Journal of the Text Encoding Initiative

Issue 12 | July 2019 - May 2020
Selected Papers from the 2017 TEI Conference

Encoding Disappearing Characters: The Case of Twentieth-Century Japanese-Canadian Names

Stewart Arneil

Electronic version
URL: http://journals.openedition.org/jtei/2301
DOI: 10.4000/jtei.2301
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Stewart Arneil, « Encoding Disappearing Characters: The Case of Twentieth-Century Japanese-Canadian Names », *Journal of the Text Encoding Initiative* [Online], Issue 12 | July 2019 - May 2020, Online since 26 August 2019, connection on 27 June 2020. URL : http://journals.openedition.org/jtei/2301 ; DOI : https://doi.org/10.4000/jtei.2301

Encoding Disappearing Characters: The Case of Twentieth-Century Japanese-Canadian Names

Stewart Arneil

ABSTRACT

The Landscapes of Injustice project seeks to encode mid-twentieth-century documents by and about the Japanese-Canadian community so they are accessible to modern audiences. The fundamental problem is that some of the kanji used at that time have been replaced since then by different kanji, and others have been removed from lists of formally acceptable characters. This report documents our efforts with two technologies designed to address this situation. The first is the Standardized Variation Sequence (SVS) feature of Unicode. Our work revealed that this set of variation sequences does not completely cover the old and new glyph pairs identified by the Japanese authorities, and that the pairs formally identified by the Japanese authorities do not completely cover all the new glyph forms in general use. We turned to TEI's `<charDecl>`, `<glyph>`, and `<mapping>` elements as a second technology to augment the support provided by Unicode. Lastly, we dealt with the issue of

finding suitably qualified people to do the markup. The result is markup which retains the original glyphs and relates them to the modern glyphs, so that in our output products we will be able to support search and display using either form of the glyph.

INDEX

Keywords: kanji, unicode, glyph, variant, Japanese

1. The Problem of Disappearing Japanese Characters

1 *The Landscapes of Injustice*[1] project seeks to integrate data from various sources (such as oral histories, court records, government minutes, land title documents, maps, community directories, and personal letters) to capture multiple perspectives on events affecting Canadians of Japanese descent in the 1940s and to create products based on that research for modern academic and public audiences. The Japanese-language documents (for example, community directories) used kanji (Chinese characters used in Japanese script) which were perfectly acceptable at the time, but which have since been superseded (either officially or practically) by other kanji glyphs. The project's concern with the changing forms of kanji over the twentieth century is primarily practical, rather than a scholarly focus.

2 In 1946 and 1981 the Japanese government specified simpler forms (known as *shinjitai* kanji) for certain characters and deprecated their older, traditional forms (known as *kyūjitai* kanji) for many purposes such as education and government publication (Agency for Cultural Affairs 2010). Although the kyūjitai kanji were not banned, the obsolete kyūjitai kanji have become unreadable to more and more readers over time, thus making texts including them difficult for modern readers, but at least there is a recognized mapping from new form to old form. In addition to the officially recognized shinjitai-kyūjitai pairs of kanji, there are other forms which are outside the lists of current kanji as identified by the Japanese government in Agency 2010. These *hyōgaiji* kanji may still appear (particularly in names), or in some cases may have counterpart modern forms. We have so far found just over 1,000 instances of what I call non-conventional kanji, consisting of just over 110 different shinjitai-kyūjitai pairs (some of which appear more than once in our documents) and about 5 hyōgaiji (all of which are single instances).

3 Our pre-1945 source documents include both classes of non-conventional kanji forms, particularly in personal names. Personal names are especially problematic as they are proper nouns, and as such the correct reading is dependent almost entirely on the characters in the name rather than grammatical or other context clues. The project is particularly sensitive to the representation of names as the community involved was largely erased as a community from Canadian society in the 1940s. Changes to the kanji thus risk the names of the individuals affected being "disappeared" from the historical record we are creating in a way which echoes the disappearance from history suffered by the actual community. More practically, people searching for specific names may not find the records they seek due to a mismatch of kanji, and similarly for people reading results who do not recognize a name.

2. Representing Disappeared Characters in Unicode

4 The project's focus is on the historical treatment of the Japanese Canadian community, and not the evolution of the Japanese language, so we sought the simplest solution that would meet our needs. Initial research suggested exploiting features in the Unicode character encoding standard.

