GENETIC JOYCE STUDIES – Issue 17 (Spring 2017)

Introduction to *The* Finnegans Wake *Genetic Research Archive* (in Progress) Part 1. The Document Data

Mikio Fuse

1. Introduction

The Finnegans Wake *Genetic Research Archive* started in 2012.[1] It is basically a single-handed project, whose aim is to reorganize my personal *Finnegans Wake* genetic research routines, to make them more efficient and more sustainable. Because it is web-based, it is expected that, if all conditions are met, the archive may serve as a shared tool for serious researchers. After five years, the archive still needs to be filled with a great number of essential document data, but the basic question of how to encode and store the document data has been solved. In this first installment of my introduction to the archive, I would like to describe how and in what format the document data have been created, stored, and maintained. How the archive mobilizes the data and what it looks like from the user's point of view (the user interface) are topics that I will develop later.

2. An overview of the archive's document data

Key components of the archive are two types of files on the server: XML document files and CGI script files. The former store the data of various kinds of documents involved in the genetic process of *Finnegans Wake* (hereinafter *FW*), while the latter mobilize the document data for specific purposes and present the results in HTML web pages. In this article, I will focus on the former static aspect of the archive, explaining the variety of XML document files created and stored in this archive.

The XML files that encode the documents involved in the genetic process of *FW* are grouped into five categories and are stored accordingly in five separate subfolders in the **"doc" directory**.

2.1 The **FW.xml** file, placed in the **"FW" subfolder**, is the document data file of the printed text of *FW*[2]. Aside from such basic information as the text content, Book, Chapter and Section divisions, line and page numbers, it also stores the information added on the editor's responsibility about those "segments" in the *FW* text that are to be referred to when presenting or analyzing the genetic process involving them.

---

[1] Special thanks are due to Asanobu Kitamoto, Kiyonori Nagasaki, Elena Pierazzo, and James Cummings for their instruction and advice on the technicalities of digital archiving.

[2] The text is based on the 1939 Faber edition.

2.2 The **"SO" subfolder** stores all the different source documents that have been suggested as the sources of the Notebook entries. As in the *FW* document file, the Source document files also contain the "segment" information added by the editor. As explained below, the files are so named that they serve as short-hand identifiers of individual source documents.

2.2.1 The filenames of book sources are shorthand reminders of their authors' names, titles and publication years. For example, the file for Eugene O'Neill's *The Hairy Ape*, published in 1922, is: **SO_ONeill-HairyApe-1922.xml**.

2.2.2 For newspaper sources, one file corresponds to one newspaper page that is known to include the sources for Joyce's Notebook entries. For example, **SO_NP-FJ-19240204-06.xml** is the document file for page 6 of the 4 February 1924 issue of the *Freeman's Journal*.

2.2.3 For encyclopedia sources, the files are created by articles and are named after the article titles. For example, **SO_EB11-CETACEA-1911.xml** is the document file of the article on "Cetacea" in the *Encyclopedia Britannica*, 11th edition.

2.3 The Notebook document files, stored in the "**NB" subdirectory**, are named after the Buffalo Notebook numbers, such as **NB_B03.xml** and **NB_C01.xml**. Not only do they contain the data for the Notebook documents themselves (e.g. transcriptions, unit breakups and cross-out color information), but they also include crucial information about the genetic process: where a particular unit came from (Source), where it was directly used or transferred (Notebook, Notesheet, Manuscript), and where it ended up in the printed text (*FW*). Those essential analytical dossier data can be found in the header section of each Notebook document file.

2.4 The **MS subfolder** stores all the Manuscript document files. They are created for each level of each Section of each Book and Chapter of *FW*. For example, **MS_1-5-4_3.xml** is the document file for the level 3 Manuscript of Book I, Chapter 5, Section 4. The files **MS_1-5-4_3p.xml** and **MS_1-5-4_3x.xml** are respectively for the duplicate manuscript ("p" for plus) and extradraft material ("x" for extra) of the same section-level.

2.4.1 It should be noted that the editor has chosen to create the Manuscript document files not by individual Manuscript pages (the physical entities) but by draft levels (analytical concepts), citing the Manuscript passages, more often than not from more than one Manuscript pages, that correspond to each page and line of the printed text of *FW*. The choice of this analytical approach to Manuscript materials is in accordance with the requirements of this "genetic research" archive. In order to describe the genetic process of *FW*, it is essential for the editor to manually collate the Manuscript text and the printed text by analyzing the correspondence in terms of the *FW* page and line numbers. In

this archive the process of collation is finished before the Manuscript files are created, not after, so that the exact physical properties of individual Manuscript pages, including line breaks, are not encoded in the Manuscript files. The editor expects that the lack of physical information will be compensated by enabling the user to refer to the facsimile images of corresponding Manuscript pages.

