaceGrp/surface/zone/pb|lb|add

ページ内の物理的なテキスト構成要素の背景化は TEI が「空要素による構造枠の記述」方法の一つとして提示する、TEI に準拠したオプションである[3]。たとえば MS 文書で additions は図 3 のように記述されることになる。(例で使用している記号*や[]については後述する。)

もしこの方法を取らなかった場合、行に収まる addition は<add>、行をまたぐものは<addSpan>と二種類の要素を使い分けなければならない。膨大な数の additions を扱う本研究の場合、恣意的かつ無意味に煩瑣なタグ付け作業を強いられることになってしまう。

```
<lb/>above you peregrines.  
*<add>ay [voch*] to  
rumanescu[.]</add>  
<add>See the leabhour <lb/>of my  
generations!</add>
```

図 3 行の空要素化による add 要素の前景化

3.3 文書間の関係の記述

次にヘッダーについて。TEI がヘッダー部分に用意する記述項目のうち、特に本研究の目的上重視されるのは次の二つである:

- A. /TEI/teiHeader/sourceDesc/bibl/date
- B. /TEI/teiHeader/profileDesc/creation/listChange/change/graph

前者は個々の文書の作成年代に関する情報記述の場であり、後者は文書の生成過程上の変化に関する情報をグラフの形で記述する場所である[4]。

テキスト生成過程の分析上諸文書間の時間的前後関係をつけていこうとする場合に、A に入れてあるそれぞれの文書の作成年代が決定的に重要なカギになることは言うまでもない。

一方 SO から FW に至るまでの関係文書間の関連情報については、TEI のガイドラインが例に挙げているような一文書から一文書の関係ではなく、一文書内の複数の単位から他の複数文書内の複数の単位への多対多の関係になるのだが、ルール通りヘッダーの graph を活用して図 4 のように記述することにした。

このヘッダー情報は NB 文書にのみ入れる。図 4 の例は D01 の 127.11x というユニット(これは失われた NB の復元

```
<graph>  
  <node xml:id="D01-127_11x-S0" facs="../XMLDocument/S0/S0_Low-WP-1924.xml#Low-WP-1924_036-14_seg1"/>  
  <node xml:id="D01-127_11x-NB-BD" facs="#Lb34b02d01b06c0203543_D01-127_11x"/>  
  <node xml:id="D01-127_11x-MS" facs="../XMLDocument/MS/MS_3-3A_8.xml#MS47484a-253_FW482-34_add1"/>  
  <node xml:id="D01-127_11x-FW" facs="../XMLDocument/FW/FW.xml#MS47484a-253_FW482-34_add1"/>  
  <arc from="#D01-127_11x-S0" to="#D01-127_11x-NB-BD"/>  
  <arc from="#D01-127_11x-NB-BD" to="#D01-127_11x-MS"/>  
  <arc from="#D01-127_11x-MS" to="#D01-127_11x-FW"/>  
</graph>
```

図 4 graph による文書間情報の記述

ユニットの例である)について、その SO から FW に至るまでの関連文書の場所と該当箇所の ID を各 node に記した後、node 間の関係を @from, @to で示しているが、D01 文書のヘッダーには D01 の個々のユニットについての文書間関連情報が同様の graph としてユニットの数だけ並ぶわけである。

このように特に NB 文書のヘッダーを文書間情報のベースにすることは、NB が FW 創作の事実上最初の基盤であることから理にかなっており、実際これまでのテキスト生成過程研究でも NB をベースにして文書間関連情報の記述が行われてきたので、その伝統を継承していることになる。

3.4 元データ管理場所としての Excel の利便性

FW の電子アーカイブはもともと Excel で管理していた。その理由は、Excel が膨大な種類と量のデータを人間の手作業で扱うことに適しているからである。たとえば NB や MS のデータはセルや文字の書式を工夫して視覚的に Word のように作ったり読んだりできる形で蓄積・保存できる。MS の場合は一つのワークシートの縦軸に FW のページ・行、横軸に MS のレベルをとることによって、膨大な数の MS 文書を一つのワークシート上に収めることができ、新規入力の際には前のレベルの MS の内容をコピー・ペーストしたものをもとに次のレベルの MS 文書を効率的に作成していけるという利点がある。

