

# The ENRICH Project and Non-Standard Characters

ENRICH has done its best to adhere to recommended Unicode practice and use Unicode character sets internally for its work. As most projects use Unicode whenever possible these days, to do otherwise is not best practice and may indicate work that is unlikely to be funded. However, there are many completely valid instances where for using characters that do not appear yet in Unicode or will never appear in Unicode because they go against Unicode principles. For example, precomposed characters of convenience used when studying a specific scribal variance. The ENRICH recommendation is to use TEI methods of description of these characters to preserve information about their standardization and reasons for use.

## **1. The ENRICH gBank and the Medieval Unicode Font Initiative**

To provide a usable service for the ENRICH project, and potentially the others, the project decided in its investigation of Unicode to create a gBank. This is named both for the <g> elements which might reference these character descriptions, but also the TEI's **Gaiji** module for non-standard characters which contains these elements.

The ENRICH Gaiji Bank has obviously benefited from the [TEI P5 Guidelines](#) upon which it is based. But less obvious is that the initial definitions have come from the [Medieval Unicode Font Initiative](#) and the work of Odd Einar Haugen, Andreas Stötzner, Alec McAllister amongst many others. The graphic files that accompany every character description are generated from the MUF1-compliant version 3 of Andreas Stötzner's [Andron Scriptor Web font](#).

## **2. ENRICH gBank in the Manuscriptorium System**

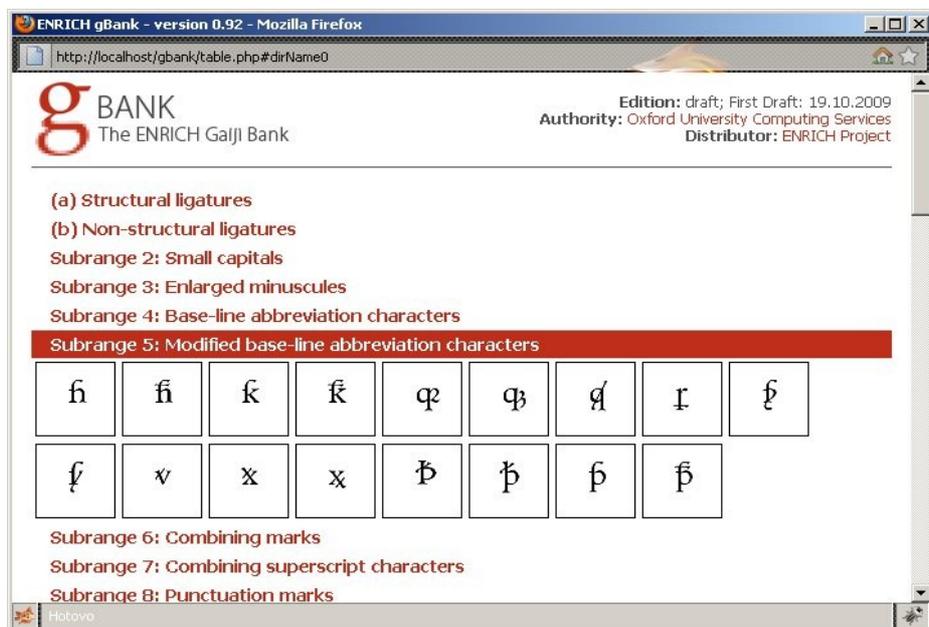
There are five applications of gBank in the ENRICH Manuscriptorium system. Each provides different service in order to enable both end-users and content authors to work efficiently with documents that require usage of special characters and glyphs, often not supported by Unicode.

