

Using the SAS System to Produce HTML-Based Text

Richard D. Langston, SAS Institute Inc.

ABSTRACT

The author has previously produced a poster session regarding the presentation of genealogical data. The next step in the process of producing a useful document of research is to use emerging HTML technology in conjunction with the World Wide Web. The author has used the SAS® System to produce a set of HTML files consisting of text and image data. The resultant files can be examined by common Web browsers and can also be converted into a serial form suitable for publication.

INTRODUCTION

In looking for ways to present my genealogical research, I found the most effective means was to provide individual family worksheets with page references where the parents were children and where the children were parents. The system I developed to produce these worksheets was extensively documented in a poster session presented at SUGI 19 (see *Proceedings*, pp. 914-923).

One of the future enhancements cited in this poster was the fact that I was interested in a graphical interface to simplify the browsing of the database. I envisioned that one could use a point-and-click interface to follow the thread of relationships through many different worksheets.

Originally I had intended to develop this using SAS/AF® Software, but with the advent of HTML and popular Web browsers, I realized that recreating the worksheets as HTML files would be a much more attractive approach. These sets of HTML files could reside on CD-ROMs and be accessed via a local browser. Further, these files could reside on the World-Wide Web and be accessible by anyone with Web access.

Thus this paper describes the mechanism I implemented in order to produce a set of HTML files to represent my genealogical research.

A further enhancement made during the development was the addition of a photographic database. HTML allows for easy integration of images into its files, so this feature is also discussed and demonstrated.

BRIEF OVERVIEW OF THE GENEALOGICAL DATABASE

My genealogical database consists of people who are blood-related to me in some way. Each person has a key associated with them so that they can be properly linked into the database. The key is based on that person's relationship to me. For example, my mother's father would have the key of MF. The key is ultimately derived based on traversing up a direct line, across to a sibling if necessary, then down through sons and daughters for the appropriate number of generations. Each sibling in a family is numbered, with distinctions made for male vs. female. Altogether, the codes are: M=mother, F=father, Bn=brother *n*, Sn=sister *n*, On=son *n*, An=daughter *n*. So the key MS101A1 would refer to my "mother's oldest sister's oldest son's oldest daughter". More correctly, it would refer to the mother's oldest sibling's oldest child's oldest child, but the codes do indicate the gender of siblings and children.

Having an egocentric key in place allows for convenient sorting of

the genealogical data for presentations such as family worksheets, indexes, and pedigree charts.

Here are some sample data lines:

FF	Langston, Ernest Albert	12APR1880	...
01NOV1937	Langston Cem, Riley's Cr NC	28DEC1911	...
FFB1	Langston, Claudius Samuel	21DEC1873	...
10AUG1883	Needham-potter Cem, Albtrsn	N	

BRIEF OVERVIEW OF THE ORIGINAL OUTPUT

The original program described in the SUGI 19 poster session was designed to produce a family worksheet for each family. A "family" was defined as a set of parents that produced a set of offspring. The parent who is/was blood-related to me is listed first, along with birth/death information and a page number referencing the family worksheet where this parent is listed as a child. The spouse is then listed after the blood-related parent. Then all children are listed, along with their birth/death/spouse information, along with page numbers where each child is listed as a parent.

Following all the worksheets is an index of all individuals (with women listed by both maiden names and married names) with the page numbers where they are listed as parents or children.

See figure 1 for an example of a family worksheet.

HTML DESIGN

Because of the page number linkage, it seemed straightforward to extend the individual worksheet data such that each worksheet became a separate HTML file. The page number references would become links to other pages, which would actually refer to other HTML files.

My intention was to produce a set of HTML files from the original output, with little change to the original program. Most all the information needed to produce the HTML files was provided in the original output, so modification of the original program would not be necessary.

I chose to use all uppercase names to simplify storage of the data on either PC-based or UNIX-based systems. I also chose to use the 3-character extension HTM for this reason. I was able to use the original page numbers of the worksheets to derive the names of each HTML file.

For example, if my grandfather Ernest Albert Langston appeared as a parent on page 12, and appears as a child on page 20, I could produce worksheet files in P012.HTM and P020.HTM, with links in each HTML file that refers to the other. The index file would consist of links back to all the other pages, and would be called INDX.HTM.

I also wanted to use "good HTML" so that a majority of web browsers could handle the HTML files correctly. This meant that I used the following design for all my HTML files:

```
<HTML>
<HEAD>
<TITLE>title of HTML file</TITLE>
</HEAD>
<BODY>
```

```
<H1>repeated title</H1>
...
</BODY>
</HTML>
```

My intention was also to limit the amount of text on a single line so that if a person wanted to examine the HTML file, it would be reasonably readable to a human.