また今回の研究で新たに必要が生じた文書間の関連情報のデータ化のための校合(collation)にも、Excel が有効に活用できることが分かった。具体的には「参照元の NB-B と転写先の NB-C のユニット単位の対応の関係付け」、および「MS の行と FW のページ・行の対応の関係付け」である。この校合作業は機械の自動判別に頼るわけにはいかず、人間がどこかでしなければならないが、図 5 のように Excel のワークシートを使うことが機能的にも視覚的にも人間による手作業に最も適している。

B34	172	7	C	SA hair grows	C02	001	1	a	0	SA has grown
B34	172	8	0	short or long	C02	001	2	b	0	short a lass
B34	172	9	D	river	C02	001	3	c	0	rive

図 5 Excel を使った B34 と C02 の校合

本研究のように既にある程度の文書データが Excel に用意されている場合、今後も元データの管理場所として Excel を使い続けるべきか、それとも TEI 文書に変換したデータを新たな元データとして管理し、XML エディタで新規入力や修正を行っていくべきかという決断を迫られることになるが、上述の諸理由によって Excel を元データ管理場所として、また一部の例外を除いて基本的に新規入力や修正などのデータ更新の場として継続使用することにした。

3.5 Perl を用いた Excel から TEI への自動変換

このように本研究では元データを Excel で管理することにしたが、ここで表形式のデータと TEI とをどのように結び付けるか問題となる。そこでプログラミング言語 Perl を用いて Excel 文書から TEI 文書を自動生成するためのプログラムを構築し、Excel 上での変更が TEI に反映されるようにした。必要に応じて校合を済ませた Excel のデータは Perl のモジュール Spreadsheet::ParseExcel と XML::LibXML を利用して必要なタグ付けがなされた形で TEI 文書に変換される。MS や NB での revision について従来使用してきた < > (add) や < < (delete) 等の非公式の記号も、Perl で変換用の適当なサブルーティーンを作って一挙に TEI のタグに変換できる。たとえば前頁で MS から挙げた例について、Excel に入れているデータは図 6 のように自動変換される。



484 29	47484a-254v-255	above you peregrines. *ay [voch*] to rumanescuf.] <See the leabhour
484 30	47484a-254v-255	of my generations!>

```

<lb xml:id="MS47484a-254v-255_FW484-29"/>above you peregrines. *<add xml:id="MS47484a-254v-255_FW484-28_add1">ay [voch*] to rumanescuf[.]</add><add xml:id="MS47484a-254v-255_FW484-28_add2">See the leabhour
<lb xml:id="MS47484a-254v-255_FW484-30"/>of my generations!</add>

```

図 6 Excel から TEI への自動変換例 (MS 文書)

上の変換例についてはいくつか補足が必要である。

まず、lb に FW のページ・行との校合結果を踏まえた有意の ID が割付けられている点について。前述のように文書間の関係情報は本来ヘッダーで graph として記述するのが TEI のルールである。NB の場合と同様 MS についてもこの方式に従うことは技術上可能であるが、今回の試作では ID の中に文書間関連情報を埋め込むことで、専用のプログラムを使えば文書間の関係を ID だけから復元可能とする方法を用いた。これは、ヘッダーに膨大な校合情報を入れるのを嫌ったこと、そしてより本質的には次に述べる add 要素への ID 割付けの便を理由に、本体部分で文書間の関係情報を補完するためである。

次に add 要素の ID として、その行の ID から派生した ID が割付けられている点について。本アーカイブの機軸となる要素として前景化した MS の add 要素や NB 等の seg 要素には当然 ID を割振って情報を駆使することになるが、すべて上図のようにその行の ID から派生した有意の ID を割付けることにした。MS の add の ID 割付けは Perl を用いて自動で行う。