The gBank is used:

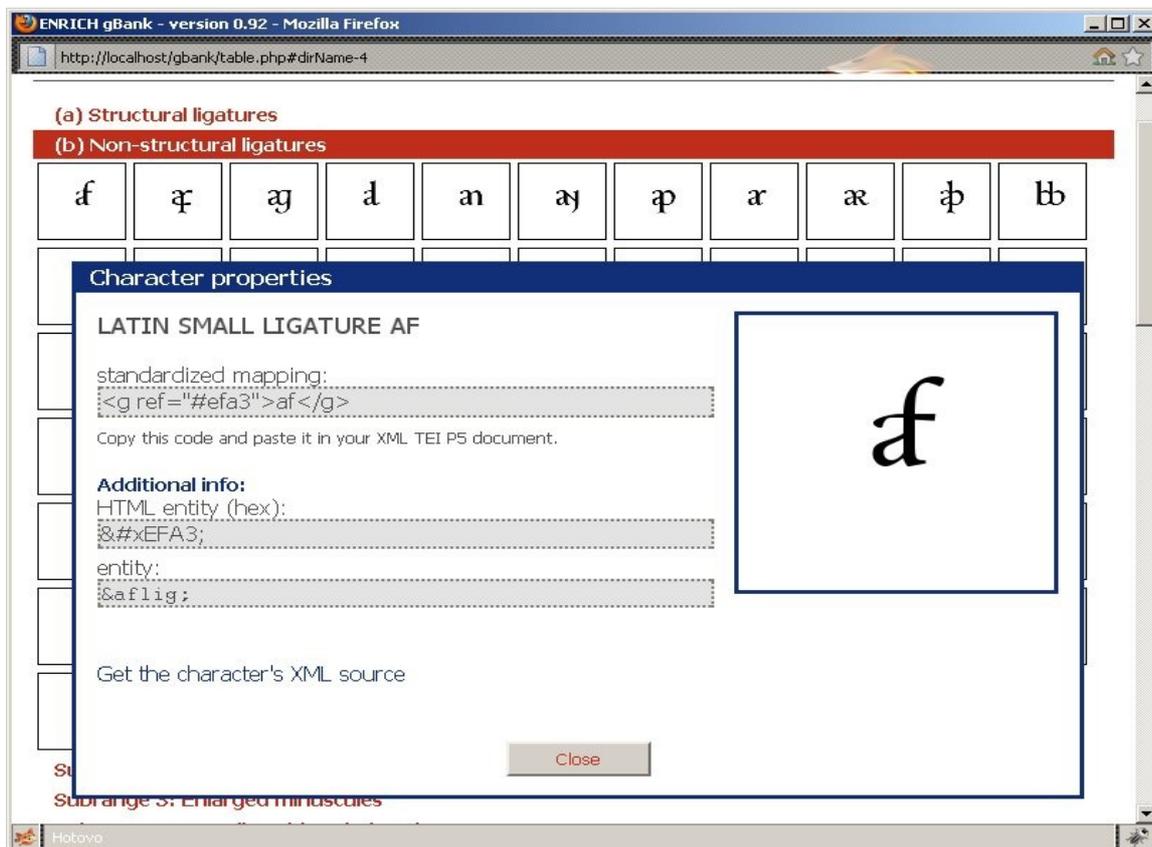
1. as a database for the newly created standalone gBank end-user interface
2. as a database for the newly created standalone gBank API interface
3. to internally enhance indexing routines and the search and retrieval system
4. to enhance the presentation layer through display of the characters covered by gBank
5. to enhance searching texts that originated without using gBank and Gaiji module

### 3. gBank End-User Interface

The standalone online application is now available for use at <http://beta.manuscriptorium.com/apps/gbank>. The application presents the content of the gBank database to end-users interested in finding a particular non-standard character. The user friendly interface displays characters ordered into sets, which can be further searched in order to find individual characters. As there are images available (for almost all of the characters) it is fairly straightforward to find a particular character. If an image is not available, a short description is displayed as a label.



The user can display a particular character description by clicking on the appropriate image. For end-users' convenience a valid XML code is displayed - this code can be copied and pasted directly in the XML metadata of the particular digital document.



The <g> element then fully substitutes the special character, the original information is transferred into the XML records without any information loss.

```
<titleStmt>
  <title>A sample title with a special <g ref="#eec6">af <>
  character</title>
</titleStmt>
```

The subsequent processing of the information represented by the <g> element is ensured for instance by the gBank API interface.