2.5 The major components of the **NS subfolder** are the two files for the two sets of notesheets involved in the genesis of *FW*, i.e., the ***transition*** **Notesheets** (**NS_NStr.xml**) and **Jahnke Bequest Notesheets** **(NS_NSjk.xml)**. Because the page numbers of both materials assigned by librarians do not agree with the actual page order in which Joyce created them, both Notesheet files encode both the "original" and the "correct" page orderings.[3]

The NS subfolder also stores the file **NSx.xml:** a collection of select pages of other miscellaneous extradraft materials that contributed to the drafting of Manuscript pages. As mentioned above, extradraft materials are also found in the Manuscript subfolder where Manuscript files with the "x" filename suffix include citations from extradraft materials corresponding to given *FW* page and line numbers. Again, this double-standard approach to extradraft materials reflects the editor's decision to make this archive a "genetic research" tool. It prioritizes the encoding of the relationship between documents rather than the encoding of individual documents *per se*. The extradraft document files in the MS subfolder serve the former purpose. However, because it is not always possible to break the text of extradraft materials into segments that neatly correspond to particular *FW* page and line numbers, there are inevitable stray words and passages that are not included in the collated extradraft document files in the MS subfolder. The file **NSx.xml** in the NS subfolder makes up for this weakness by faithfully encoding the pages of such extradraft materials, in order to resuscitate the stray passages and fragments. For example, this double-standard approach becomes relevant when encoding the extradraft materials for Book II, Chapter 4, Section 2, Level 8 ("Mamalujo"), where the text contains the germs of the printed text of *FW* in a tantalizingly fragmentary manner.

2.6 Dossier files

Aside from all the above-mentioned data files for the basic documents, the "doc" folder also has a **"dossier" subfolder** where the files for certain meta-document data are stored. As mentioned above, the Notebook document files have the basic dossier-level data in their headers, so theoretically any analysis of the genetic process should be executable employing those header data. However,

[3] The "correct" page orderings are based on the editor's unpublished analysis adopting the method suggested by Dirk Van Hulle.

some analysis requires an extensive reference to a copious collection of different Notebook header data, so that it is not practical and at times impossible to process all the lookup tasks on the web server. That is why the archive is equipped with these kinds of pre-processed data files. Currently the major dossier files installed are those related to the "orphan" additions on a particular section-level of Manuscript documents. For example, **DOS_3-3A_8-NBorphans.xml** is a collection of all those additions made on Book III, Subsection3A, Level 8, whose "parent" Notebook units have not been located yet.

3. TEI

With the exception of makeshift "dossier" files, all the XML document files mentioned in the previous section are created in accordance with the TEI [Text Encoding Initiative] Guidelines, which is "the format recommended for preservation and interchange of electronic textual resources by a number of funding bodies for arts and humanities projects" (Cummings). By adopting this open standard, the editor of the archive can not only (1) assign the tag or element names (e.g. <add>, <del>) and their attribute names (e.g. <del type="used" hand="#blue">) in a consistent and sustainable way (Preservation), but also (2) share and exchange the document files with other archivists who follow the same guidelines (Interchange).

The TEI XML guidelines are constantly updated. The archive follows Guidelines P5, version 2.5.0 (released 26 July, 2013), but the latest version as of this writing is P5, version 3.1.0 (released 15 December, 2016). The archive's document files can be updated without difficulty using a conversion script written for that purpose, but it is not in the editor's scope to do so very soon, given that there is a lot of other work of higher priority.

It is one of the landmark features of TEI P5 Guidelines version 2+ that they substantially address handwritten manuscript sources in a dedicated chapter (Chapter 10 "Manuscript Description") on top of printed sources (Chapter 4 "Default Text Structure"). In the archive, the documents in the "FW" and "SO" subfolders follow the rules for printed texts, while those in the "NB," "NS," and "MS" subfolders follow the rules for manuscripts.

As Cummings clarifies, the TEI does not intend to provide "monolithic rules" applicable to all possibilities; instead, "it intends to be customized and modified." It is therefore the editor's responsibility to clarify where he has customized and modified the rules. It is also necessary for the editor to explain where and how he has decided to encode those analytical dossier data that are not physically present in individual documents *per se*. In the following sections, I will highlight some select instances of how the TEI Guidelines have been applied in the archive's basic document files.

3.1 The encoding of text-type documents (SO, *FW*)

Joyce's Notebook entries have different types of printed texts as their Source materials, such as books, journals, and newspapers. All these materials excepting

newspapers can be encoded as the TEI's "Default Text Structure": the body of text is encoded as the **<body>** element, inside which the largest child elements are the **<p>** ("paragraph") elements, which in turn may contain the **<pb>** ("page break") and **<lb>** ("line break") elements. The same "text" encoding rules apply to the printed text of *FW*.