NB 文書についても参照元と転写先の NB の関係についての情報を本体部分のユニットの seg 要素の ID に組み入れた。

```

<seg type="unit" xml:id="Lb34b02d01b06c0200000_C02-001_a">
  $A has grown </seg>

```

図 7 NB ユニット ID への校合情報の組み入れ

図7の例ではユニット名(C02.001a)の前に、C02とその参照元(B34,B02,D01,B06)のNB名一覧、およびこれらのNBのユニットごとの校合に使ったExcelのワークシートの行番号(00000)が入っている。

ただNBの場合は先述のようにすでにヘッダー部分にも個々のユニットについて参照元/転写先のNBユニットの情報は入っているので情報的には冗長になる。ヘッダー部分のみに情報を記すべきかどうかは、今後の試作で運用の利便を検討したうえで判断したい。

最後に、テキストに*や[]といった記号が用いられている点について。これはExcelでデータを蓄積する際に「不確定」を表すために非公式に使ってきた記号で、これらは当然TEIに準拠した形で記述しなおさなければならない。将来的にはアーカイヴの大筋ができあがり、特に後述の画像の装備が充実した段階で、この部分もTEIでマークアップする計画である。

Excelからの変換で出力された諸文書は、XMLエディタとして利用しているoXygenによってXML/TEI文書として不整合がないか最終的なエラーチェックをする。oXygenはこの他、手動でSOおよびFW文書の該当箇所のタグづけを行う場合にも活用している。

以上説明した文書作成の流れをまとめると表1のようになる。

表1 文書作成の流れ

Excel ↓	- データの新規入力と更新 - MSとFWの行単位の校合 - NB-BDとNB-Cのユニット単位の校合
Perl ↓	- Excelデータの読み取り (Spreadsheet::ParseExcel) - TEI文書としてのデータ加工とコード出力(XML::LibXML) - Excelで使用している revision 記号の変換
oXygen	- MS文書の add 要素への ID 自動割付け - TEI文書としての整合性チェック - 文書関連情報にもとづく SO, FW 文書の該当セグメントの手動タグづけ

4. TEIからのHTMLビューの生成

以上のようにして用意したTEI文書は研究上の機軸となる要素がXMLとしてきちんと構造化されているので、テキスト生成過程分析の様々な具体的局面を透明化し活性化するための、目的に応じたビューを生成できるようになる。具体的には、TEI文書の中から必要な要素や属性を選び出して文書構造を変換するためのスタイルシートであるXSL文書を用いて目的にかなったHTML文書を生成するとともに、HTML文書に適切なクラスやIDを設定することで、特定の要素や属性の表示方法（例えば打消しに使われた色鉛筆の色の表示など）もCSS(Cascading Style Sheet)によって制御できる。現時点で4.1~4.3に示す三つのビューを実現している。なお、XSLTの実行にoXygen搭載のプロセッサを使うと、XSL以外のプログラミングが不要になるというメリットがあるものの、本研究ではPerlのモジュールXML::LibXSLTをXML::LibXMLと併用することで、如意の文字列操作や加工を行うとともに、4.1の場合のように新しい複合文書を作ってXSLTにかけるという処理も実現した。

4.1 Parallel NB Units

このページはNBを中心にFWテキスト生成過程の実際が巨視的にも微視的にも容易に確認できるようにしたものである。先述のNBユニットのIDに埋め込んだ補助的なNB文書間の校合情報に基づいてPerlで新たな複合NB文書を作り、参照元と転写先の関係にある個々のNB-BとNB-Cのユニットが交互に並ぶようにXSLTで呼び出している。またユニットのコラムの左側にはそれぞれのユニットの出自であるSO文書、右側にはその使用先であるNS/MS以降についての情報が呼び出せるようになっている(図8)。