#### **4. gBank API Interface**

The gBank API interface performs one important task: it returns properties of a selected characters based on request passing character's ID. The format of the request is as follows:

<http://beta.manuscriptorium.com/apps/gbank/char.php?id=eec6>

The value of id parameter identifies the particular character. The API then returns the full character description as seen below:

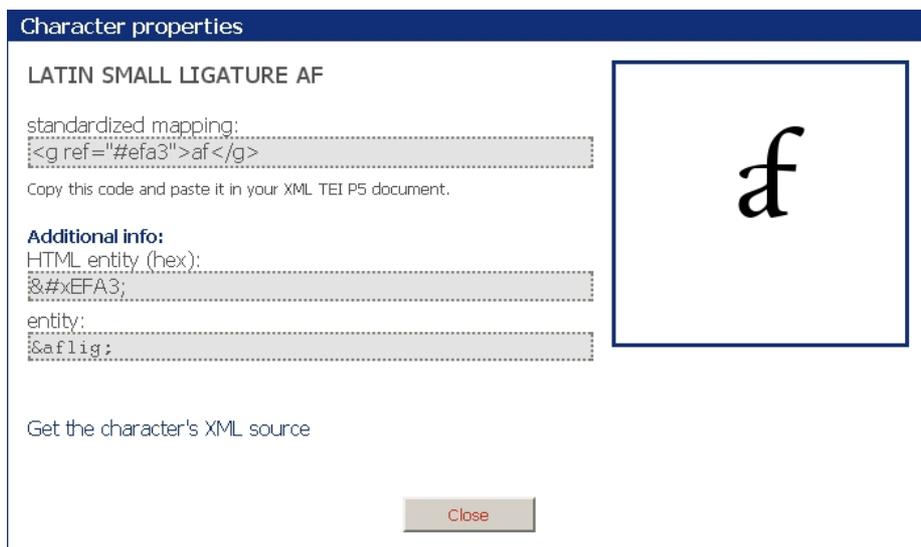
```
<char xml:id="eg-eec6">
  <desc>LATIN SMALL LIGATURE D ROTUNDA D ROTUNDA</desc>
  <charProp>
    <localName>entity</localName>
    <value>drottdrotlig</value>
  </charProp>
  <mapping type="MUFI" subtype="Unicode">□</mapping>
  <mapping type="MUFI" subtype="PUA">U+EEC6</mapping>
  <mapping type="standardized">dd</mapping>
  <graphic url="images/eec6.png"/>
</char>
```

Any application processing or retrieving the information can then again substitute the <g> elements with an appropriate information available in the <char> element (e.g. use the graphic, alternate mappings etc.)

## 5. Indexing and Searching with the gBank

### 5.1. Indexing with gBank Characters

The gBank database is also used during indexing routines within the ENRICH Manuscriptorium system. The metadata in ENRICH Manuscriptorium is indexed into a special database which enables efficient on-line searching. As it is difficult to include the non-Unicode characters into the search database, the gBank is used to substitute the <g> element with the standard mappings. For instance supposing we have following information for LATIN SMALL LIGATURE A F:



Character properties

LATIN SMALL LIGATURE AF

standardized mapping:  
<g ref="#efa3">af</g>

Copy this code and paste it in your XML TEI P5 document.

**Additional info:**  
HTML entity (hex):  
&#xEFA3;  
entity:  
aflig;

Get the character's XML source

Close

```
<char xml:id="eg-efa3">
  <desc>LATIN SMALL LIGATURE AF</desc>
  <charProp>
    <localName>entity</localName>
    <value>aflig</value>
  </charProp>
  <mapping type="MUFI" subtype="Unicode">□</mapping>
  <mapping type="MUFI" subtype="PUA">U+EFA3</mapping>
  <mapping type="standardized">af</mapping>
  <graphic url="images/efa3.png"/>
</char>
```

We can then use the standardized mapping and replace the character  $\text{af}$  with 'af' in the indexes.

### 5.2. Searching with gBank Characters

Users can do the same when they build their search query: they can simply use the standardized character mappings instead of the original character. This approach is very important for end-users, because majority of the special characters - or even those covered by Unicode - are difficult to enter into search queries using common keyboards and fonts.