For newspapers, the Guidelines recommend (1) regarding a collection of newspaper issues as a text corpus in which an individual issue qualifies as normal text material, like a book or a journal,[4] and (2) marking individual articles by the division tags (** **) and their headers and bylines by the **<head>** and **<byLine>** tags.[5] The normal "text" structure tags like **<p>**, **<pb>** and **<lb>** can be used inside the ** ** tags.

Because newspaper columns, unlike encyclopedia columns, are not always used one by one from page top to page bottom, and because it is customary in *FW* genetic studies to refer to newspaper sources by page and column numbers, the editor uses the values of the **xml: id** attribute of the ** ** elements to explicitly refer to the "blocks" and "columns" where the articles, advertisements, etc., can be found. For example, the first advertisement on Page 6, Columns 1-2 of the 16 May 1925 issue of the *Connacht Tribune* is encoded within the division tag ** **, where the **n** attribute indicates the column number(s) and the **block** suffix in the **xml:id** value refers to one of the "blocks" the editor has mapped out on the thumbnail image of the original newspaper page:

Aside from the question of how to encode those structural features unique to newspapers, the newspaper Source raises a more basic problem for this archive: the editor has to decide how extensively and exhaustibly he wants the archive to be able to *internally* refer to newspaper documents as they are encoded and stored *inside* the archive.

---

[4] "TEI: P5 Guidelines" 15.1. As mentioned above in Section 2.2.2, the newspaper document files are currently created by newspaper pages. At an advanced stage, they might well be amalgamated into corpus files by newspaper titles. Individual "article" files of an encyclopedia are also subject to future amalgamation into a text corpus file.

[5] "TEI: P5 Guidelines" 4.1. (esp. 4.1.4) and 4.2.1.

Because it is a genetic "research" archive, the editor thinks it is not enough to show where and how the passages in the Source were used in the genetic process; the archive should also enable the researcher to discover more source passages that have eluded attention. In this light, it is ideal to encode not only the articles, advertisements, etc. that include source passages but all the pages of all the issues that include the sources, and, if possible, all the pages of all the other issues that may include yet-to-be discovered sources. That, however, is an impossible dream. For the moment, the editor's practical solution is to include some sample pages where Notebook sources have already been discovered, while leaving the researchers who have access to libraries and/or external newspaper archives to independently check the resources that have not been included in the archive.

On top of encoding the basic "text" structure of Source documents and the printed text of *FW*, the editor adds the **<seg>** ("segment") tags to encode those parts of the Source text that are the "parents" of Notebook units, and those parts of the *FW* text that are the "children" of consequent Manuscript additions. These **<seg>** tags are major editorial additions based on his or his precursors' analysis. Here are the examples:

Addition of **<seg>** tags in the Source book file **SO_Heard-Narcissus-1924.xml** (Gerald Heard, *Narcissus: An Anatomy of Clothes*, 1924):

```
<p><lb xml:id="Heard-Narcissus-1924_013-2"/><s rendition="#center #large">A METAPHYSIC
                OF MODE</s></p>
<p><lb xml:id="Heard-Narcissus-1924_013-3"/>If, then, we may assume the psychological
                <lb xml:id="Heard-Narcissus-1924_013-4"/>commonplace that the unperceived is
                <seg xml:id="Heard-Narcissus-1924_013-4_seg1"><hi rendition="#italic">ipso <lb
                        xml:id="Heard-Narcissus-1924_013-5"/>facto</hi></seg> the vital, there
        can be no more <lb xml:id="Heard-Narcissus-1924_013-6"/>striking example of it than
```

Addition of **<seg>** tags in ***FW*.xml**:

```
<p><lb xml:id="FW003-15"/> The fall (bababadalgharaghtakamminarronnkonnbronntonner- <lb
                xml:id="FW003-16"/>ronntuonnthunntrovarrhounawnskawntoohoohoordenenthur- <lb
                xml:id="FW003-17"/>nuk!) <seg xml:id="FW003-17_seg1">of a once wallstrait
                oldparr</seg> is retaled early in bed and later <lb xml:id="FW003-18"/>on life
        down through all christian minstrelsy. The great fall of the <lb xml:id="FW003-19"
```