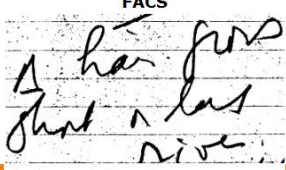
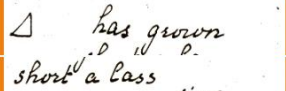
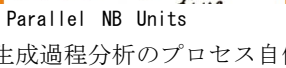
SOURCE	UNIT	COLOUR	ENTRY	FACS	NS	MS	FW
B34-172_	C		\$A hair grows short or long				
	C02-001_a		\$A has grown				
	C02-001_b		short a lass				

図8 Parallel NB Units

このページについては、アーカイヴの目指す「生成過程分析のプロセス自体を可視化することによる、自己精錬の促進」という観点から、以下のような用途が特筆できる。

4.1.1 NB-BとNB-Cのユニット番号の割付けの整合性チェック

本来参照元のNB-Bと転写先のNB-Cのユニットは一対一で対応していなければいけない。言い換えれば、このページで背景が白色のNB-Bのユニットと背景がオレンジ色のNB-Cのユニットが一行ずつ交互に出てこなければならない。ところが実際にこのページを見てみると、図8の例も含め、そうならないところが随所にある。

その原因は一つには、必ずしもすべての NB のユニット割付けについての公式な合意がまだ形成されていないことにある。一応現時点で公式のユニット割付けがなされているものについては a, b, c と小文字のアルファベットを使い、そうでないものについては扶瀬が A, B, C と大文字で割付けたものを使っている。もしユニットの割付けが一对一で対応していない場合、小文字の割付けに合わせて大文字の割付けを修正するというのが全うなようだが、実際には公式の割付けも絶対正しいとは限らず、最終的な判定のカギになるのは親となる SO の該当パッセージの文章内容である。SO が判明していないユニットについては、ユニット番号の割付けの不整合は今後の要検討箇所として残しておくのが妥当である。

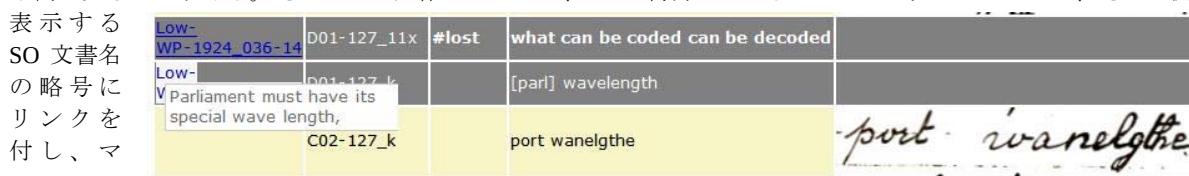
4.1.2 NB-B と NB-C のテキスト・画像両面での比較

NB 文書の各ユニットは surface 上の小区分である zone として扱うが、zone の属性でファクシミリ画像の該当箇所の座標を指定することができるので、試作ではこれを参照してユニットごとに該当箇所を書き起こしテキストの右側に CSS Sprites と呼ばれる技法を用いてページ画像の該当部分を呼び出せるように工夫した。

先述のように NB-C の読みが NB-B の読みと食い違っているという現象は、転写した Raphael 夫人の写し間違えと断定できる場合はそれとして客観的にデータとして残しておくかなければならないが、実際にはもとの NB-B のジョイスの筆跡が相当の走り書きであることから NB-B の判読自体もつねに正しいとは限らない。この場合も、最終的な判定のカギになるのは SO の該当パッセージの内容であり、SO が判明していない限り NB-B と NB-C の読みの食い違い箇所は今後の要検討箇所として残しておくのが妥当である。