Note also that I placed carriage returns and line feeds (hex 0D and hex 0A to simplify viewing of the HTML files on PC platforms.

I also wanted to avoid any of the specialized HTML that is only recognizable by certain browsers. My intention was to ensure that even those users confined to a line-oriented browser such as *lynx* would be able to successfully navigate the pages.

See Figure 2 for the browser output of the HTML version of the sample worksheet in Figure 1. Figure 3 provides the HTML text that corresponds to the output in Figure 2.

INCLUDING PHOTOGRAPHS

With the migration to HTML, I realized that I would have a good opportunity to introduce photographic images into my presentation of genealogical data. HTML has built-in capabilities to incorporate images via its IMG link syntax.

I had acquired many photographs over time and had scanned many of these into JPEG format, so referencing them inside of HTML files would be a very simple matter.

However, the bigger problem with incorporating photographs was to ensure that they were properly associated with the right people. This meant that I could not simply provide a text stream that described the photograph. Instead, I needed to provide a standardized list of people in the photograph so that their images would appear in the proper HTML files.

I chose to continue using the egocentric key system to properly identify the individuals in photographs. I had considered using full names, but realized that there were several people in my database whose name was identical to others in the database. It was less confusing to use keys instead, although there was some burden involved in locating the keys and ensuring they were entered properly.

To simplify the association of images with a description of the images, I chose a naming convention for the images and descriptions. All images are JPEG files, so image names are of the form *PICnnnnn.JPG*, with *nnnnn* being an arbitrary number. A single text file, *desc.txt* is used to store the descriptions of the images, along with the list of people pictured in the image. An example entry is as follows:

```
IMAGE NAME: PIC00088.JPG
DATE:      ca. 1907
PICTURED:  LEFT TO RIGHT: MFS1, MFB4, MFF, MF,
           MFS3, MFB5, MFM, MFB2
LOCATION:    Probably Point Caswell, NC
INFO:
```

This photograph is of the James William Flynn family. his was obtained from the Malpass family, related to James William Flynn through his sister. We determined the date of this photograph based on the apparent ages of the younger children with the knowledge of their birthdates.

The HTML file created from this text would be called

PIC00088.HTM, to correspond to the JPEG file name. See Figure 5 for the browser output of this HTML file. Note that the names are substituted for the codes in the final output. Note also that the image does not appear in the browser output, but instead is referenced via a link.

LINKING PHOTOGRAPHS

My intention was to ensure that if one wanted to view all the images associated with an individual, that all the images would not appear on the same Web page. This is because the download time for images can be a problem for those connected at slower speeds. Instead, I provide a single HTML file for each individual that has associated images, with links within the HTML file to the corresponding images. The description of the image appears, but the viewer must click on the link to actually view the image.

These HTML files are named *HPnnnnnn.HTM* (which stands for "Has Picture") and are arbitrarily numbered. They are linked to the individual in all worksheets on which the individual is referenced. Refer back to Figure 2 for the browser output. Note that the "PICTURES" link appears in the parent section. This is how the link is provided where the individual appears as a parent on a worksheet. Now note the highlighting of the birth number for several of the children listed on the worksheet. This is how the link is provided when an individual appears as a child on a worksheet.

See Figure 4 for the browser output of a sample *HPnnnnnn.HTM* file.

These links are not the most aesthetically pleasing, but my desire was to retain the original look of the text output. I may redesign the worksheet layout in the future to better accommodate the picture links.

DESCRIPTION OF THE SAS PROGRAM

The SAS program is designed to read from the worksheet listing produced by the earlier SAS program documented in previous articles. The earlier SAS program was altered slightly to provide an additional file that indicates an egocentric key based on the line of worksheet text. The individual's name is also supplied in this additional file. With these additional data values, the new SAS program can link up the picture texts with the proper names.

The SAS program produces formats (via the *CNTLIN=* option) that allow for convenient translation of worksheet position into keys and from keys into names. The picture description file is read and a *PICnnnnn.HTM* file is generated for each picture. A SAS data set is created that provides the keys for all people in each picture. This is matched up with the worksheet text so that when each *Pnnnnnnn.HTM* file is written out, the links for *HPnnnnnn.HTM* files can be added. The links to other worksheets is easy, since the page numbers are already provided in the worksheet text, and the *nnnnnnn* of *Pnnnnnnn.HTM* corresponds exactly to a page number. The *HPnnnnnn.HTM* are produced by re-sorting the list of individuals by key and producing one *HPnnnnnn.HTM* file for each individual. The information for each appropriate image is provided in each *HPnnnnnn.HTM* file.

The SAS program is available at the Web site whose URL is provided in the INTERNET LOCATION section below.