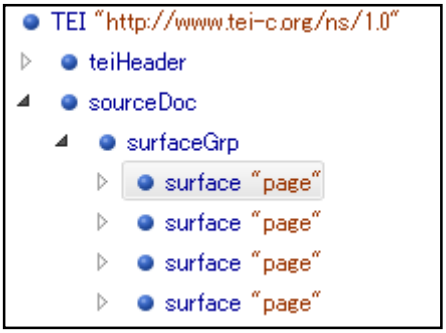
For the *FW* document file, the **<seg>** element also helps to mark the point of so-called transmissional departure, where a given MS addition has no corresponding "child" segment in the printed text. In that case, an "empty" **<seg>** element, namely **<seg/>**, is added to that point, and is given the value "**TransDep**" in its **type** attribute. See the following example:

```
called no name at all. Together. <lb xml:id="FW044-15"/>Arrah, leave it to Hosty,
frosty Hosty,<seg xml:id="FW044-15_seg1" type="transDep"/> leave it to Hosty <lb
        xml:id="FW044-16"/>for he's the mann to rhyme the rann, the rann, the rann, the
king <lb xml:id="FW044-17"/>of all ranns.<seg xml:id="FW044-17_seg1" type="transDep"
/> Have you here? (Some ha) Have we where? (Some <lb xml:id="FW044-18"/>hand) Have
you hered? (Others do) Have we whered? (Others dont) <lb xml:id="FW044-19"/>It's
```

3.2 The encoding of manuscript-type documents (NB, NS, MS)

Although the TEI started as "Text" Encoding Initiative, it now also addresses "Manuscript" encoding. That is a breakthrough for the genetic archivist, for handwritten materials like Notebook, Notesheet and Manuscript documents do have a number of specific "document" (as against "text") features to encode, and the editor can confidently turn to the TEI Guidelines to create consistent encodings.

The basic structure of TEI Manuscript encoding differs from that of TEI Text encoding in that the main body is encoded as the **<sourceDoc>** element, instead of the **<text>** element. As the element structure outline below shows, the largest child element inside the **<sourceDoc>** element is **<surfaceGrp>**, which contains individual **<surface>** elements.[6] Each **<surface>** refers to an individual document "page," thus its **type** attribute has the value "**page**."



The **<surface>** element has two major child elements: the **<graphic>** element is to indicate the location of the facsimile image of the page, and the **<zone>** elements are to define particular areas on the page. In Notebook and Notesheet document files, the **<zone>** elements refer to individual units, while in Manuscript files the **<surface>** element has only one child **<zone>** referring to the selfsame page surface, because the editor has chosen not to encode any smaller partitions on the Manuscript page.[7]

In the following sections, I will explain the application of the TEI manuscript

---

[6] "TEI: P5 Guidelines" 11.1.

[7] This is another instance of the editor's choice of sacrificing the physical, as against textual/analytical, encoding of Manuscript pages *per se*. In a truly physical encoding, the archivist may want to define zones where marginal and interlinear additions have been made, so that the analysis of the zone data, the presentation of the images of the zones, etc., may eventually be facilitated.

encoding rules to each of the three kinds of manuscript-type documents and where I have customized and modified the rules.

3.2.1 Manuscript

The archive prioritizes the analysis of the genetic process at the expense of the physical description of Manuscript pages *per se*. The policy is reflected in the use of the **<lb>** ("line break") elements inside the **<zone>** element in the Manuscript file. They do not encode the line breaks of the handwritten lines on the Manuscript page but those of the corresponding *FW* text. As the following example demonstrates, each line break has a unique **xml:id** value which explicitly indicates the corresponding *FW* page and line numbers:

```
<surface type="page">
   <graphic url="./images/MS_47471a-003.jpg"/>
   <zone>
      <lb xml:id="MS47471a-003_FW003-17"/><add xml:id="MS47471a-003_FW003-17_add1">of
         a once wallstreet oldparr</add> is retaled early in bed and later <lb
         xml:id="MS47471a-003_FW003-18"/>on life down through all christian
      minstrelsy. The great fall of the <lb xml:id="MS47471a-003_FW003-19"/><del
```

The above example includes an instance of the **<add>** element which marks an addition. Each **<add>** element has a unique **xml:id** that derives from the **xml:id** of the line break immediately preceding it, with the suffix **add1** (or, if it is already used, **add2**) attached to the **xml:id** of the **<lb>** element.