4.1.3 さらなる SO 発見の糸口

上記 1、2 で分かるように、NB の判読には SO の特定が決定的なカギを握る。SO 文書自体の発見には今日の書籍・新聞の検索サービス(Google Book Search, Internet Archives, Irish Newspaper Archives など)が威力を発揮するが、一応 SO が判明している場合でも、先行研究者が見落とした同一の SO 文書中のさらなる SO パッセージが見つかる例が少なからずある。そこでこの試作ページでは、SO が判明している NB のユニットについては、その左側に



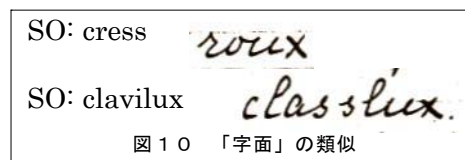
表示する SO 文書名の略号にリンクを付し、マウスオーバーで該当パッセージを tooltip として呼び出して確認できる一方(図 9)、クリックすると SO 文書自体が呼び出せるように工夫した。

図 9 tooltip を用いた該当 SO パッセージの呼び出し

クリックで開かれる SO 文書のページでは SO パッセージとして特定された部分がすべて太字で強調され、クリックした該当ユニットの出自となる SO パッセージにターゲットとして焦点が当たるようになっている。もし、その前(ないし後)のユニットで SO が判明していないものがあり、かつそれよりさらに前(ないし後)のユニットの SO パッセージが同一文書内に特定されている場合は、特定されて太字になっている二つの SO パッセージの間にある本文を精査して、SO の見つかっていないユニットの語句に合致するものがないかチェックすることができる。これは膨大な量の SO 文書を難なく扱える電子アーカイヴならではの研究支援ツールになるはずである。

4.1.4 ノートブック D01 の復元とその根拠の透明化

4.1.3 の手法は NB 一般の SO 特定方法として有効だが、特に散逸した D01 のさらなる SO パッセージの発見に際しては、上述の画像と併用することでユニット復元に向けて利用価値がさらに高くなる。たとえば Raphael 夫人が“roux” (C02.150g) と転写したり“classlux.” (C02.150k)と転写したりしているユニットについて、その間に連続するユニット、“artesian wells” (C02.150h)、“to harbour” (C02.150i)、“Thomas Witford.” (C02.150j)についてまず新聞アーカイヴの検索によって SO 文書としての新聞記事(Irish Independent, 19 May 1925, p. 6, col. 4)が特定できていたものだが、4.1.3 の最後に書いた方法では肉眼でも文字検索でも“roux”や“classlux”に合致する語句はどこにもみつからない。しかし、テキストの中に(特にそれが活字ではなくジョイスの走り書きで書かれた場合を想像して)「字面」上これらの単語と形が似ている単語を探せば、同じ記事の中に“cress”や“clavilux”を見つけることは肉眼でも(あるいは肉眼でこそ)可能であり、特に前者のような微妙な事例の場合は、活字でなく画像の参照が「字面」の類似性を確認するために必要不可欠である(図 10)。この例は本研究に着手する以前に発見した事例ではあるが、失われた D01 のユニットとして(“roux”や“classlux”でなく)“cress”や“clavilux”を復元する場合に、その根拠がいつでも再確認できるようになっていることの意義は大きい。



ちなみに手稿類の書き起こし text に対して、それを物理的な document として記述するための facsimile の重要性は、TEI の Manuscript SIG でも指摘されていたし[5]、Samuel Beckett Digital Manuscript Project ではこの種の議論

を踏まえて linear transcriptions, digital facsimiles に加えて、さらに両者のハイブリッドとしての topographic transcriptions を提示するという周到な配慮までなされているが[6]、今回の試作で導入したような画像と書き起こしテキスト(そしてその SO 文書のテキスト)を至近距離で比較参照できるようにする工夫は、専門性の高い特定の目的にかなった、実用的な併記法と考える。