Therefore as a result of our analysis and tests we provide standardized mappings to basic ASCII for each character in the gBank suite, using the principle that they should be able to be easily typed by a common keyboard and are covered all common fonts. The only exceptions to these are the medieval thorn 'þ' and eth 'ð' characters which are included

because there are no easy basic ASCII transliteration and as they were present in extended ASCII they are present in most fonts. In all standardization any ligatures have been decomposed into their component parts and any combining non-alphabetic modifiers (accents, etc.) have been removed. All combining alphabetic characters have been decomposed as separate characters.

For instance considering the example above: having *af* with 'af' standardized mappings then the user can simply enter 'abca*af*def' string into the query line and as a result not only abca*af*def will be found, but also the 'abca*af* def' will be found too. Of course, the search result would be wider if using this approach, but the number of overabundant records will not be significant because of the limited size of the gBank database.

These features are implemented into ENRICH Manuscriptorium, but have not yet been extensively tested with real world examples.

### **5.3. Advanced Search Features Using gBank Characters**

There are many metadata sources that do not use TEI P5 to create their primary metadata. Therefore they do not use the <g>element within their markup or provide descriptions as <char> elements (e.g. MARC, or other format sources including those retroconverted from print sources). However, even for these sources the gBank suite can be used to make search tasks more efficient. In most cases these sources use their own standardized mappings to transcribe special characters not available on a standard keyboard. When aggregating documents from a particular source its mappings can be analyzed and their individual standardized mappings can be linked with gBank standardized mappings. This way a table of equivalences in variation can be created and placed within the search system. This has already been implemented and is in use in the ENRICH Manuscriptorium system.

As a result of this work all end-users querying the ENRICH Manuscriptorium using gBank standardized character mappings will be able to retrieve documents from sources using different mapping approaches (and vice versa).

## 6. Support of gBank in the Presentation Layer

The final but significant task when implementing gBank into Manuscriptorium was to analyze and prepare the rendering and presentation of texts and descriptions within the end-users interface.

### 6.1. Use of Images and Standardized Mappings

The system again uses the incorporated gBank database and replaces the <g> element either with an image (preferred) or standard mappings (where images are not available). So the user can read the information in the most natural and comfortable way. This approach is implemented and working in the ENRICH Manuscriptorium system.

### 6.2. Use of TTF and CSS 3

In rendering non-standard characters in the end-user's browser, there is the possibility to use a dedicated TrueType Font (TTF) which is capable to display the texts in combination with CSS 3.

In creating images for display, all the graphic files with examples characters are based on the MUFI-compliant version 3 of Andreas Stötzner's [Andron Scriptor Web](#) font. They were converted from TTF to SVG using Apache Batik and then from SVG to PNG for online display.

Additionally, the CSS 3 recommendation enables the use of web font rather than a font residing on the end-user's computer using the @font-face rule:

```
<style type="text/css"> @font-face { font-family: AndronScriptorWeb; src: url('Andron Scriptor Web.ttf'); } .mufiFont { font-family: AndronScriptorWeb; } </style>
```

This way it would be possible to display texts using dedicated fonts that supply the correct characters for the Unicode Private Use Area. Unfortunately the browser support for TTF web fonts currently is rather low (supported by: Mozilla Firefox 3.5+, Opera 10, Safari 3.1, Safari 4; not supported by :IE (all versions - support only Embedded Open Type), Opera 9, Google Chrome 3.0, Mozilla Firefox 2, Mozilla Firefox 3.0.) Therefore this way of usage is not recommended at present, but as more browsers support TTF web fonts, then the suggested CSS3 approach could be successfully applied.

Another possible way of using TTF would be to let the users to install dedicated font into their systems. To check whether the font is installed a javascript detection of the font's presence in the system could be implemented. This would help to decide whether use images or dedicated font during the presentation.

**Note:** This approach is tested and ready for implementation if the end-users decide that they require it. However, bearing all this in mind, the use of standardized mappings as described above is the preferred recommendation at the moment.