Notebook units do not always find their children in Manuscript "additions." If a Notebook unit finds its way into the base text of a Manuscript, the corresponding passage in the Manuscript is marked as a **<seg>** element, with its **xml:id** assigned according to a policy similar to that of the **<add>** element. See the following example:

```
ancient <del type="cancelled">of</del><add xml:id="MS47471a-008_FW007-16_add2"
   >from out</add> the ages of the Agapemonides, <lb
   xml:id="MS47471a-008_FW007-17"/>he is <seg
   xml:id="MS47471a-008_FW007-17_seg1">smolten</seg> in our mist, woebecanned
and packed <lb xml:id="MS47471a-008_FW007-18"/>away. So that meal's dead off<add
```

3.2.2 Notebook

Unlike Manuscript document files, Notebook document files faithfully encode their physical features: the **<surface>** element refers to a Notebook page, and the **<lb>** elements refer to the actual line breaks on the page. While these two elements encode the physical features of Notebook pages, Notebook units, encoded as **<zone>** elements, are editorial additions based on the editor and his precursors' analysis, and are subject to revision in case of a new Source discovery. Here is an example showing the encoding of the first two units of Notebook B6:

```
<surface type="page">
        <graphic url="./images#NB_B06-001.jpg"/>
        <lb xml:id="Lb34b02d01b06c0204847_B06-001-1"/>
        <zone type="unit" ulx="0" uly="0" lrx="0" lry="0"><graphic
                        url="./images/NB_B06-001.jpg"/>
                <del type="used" hand="#red">
                        <seg type="unit" xml:id="Lb34b02d01b06c0204847_B06-001_a">
                                <del type="cancelled" cert="unknown">p*ay</del><add>pray</add> fervently
                                they [may] <lb xml:id="Lb34b02d01b06c0204848_B06-001-2"/> not depart
                                this life <lb xml:id="Lb34b02d01b06c0204849_B06-001-3"/> till they have
                                ---- </seg>
                </del>
        </zone>
        <lb xml:id="Lb34b02d01b06c0204850_B06-001-4"/>
        <lb xml:id="Lb34b02d01b06c0204851_B06-001-5"/>
        <zone type="unit" ulx="0" uly="0" lrx="0" lry="0"><graphic
                        url="./images/NB_B06-001.jpg"/>
                <del type="used" hand="#red">
                        <seg type="unit" xml:id="Lb34b02d01b06c0204851_B06-001_b"> treasured
                                unkindly words </seg>
                </del>
        </zone>
```

The above example encodes five line breaks (lines 1 to 5). Each of them has a unique **xml:id** that not only indicates the page and line numbers explicitly (e.g. "**B06-001-1**") but is prefixed by a curious code like "**Lb34b02d01b06c0204847**," where the string "**b34b02d01b06c02**" indicates that the group of Notebooks B34, B2, D1 and B6, finds its scribal transcription in Notebook C2 in that order. The last serial number "**04847**" in the above example refers to the line number of the Excel spreadsheet used in collating the group of authorial Notebook entries with the entries of the scribal Notebook. (For a snapshot of this Excel spreadsheet, see 4.2).

In the above example, each **<zone>** (=unit) has a set of four attributes that define the coordinates of the four corners of the zone on the facsimile image of the Notebook page.[8] These coordinates will be useful when it comes to comparing the image of a particular authorial Notebook entry with that of the corresponding scribal Notebook entry, using the CSS image sprite technique.

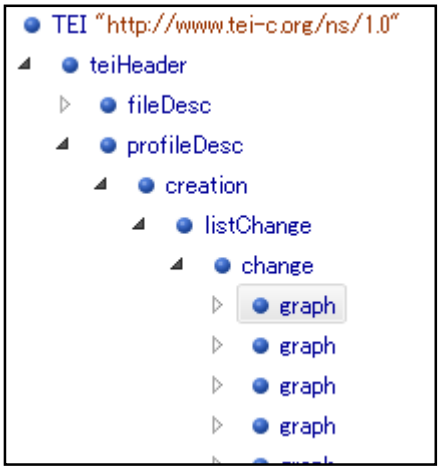
Each unit **<zone>** has a child **<del>** ("delete") element whose **type** attribute specifies that the deletion was meant to mark the entry as "**used**." The **hand** attribute of the same element encodes the cross-out color, like "**#red**" or "**#blue**." If the unit is not crossed out, the value is "**#N/A**"; if the uncrossed-out unit proves to have been used as an "x-deleted" unit, the value is changed to "**#x**."

The transcription of a Notebook unit is encoded in the child **<seg>** element. Its **xml:id** consists of the Notebook page and unit code (e.g. **B06-001_a**) preceded by the above-mentioned code indicating the group of the authorial and scribal Notebooks. If the unit code is assigned by the editor himself, the alphabetical code is

[8] In this example their values are all null because the editor has not yet provided them.

given in upper case.