以上はユニットの左側に呼び出した SO 文書関係の情報についてだが、右側に呼び出す NS, MS, FW 文書の情報についても同様に二種類の表示が可能なリンクを用意すればよい。ただし MS 文書については、リンクをクリックしたときに SO と同様にすべての add 要素を太字にした形でページを呼び出しても、利用価値はあまりない。SO から NB への転記の場合は、一つの SO を読みながら一つの NB へ順次語句が写されていったので、双方のページ・行が並行して先に読み/書き進むようになっているが、NB から MS への転記の場合は一つの NB から複数の MS へ、また一つの MS へは複数の NB から、転記が行われるからである。次項に示すビュー NB Parents for MS Additions は、Parallel NB Units の右側の MS 文書情報について、リンクをクリックしたとき開かれるべきものとして考案したものである。

4.2 NB Parents for MS Additions

このページは FW の Book III, Subsection 3A の level 8 の MS について、そのすべての additions のうち親となる NB のユニットがどの程度判明しているかが一目で確認できるようにしたものである。MS の最初の試作として数多くある MS 文書の中から特にこの文書を選んだのは、散逸したノートブック D01 のユニットのうち C02 に転記された形で残っていない(すなわち、D01 上で色鉛筆で消され、NS ないし MS に使用された)ユニットの復元として Henkes が提案したものうち最も多くがこの MS を使用箇所としていたからである [7]。

このページは XSLT でこの MS 文書のすべての add 要素の ID について、NB 文書のヘッダーの文書間情報を参照し、合致する NB ユニットの情報を呼び出している。参照用に読込む NB 文書としては、*Finnegans Wake Notebook Editions* の編者の一人 Vincent Deane から得た情報に従って、この MS を使用箇所とする NB として判明している 14 冊の NB 文書を使っている。Henkes の提案する D01 の復元ユニットのうち、MS での使用先がほかの NB の使用先としてすでに特定されているものは却下されなければならないということになる。

このページの用途は失われた NB-D の復元の蓋然性チェックだけにとどまらず、一般に NB のユニットの使用先特定のための有力なツールとしても利用価値が高い。MS の左に親 NB ユニット名を表示するセルは、そのユニットが使用済みであることを示すためにつけられた打消し線の色を表すようにしてあるので、所定のセクション・レベルの MS 執筆の際にジョイスが参照した NB ユニットの打消しに使った色鉛筆の色に一定の傾向があることが分かり(例えば図 1 1 で B17 の三つのユニットはすべて青)、同じ NB からの使用例の数と併せて、色鉛筆の色が合致しているかどうか、NB ユニットの使用箇所特定の際の重要な判断基準になるわけである。

B17-077_A	MS47484a-253_FW482-30_add1 He is cured by faith who is sick of faith
	MS47484a-253_FW482-31_add1 a writing
	MS47484a-253_FW482-32_add1 still more learned
	MS47484a-253_FW482-32_add2 the raiding
D01-127_11x	MS47484a-253_FW482-34_add1 What can't be coded can be decoded
	MS47484a-253_FW482-35_add1 when the if an ear aye sieze what no eye ere grieved for.
	MS47484a-253_FW482-35_add2 if an
B17-015_D	MS47484a-253_FW482-36_add1 the doctrine obtains
	MS47484a-253_FW483-02_add1 , * take it upon myself to let me suggest,
B17-015_H	MS47484a-253_FW483-02_add2 take it upon myself to

図 1 1 NB Parents for MS Additions

4.3 MS Line Diffs

このページは 4.2 で取り上げた FW の Book III, Subsection 3A, level 8 のテキストを一つ前の level 7 のテキストと(該当 FW テキストの)行ごとに比較して、語句の異同をチェックしやすくしたものである。MS 文書の行の ID に埋め込まれた補助的な文書間情報を参照して、同一の FW のページ・行に対応するテキストの異同を Perl で調べ、内容が違う行をハイライト表示している。

このページの用途としては、書き起こしや校合のミスの発見のほか、一度先行する MS で作者によってキャンセルされたあと(あるいはタイプストや植字工によって見落とされたあと)、後続の MS で再び addition として復活した語句を見つけることにより、通常の新規の additions と restore された additions を峻別し、後者を 4.2 のデータ分析から除外することによって、分析の精度をさらに高めることにある。