FUTURE PLANS

Because of the fact that the HTML files can be used in conjunction with browser software to produce camera-ready

images, it is my intention to eventually create a publishable document containing the worksheets, index, and photographs. And because everything is linked together, the HTML files and images can as well be placed on a CD-ROM to be distributed to family members, libraries, and genealogical archives. This is also in my long-range plan for disseminating my research.

INTERNET LOCATION

In the meantime, the entire set of HTML files and images currently reside on the Internet and are available to anyone with Internet access and a browser. As is always the case with Web addresses, this URL is subject to change. And because a genealogy is constantly being updated, the contents are certainly changing all the time.

The URL for accessing the starting page is

<http://www.webworldinc.com/langston>

Note that the SAS program discussed in this paper, along with a PostScript version of the paper itself, can be found on this Web site. Check the above URL for a link to the proper files.

CONCLUSION

HTML is well-suited for the presentation of genealogical data. And the SAS System is equally well-suited for managing the data. Allowing the SAS System to generate HTML files for presentation provides the combination of these capabilities that result in a unique yet extendable process.

BIBLIOGRAPHY

Langston, Richard D. (1994) "A Genealogical Data Management Program Using the SAS System," *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., 914-923.

SAS and SAS/AF are registered trademarks of SAS Institute Inc. in the USA and other countries. Other brand and product names are registered trademarks or trademarks of their respective companies.

ACKNOWLEDGMENTS

I would like to thank Ben Conner of Web World Inc. for providing the web space for my genealogy project.

```
<HTML>
<HEAD>
<TITLE>Rick Langston Genealogy - Page 1 </TITLE>
</HEAD>
<BODY>
<H1>Rick Langston Genealogy - Page 1 </H1>
<P>
Click <A HREF="INDX.HTM">here</A> to go to INDEX.
<BR><PRE>
FAMILY MEMBER: Langston, Ernest Albert      PAGE: <A HREF="P0000012.HTM"> 12 </A> <A HREF="HP0000004.HTM">(PICTURES)</A>
SPOUSE:        Scott, Julia Bertha          PAGE: <A HREF="P0000024.HTM"> 24 </A> <A HREF="HP0000005.HTM">(PICTURES)</A>
NAME           BIRTH                        SPOUSE      MARRIAGE  DEATH    PAGE    AGE
```

```
1. <A HREF="P0000002.HTM">Langston, Elgin Frederick</A>      03DEC1912 Helen Brice
2. Langston, Oris Elizabeth      17OCT1914 .N      23JUN1916
3. <A HREF="P0000085.HTM">Langston, Fannie Allene</A>      23JUN1916 Bernard Lewis
4. <A HREF="P0000087.HTM">Langston, Edna Kathleen</A>      14OCT1918 Oscar Lewis
<A HREF="HP0000003.HTM"> 5.</A> <A HREF="P0000007.HTM">Langston, Gerald Ross</A>
      .      Udell Putrell
<A HREF="HP0000001.HTM"> 6.</A> <A HREF="P0000092.HTM">LANGSTON, EDWARD ADOLPH</A>
<A HREF="HP0000024.HTM"> 7.</A> <A HREF="P0000090.HTM">Langston, Ethel Mae</A>
8. <A HREF="P0000009.HTM">Langston, Ernest Albert</A>      27MAY1925 Hazel Howard
```

07MAR1937	22DEC1978	2	66 YRS	0 MOS	19 DAS	
1 YRS	8 MOS	6 DAS				
18MAY1935	01MAR1960	85	43 YRS	8 MOS	7 DAS	
22MAY1937	10NOV1976	87	58 YRS	0 MOS	27 DAS	
03APR1920	Kate Savage		20SEP1941	29MAR1991		7 70 YRS
14JAN1923	HAZEL FLYNN		26JUN1943	22AUG1986		92 63 YRS
14APR1924	Sinclair Mitchell		16DEC1944			90
01SEP1991		9	66 YRS	3 MOS	5 DAS	

GENERATION: 1

RELATIONSHIP OF 1-8 TO RICHARD D. LANGSTON: PARENT/aunts & uncles

```
</PRE>
</BODY>
</HTML>
```

FIGURE 1: SAMPLE WORKSHEET

FAMILY MEMBER: Langston, Ernest Albert PAGE: 12
 BORN: 12APR1888 BIRTHPLACE: Seven Springs, NC
 DIED: 01NOV1957 BURIED: Langston Cem, Riley's Cr NC
 AGE: 57 YRS 6 MOS 20 DAS

SPOUSE: Scott, Julia Bertha PAGE: 24
 BORN: 24DEC1885 BIRTHPLACE: Riley's Cr, NC
 DIED: 21OCT1927 BURIED: Riley's Cr Ch Cem
 AGE: 41 YRS 9 MOS 27 DAS