As mentioned above, the most important application of the TEI XML rules to the archive is the use of the header to describe the whole genetic process involving each Notebook unit. The analytical dossier data are collected in the **<change>** element whose child **<graph>** elements encode the information on the genetic origins and destinations of individual Notebook units.[9]



For a simple example, below is the first **<graph>** element in the file for Notebook B6. It contains three **<node>** elements followed by two **<arc>** elements.

```
<change><graph>
        <node xml:id="B06-001_a-NB-BD" value="#Lb34b02d01b06c0204847_B06-001_a"/>
        <node xml:id="B06-001_a-MS1"
              value="./doc/MS/MS_3-2C_1.xml#MS47482b-017v_018_FW472-28_add1"/>
        <node xml:id="B06-001_a-FW1" value="./doc/FW/FW.xml#FW472-28_seg2"/>
        <arc from="#B06-001_a-NB-BD" to="#B06-001_a-MS1"/>
        <arc from="#B06-001_a-MS1" to="#B06-001_a-FW1"/>
    </graph>
```

These data on "nodes" and "arcs" define a graph describing how the child of unit *(a)* on page 1 can be located in the Manuscript file for Book III, Subsection 2C, level 1, in an addition made on the set of Manuscript pages 47482b-17v and 18, and how the addition corresponds to a segment starting at *FW* 472.28 in the FW.xml file:[10]

If a given path of transmission is not certain, the **<arc>** element encoding that path will have the **cert** attribute with the value "**unknown**." Because the TEI Guidelines do not expect the "cert" attribute to be used for the "arc" element, this is the first (and so far only) case of modifying the TEI rules in this archive.

---

[9] "TEI: P5 Guidelines" 19.1.

[10] Should the parent Source of the Notebook unit be discovered in the future, the **<graph>** would have another **<node>** for the corresponding segment in the Source, along with a new **<arc>** element to indicate the direction of transmission from the Source segment to the Notebook unit

3.2.3 Notesheet

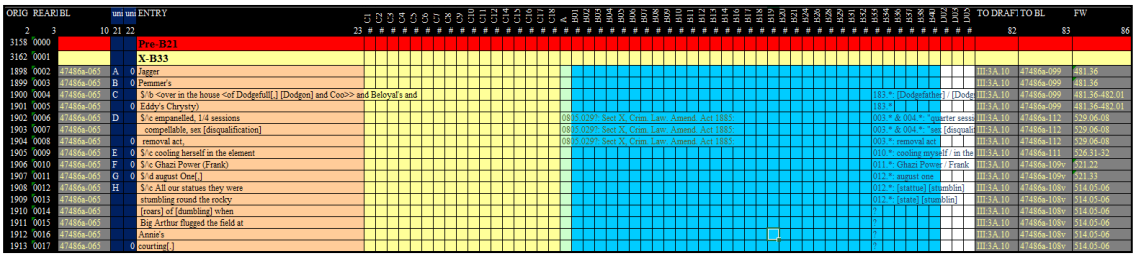
The encoding of the two extensive Notesheet materials, *transition* Notesheets and Jahnke Bequest Notesheets, follows virtually the same rules as Notebook encoding, except for the following three points:

(1) The Notesheet document files have no **<change>** element in the headers to describe the paths of genetic transmission, because it is the job assigned to the Notebook headers exclusively.

(2) While the **xml:id** of a line break in a Notebook file derives from the Excel spreadsheet used in collating the scribal Notebook entries with the authorial Notebook entries, the **xml:id** of a line break in a Notesheet file derives from the Excel spreadsheet used in analyzing the Notebook origin of each Notesheet unit in order to establish the "correct" order of the set of Notesheet pages. For example, below is the beginning of the first **<surface>** (=page) of *transition* Notesheets whose first line break has the **xml:id** value "**NStr_47486a-065-1_0002X-B33_1898**."

```
<surface type="page">
    <graphic url="./images/NStr_47486a-065.jpg"/>
    <lb xml:id="NStr_47486a-065-1_0002X-B33_1898"/><zone type="unit"
        ulx="0" uly="0" lrx="0" lry="0"><del type="used" hand="#NA"><seg
            xml:id="NStr_47486a-065_A">Jagger</seg></del></zone><lb
        xml:id="NStr_47486a-065-2_0003X-B33_1899"/><zone type="unit" ulx="0" uly="0"
        lrx="0" lry="0"><del type="used" hand="#NA"><seg xml:id="NStr_47486a-065_B"
            >Pemmer's</seg></del></zone><lb
        xml:id="NStr_47486a-065-3_0004X-B33_1900"/><zone type="unit" ulx="0" uly="0"
        lrx="0" lry="0"><metamark>b</metamark><del type="used" hand="#x"><seg
            xml:id="NStr_47486a-065_C"> $/¥b <add>over in the house <add>of
                Dodgefull[,] [Dodgon] and Coo</add></add> and Beloyal's
            and<lb xml:id="NStr_47486a-065-4_0005X-B33_1901"/> Eddy's
        Chrysty)</seg></del></zone><lb
```

The segment "**47486a-065-1**" indicates that it is the first line of British Library Additional Manuscript 47486a-065. The code "**0002X-B33**" refers to the line number on the Excel spreadsheet when it is sorted by the rearranged "correct" order ("**0002**") as well as the arbitrary stack name the editor used in grouping the pages according to their Notebook origin ("**X-B33**"). The last suffix "**1898**" refers to the line number on the Excel spreadsheet when it is sorted by the "original" order as they are arranged in the *James Joyce Archives*.