3-3A_8	3-3A_7
ly a wall went forth. Pure Ywan Yawn lay low. On the [FW474-01_MSPinceton-19]Lowly, longly a wall went forth. Pl	
[FW474-02_MS47484a-156]mead of the hillock lay, [FW474-02_MSPinceton-19]mead of the hillock lay,	
wallet to his side, an arm loose, by his staff of citron [FW474-03_MSPinceton-19]brief wallet to his side, an arm loo	
FW474-04_MS47484a-156]briar, tradition stick-pass-on.. [FW474-04_MSPinceton-19]briar.	
[FW474-05_MS47484a-156]Most [FW474-05_MSPinceton-19]Most	
'ully (but, my dear, how successfully!) to wall he did, [FW474-06_MSPinceton-19]distressfully (but, my dear, how su	
; of a Lucan tinge, quickrich, ripely rippling, unfiled, [FW474-07_MSPinceton-19]his locks of a Lucan tinge, quickrich	
shbetasselled lids on the verge of closing time, whiles [FW474-08_MSPinceton-19]those lashbetasselled lids on the vi	
of his sidewiseopen mouth the breath of him, evenso [FW474-09_MSPinceton-19]ouze of his sidewiseopen mouth th	
ing as the princeldest treble treade or lichee chewchow [FW474-10_MSPinceton-19]anguishing as the princeldest treble	
d buy. Yawn in a semiswoon lay awailing and (hoooh!) [FW474-11_MSPinceton-19]urse could buy. Yawn in a semisw	
honeyful swoothed! (phew!) which [with] earpiercing [FW474-12_MSPinceton-19]what helpings of honeyful swoothe	
dulcitude! As were you suppose to go and push with [FW474-13_MSPinceton-19]dulcitude! As were you suppose to	
hand upinto his fleshaplushfleshaplush cushionettes [FW474-14_MSPinceton-19]your bluntblank pin in hand upinto	
me chubby beyholdboybold love of an angel. Hwoah! [FW474-15_MSPinceton-19]of some chubby [boyhold] love of	

図 1 2 MS Line Diffs

5. まとめと展望

以上のように、TEI に準拠して作成した諸文書を、目的の電子アーカイブの基礎データとして運用し、テキスト生成過程分析の重要な局面を透明化した研究支援ページを様々なビューとして作成していくことが実用レベルで十分可能であることが確認できた。今回の三つのビューはいずれも Excel だけでは可視化できなかったもので、特に、4.2 で示した MS の additions についての緻密な分析は、これまでせっかく Excel にデータを蓄積しながら実際にはほとんど活用できていなかった additions についての情報がフルに活用できるようになったという意味で画期的である。これも含めて、Excel に蓄積してきたデータを汎用性の高い形式で出力・保存・加工、最終的には共同研究のベースとして共有・交換できるようになることは有意義であり、特に量的にも質的にも一個人の働きだけでは到底完結しえない仕事として、研究データを今後も蓄積・精錬していこうとする場合、客観的なルールに従ってデータを残しておくことは研究の持続性という観点からも安心なことである。

今回の P5 2.1.0 へのバージョンアップによって、TEI がテキスト生成過程に関わるデータ記述方法を充実させたのは大変好ましいことであるが、実際にアーカイブを試作してみて気が付いたこととして、不確定性を記述する手段が揃っていないという点を指摘しておきたい。例えば、ヘッダーの graph に文書間の関連情報を記述する場合、node を矢印でつなぐ arc の属性として @from, @to だけでなく、関係の不確定性を記述するための属性（例えば @cert）が欲しい。NB の SO や MS/NB での使用箇所について、複数の候補があって一つに絞りきれなかったり、一つの候補についてより確かな証拠がないと使用箇所と断定できないケースが少なからずあるからである。TEI は個別の文書内情報の記述方法については不確定性の問題も含めて十分開発が進んでいるが、文書間情報のそれについては今後のさらなる開発に期待できるかもしれない。