MARRIED: 28DEC1911

CHILDREN							
NAME	BIRTH	SPOUSE	MARRIAGE	DEATH	PAGE	AGE	
1. Langston, Elgin Frederick	08DEC1912	Helen Brice	07MAR1937	22DEC1978	2	66 YRS 0 MOS 19 DAS	
2. Langston, Oris Elizabeth	17OCT1914	"	"	23JUN1916	NI	1 YRS 9 MOS 6 DAS	
3. Langston, Fannie Allene	23JUN1916	Bernard Lewis	18NOV1935	01MAR1960	88	43 YRS 8 MOS 7 DAS	
4. Langston, Edna Kathleen	14OCT1918	Oscar Lewis	22MAY1937	10NOV1976	87	58 YRS 0 MOS 27 DAS	
5. Langston, Gerald Ross	03APR1920	Kate Savage	20SEP1941	29MAR1991	7	70 YRS 11 MOS 26 DAS	
6. LANGSTON, EDWARD ADOLPH	14JAN1925	HAZEL FLYNN	26JUN1945	22AUG1986	92	63 YRS 7 MOS 8 DAS	
7. Langston, Ethel Mae	14APR1926	Sinclair Mitchell	16DEC1946	"	90		
8. Langston, Ernest Albert	27MAY1925	Hazel Howard	"	01SEP1991	9	66 YRS 3 MOS 5 DAS	

GENERATION: 1
 RELATIONSHIP OF 1-8 TO RICHARD D. LANGSTON: PARENT/aunts & uncles

FIGURE 4: BROWSER OUTPUT FOR PICTURE LIST

Rick Langston Genealogy – Image List for ALMEDA BELL LARKINS FLYNN

Click [here](#)¹ for image.

IMAGE DATE: ca. 1950

LOCATION: Her house?

INFO: several great-grandchildren also present; I will try to identify them at some point

Click [here](#)² for image.

IMAGE DATE: ca. 1950

LOCATION: Atkinson, NC

INFO: This photo was taken at Almeda's homeplace.

Click [here](#)³ for image.

IMAGE DATE: ca. 1907

LOCATION: Probably Point Caswell, NC

FIGURE 2: BROWSER OUTPUT OF SAMPLE WORKSHEET

Rick Langston Genealogy – Page 1

Rick Langston Genealogy – Page 1

Click [here](#) to go to INDEX.

FAMILY MEMBER: Langston, Ernest Albert		PAGE: 12 (PICTURES)				
SPOUSE: Scott, Julia Bertha		PAGE: 24 (PICTURES)				
NAME	BIRTH	SPOUSE	MARRIAGE	DEATH	PAGE	AGE
1. Langston, Elgin Frederick	03DEC1912	Helen Brice	07MAR1937	22DEC1978	2	66 YRS 0 MOS 19 DAS
2. Langston, Oris Elizabeth	17OCT1914	"	"	23JUN1916	NI	1 YRS 9 MOS 6 DAS
3. Langston, Fannie Allene	23JUN1916	Bernard Lewis	18NOV1935	01MAR1960	88	43 YRS 8 MOS 7 DAS
4. Langston, Edna Kathleen	14OCT1918	Oscar Lewis	22MAY1937	10NOV1976	87	58 YRS 0 MOS 27 DAS
5. Langston, Gerald Ross	03APR1920	Kate Savage	20SEP1941	29MAR1991	7	70 YRS 11 MOS 26 DAS
		Udell Putwell	"	NI		
6. LANGSTON, EDWARD ADOLPH	14JAN1923	HAZEL FLYNN	26JUN1945	22AUG1986	92	63 YRS 7 MOS 8 DAS
7. Langston, Ethel Mae	14APR1926	Sinclair Mitchell	16DEC1946	"	90	
8. Langston, Ernest Albert	27MAY1925	Hazel Howard	"	01SEP1991	9	66 YRS 3 MOS 5 DAS

GENERATION: 1
 RELATIONSHIP OF 1-8 TO RICHARD D. LANGSTON: PARENT/aunts & uncles

FIGURE 5: BROWSER OUTPUT FOR PICTURE LIST

Rick Langston Genealogy – Image 88



IMAGE DATE: ca. 1907

LOCATION: Probably Point Caswell, NC

PICTURED: LEFT TO RIGHT: Leola Bell Flynn Rivenbark, Roderick Lennox Flynn, James William Flynn, George Dallas Flynn, Winifred Elizabeth Flynn Stokely, Paul Augustus Flynn, ALMEDA BELL LARKINS FLYNN, James William Flynn