(3) The encoding of the third unit in the above example has a **<metamark>** element as a child element of the unit **<zone>**. It encodes the "sigla" Joyce marked on the unit.

4. The creation and maintenance of document files

In this final section, I will explain how the XML document files are created and, once they are created, how they are maintained and updated.

4.1 The creation of XML document files

The XML document files are not created from scratch. They are automatically created from the Excel files where the editor has stored the transcription and analysis data. This method of automatic conversion saves a lot of time for a virtually single-handed archiving project like mine. The transformation of the data on Excel spreadsheets to XML files is executed by various scripts written for that purpose in Perl. Although it is not the latest programming language, it is versatile in manipulating text data and is conveniently equipped with modules like **Spreadsheet::ParseExcel** and **XML::LibXML**.[11]

The files converted from Excel spreadsheets should not only be valid XML documents but should also be validated against the schema that defines the TEI P5 rules. This validation is to be easily executed by opening the created file in an XML editor called **Oxygen**[12] and associating the file with the TEI schema for validation. Oxygen is particularly helpful in checking the converted Manuscript document files where there are many instances of additions (**<add>**) and deletions (**<del type= "cancelled">**). In the Excel spreadsheet, the editor has used specific combinations of less-than (<) and greater-than (>) signs to mark the segments that are added or deleted.

[11] The former helps obtaining the values of specified cells of the target Excel spreadsheet, while the latter helps obtaining the values of specified elements and attributes of the target XML file, as well as modifying the elements, attributes and, if necessary, the whole structure of the XML file. The latter module also plays a vital role in the web page presentation of obtained/manipulated data.

[12] https://www.oxygenxml.com/

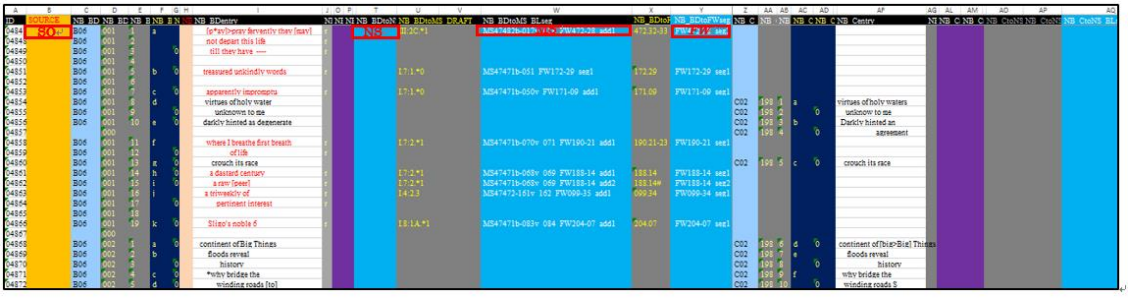
Bk-(	Sec	Page	Line	Text	0 DATE 0	BL 0	0
3-2	A	452	25	breadth from pride I am (breezed be the healthy same!) for 'tis a	#240300	* 47482b-006v	it is
3-2	A	452	26	grand thing (superb!) to be going to meet a king, not an everynight	#240300	* 47482b-006v	grand to be going to meet a king. <<<Not a
3-2	A	452	27	king, nenni, by gannies, but the overking of Hither-on-Thither	#240300	* 47482b-006v	king only in name but> the king of
3-2	A	452	28	Erin himself, pardee, I'm saying. Before there was patch	#240300	* 47482b-006v	Greater Dublin, too, the first Humphrey.>
3-2	A	452	29	at all on Ireland there lived a lord at Lucan. We only wish	#240300	* 47482b-006v	I wish
3-2	A	452	30	everyone was as sure of anything in this watery world as we are	#240300	* 47482b-006v	everyone was as sure of anything <in the real world> as I am
3-2	A	452	31	of everything in the newlywet fellow that's bound to follow. I'll	#240300	* 47482b-006v	of everything <in the other>>.
3-2	A	452	32	lay you a guinea for a hayseed now. Tell mother that. And tell	#240300	* 47482b-006v	<Tell mother that.