特にテキスト生成過程研究で扱う文書については、その性質上判読内容が確定できないケースや、文書の一部ないし全部が不明・未発見であるケースはごく普通にあるが、これはそれ自体の不確定性の問題として自己完結するものではないことを、この仕事に携わる研究者の立場から強調しておきたい。個々の文書内にみられる判読の不確定性は、文書間情報の緻密な分析によって解明されることが多々ある。今回の試作で透明化させた研究の局面に即して言えば、NB の判読不明箇所は SO の判明や MS の使用箇所の特定によって判読可能になることがよくあり、また MS での使用箇所候補が複数あって確定できない場合には、すでに確定しているほかの文書間情報を参照枠として総動員することによって確定の可能性を高めることができる。よりスケールの大きな「不確定」としての未発見 SO の発見や、散逸した NB-D を復元していく作業についても同様である。この意味でも、個別の文書だけでなく文書間情報についても、不確定要素を確定要素とダイナミックな相補関係にあるものとして同等にきちんと記述しておくことは大変重要なのである。

最後にこれからの試作に関して、次にぜひとも達成したいのはヘッダーに入れてある各文書の作成年代の有意義な運用である。時間のパラメーターを単独文書内の記述に精緻に盛り込む試みはすでに Pierazzo によってなされているが[8]、文書間情報のそれについては P5 にも説明や実例はほとんど出てこない [9]。先述のようにこの情報の活用はテキスト生成過程分析には必須のカギなので、この活用例をぜひとも実現しておきたい。

参考文献

- [1] 扶瀬「ジョイスとサイバースペース」『英語青年』2005年1月号、6-7.
- [2] P5: *Guidelines for Electronic Text Encoding and Interchange*. Version 2.1.0. Last updated on 17th June 2012. <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>.
- [3] “20.2 Boundary Marking with Empty Elements.” P5, 613+.
- [4] “2.2.7 The Source Description” and “19.1 Graphs and Digraphs.” P5 30+ and 581+.
- [5] “1.3 Document vs. Text.” *An Encoding Model for Genetic Editions*. <http://www.tei-c.org/Activities/Council/Working/tcw19.html>.
- [6] “Editorial Principles and Practice: Facsimiles and Transcriptions.” Dirk van Hulle and Mark Nixon (dirs.), *Samuel Beckett Digital Manuscript Project* (2011). <http://www.beckettarchive.org/editorial.jsp>.
- [7] Robbert-Jan Henkes and Mikio Fuse, “Inside D1.” *Genetic Joyce Studies* 12 (2012). http://www.antwerpjamesjoycecenter.com/GJS12/GJS12_Henkes_Fuse.html.
- [8] Elena Pierazzo, “Digital Genetic Editions: The Encoding of Time in Manuscript Transcription. Marilyn Deegan and Kathryn Sutherland (eds.), *Text Editing, Print and the Digital World* (Ashgate, 2009), 169-86.
- [9] P5 2.1.0 は “11.3.1.5 Substitutions” で、ページ内の字句の改訂について add/delete 要素の @seq に時間的パラメーターを緻密に組み入れて記述する方法を提案し、これに関連する話として Joyce の *Ulysses* の共観版(synoptic edition)で MS のすべてのレベルを番号のついた編集記号を駆使することによって一つの(紙の)ページ上に記述した Hans W. Gabler の仕事を引き合いに出し、“11.7 Changes”のヘッダーにおける stage 情報の記述方法を参照させて、Gabler と同様の方法で文書間レベルのテキストの変化を記述できる可能性を示唆しているが、実際にその情報をどの MS 文書のヘッダーにどのように入れるのかなど、実例を含めてより具体的な説明が欲しいところである。