The Perl script written for the Manuscript conversion replaces them to **<add>** and **<del>** elements according to the replace algorithm configured for that particular purpose. However, the results are more often than not imperfect because the Excel data sometimes include a very complex structure of less-than and greater-than sign combinations. When a file including conversion errors is opened in Oxygen, the XML editor shows red wavy lines wherever the TEI XML structure is corrupt. In that case, the editor goes back to the Excel spreadsheet (and the facsimile) and manually corrects the errors.

4.2 The data maintenance (corrections and updates)

Even if valid document files are created, they are subject to modifications and updates. There are two "bases" where the data are modified and updated: one is the Excel spreadsheet and the other is the XML document file.

When a new path of genetic transmission has to be added or when an old path should be modified, the Excel spreadsheet for the relevant Notebook is opened. For each unit the spreadsheet has cells for the **xml:id** value of its parent segment in the Source document file and the **xml:id** values of its child and further offspring in the Notebook, Notesheet, Manuscript and *FW* document files. Below is a sample snapshot of the Excel spreadsheet showing the units on the first page of B06:



Next, relevant Source, Notebook, Notesheet, Manuscript and *FW* document files are opened and the **xml:id** values of the corresponding segments, units and additions of the related documents are copied and pasted to the Excel spreadsheet cells. In the above example, all crossed-out units have known child additions and segments in Manuscript documents along with the corresponding segments in *FW* (all in terms of the **xml:id** values of the additions and segments), whereas the orange cells for Source documents are yet to be filled in (in terms of the **xml:id** values of the corresponding segments in the

documents). Thus, it would be the orange cells where the updates should be made, should the Source(s) be discovered. When the Excel spreadsheet is updated, it is converted to a new XML Notebook document file whose header has now updated **<graph>** elements.

In this way, the most vital information on the genetic process of transmission is always stored and updated in the Excel spreadsheets for Notebook documents. That is definitely an easier and more efficient method than manually updating and modifying individual XML document files directly.

Correction of transcription errors and revision of unit code assignment in Notebook and Notesheet document files are also made on the related Excel spreadsheet, from which new document files are generated to replace the old ones.

Meanwhile, because the Source, Manuscript and *FW* document files are converted from Excel spreadsheets once for all, all the corrections of errors are made on the XML document files. These files are also subject to manual updates when new **<seg>** elements have to be added to mark them as nodes involved in new or modified paths of genetic transmission.

5. Conclusion

This first installment of my introduction to *The* Finnegans Wake *Genetic Research Archive* has concentrated on the static aspect of document data creation and storage. The smooth and seamless transition from the old method of storing the document data on Excel spreadsheets to the new method of creating TEI XML document files from the old Excel data, while still using the Excel spreadsheets for data storage and maintenance, has opened up a number of new prospects.

The most significant new departure is both the speed and accuracy with which the archive will be able to help the researcher find new genetic paths of transmission and verify those already established. Key features that were not available in the old method are that the archive distinguishes all the possible nodes of genetic transmission (the Source segments, the Notebook and Notesheet unit zones, the Manuscript additions and segments, the *FW* segments) by their unique ids, and that each document that includes those nodes has information on the document's date (if available[13]) in the header, so that it becomes possible to match childless elements (e.g. Notebook units) with parentless elements (e.g. Manuscript additions) only from among selected candidates.[14]

The new hope demands old patience, however. There remain two major tasks to do in terms of static document data storage. One is the installment of all the Manuscript document files that will take several more years. The other is the installment of all the Notebook

[13] The dates are based on those given in the *James Joyce Archive* and, if available, the Brepols *Finnegans Wake* Notebook Editions.

[14] In my old method, the matching relied on a meta-document text search tool, where all it can suggest was simple string matches only.

document files. Although the editor is in possession of the transcriptions and analyses of many Notebooks, thanks to Geert Lernout and other scholars' generous help, they have yet to be collated (if applicable) with their related sibling and child Notebooks before individual Notebook document files are created.

Moreover, even when those two tasks are completed, the archive is to remain a work in progress, because the editor thinks the transcription and analysis data should always remain open to revision, and that this openness should be what qualifies the archive as a genuine "research" tool, rather than a definitive edition.

For these reasons, the editor thinks the project should sooner rather than later enter into a stage where it allows at least a discreetly limited number of users to test it out and report how it helps (or not) with their genetic research work, how easy (or not) it is to suggest emendations, etc. However, we can talk about all these possibilities only after the archive's user interface has been consolidated. And that is the subject of the second part of my introduction to this archive.

Works Cited

Cummings, James. "The Text Encoding Initiative and the Study of Literature." *A Companion to Digital Humanities*. Ed. Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell, 2004. http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-6-6

"TEI: P5 Guidelines." *TEI: Text Encoding Initiative*. http://www.tei-c.org/Guidelines/P5/