5 Unicode has a remarkably complex treatment for mapping certain non-conventional to conventional kanji (Unicode Consortium 2018a, 23.4, 872–74), the full details of which are beyond the scope of this paper. It uses what are known as Standardized Variation Sequences (Unicode Consortium 2018b). Even the following simplified consideration raises problems with this approach for our situation.

6 We want to preserve the forms as found yet maintain an association with a conventional form where one exists. A Standardized Variation Sequence consists of one entity for the conventional form of the kanji (e.g., 社) followed immediately by one of several other entities (︀, ︁, and so on), yielding, for example, 社︀. Unicode also specifies lookup tables to map from the conventional form to the non-conventional form. Note that the non-conventional form is not explicitly encoded in the document, so this approach precludes an application normalizing a non-conventional form to a conventional one in inconsistent or unpredictable ways—which of course is helpful to us. However, we are still at the mercy of (1) font developers and the degree of support they have built in to their fonts for variation sequences, and (2) application developers and the extent to which the application tries to locate a font that

supports the variation sequence (Lunde 2015). For an example with visibly clear results, note in the following how the Firefox browser differs from Chrome and Safari in representing the Standardized Variation Sequence (in the last row). I use an image to display these text characters, because the very problems of inconsistent font and application support may otherwise corrupt which forms the reader sees:

Figure 1. How three browsers display five encodings for conventional and non-conventional forms.

Encoding	Firefox displays	Chrome displays	Safari displays
社 (kyūjitai)	社	社	社
社	社	社	社
社 (shinjitai)	社	社	社
社	社	社	社
社︀	社	社	社

The Firefox implementation is at the time of writing more sophisticated than the other browsers in that it can search for a font supporting the SVS and display the correct form; the other browsers require that a font supporting the SVS be specified.

7 Differences in support are apparent in searching, too. We searched the five encodings listed above for each of the two kanji. Chrome and Safari ignore the variant sequence and thus treat the two glyphs as interchangeable (whether searching for "社" or "社," all five instances of either character are found). That is generally the desired behavior for all but scholars of historical Japanese. Firefox pays attention to the variant sequence, but it also fails to normalize as it should, so when we searched for "社" we got no hits, but when we searched for "社" we got three hits, one of which was the Standardized Variant, which as just noted is displayed to the user as "社." These findings are summarized in table 1:

Table 1. Hits for five instances of conventional and non-conventional kanji in various browsers.

Search for	Firefox finds	Chrome finds	Safari finds
社	0	5	5
社	3	5	5

Beyond the specific details of our examples, the main problem is the inconsistency of support in applications and the difficulty of using Unicode Standardized Variation Sequences in processing environments.

8 Even if support for Standardized Variants were robust and consistent, it would be inadequate for our data set because few of the shinjitai-kyūjitai pairings and virtually no hyōgaiji forms we discovered in our data appear in the Standardized Variant list. As shown in table 2, under 20% of the non-conventional forms in our data appear on the Standardized Variants list, while over 80% do not.

Table 2. Frequency of three types of pairings of non-conventional and conventional kanji.

Description	# of pairs (%)	non-conventional kanji (code point)	conventional kanji (code point)
kyūjitai with shinjitai counterpart specified by Standardized Variation Sequence	22 (19%)	社 (FA4C)	社 (793E)
kyūjitai not unifiable with shinjitai, encoded in Unicode as separate CJK unified ideograph	91 (77%)	會 (6703)	会 (4F1A)
hyōgaiji (no Standardized Variant counterpart, but likely one in JIS standards, e.g., JIS X 0208)	5 (4%)	塲 (5872)	場 (5834)

With these results, we could not rely on the Standardized Variant approach. We turned to a more elaborate, explicit encoding that would cope with the classes of kanji forms described above and summarized in table 2 to make our intentions clear regardless of subsequent processing or display applications.

3. Representing Disappeared Characters in TEI

9 We were already using TEI to encode the documents, so we needed to find and implement TEI markup to capture the three classes of problematic kanji. Specifically, we employed the gaiji module's `<charDecl>`, `<g>`, `<glyph>`, and `<mapping>` elements to represent each non-conventional kanji, the conventional kanji associated with that non-conventional kanji (if one exists), and whether the mapping appears in the kyujitai-shinjitai list and/or the Standardized Variant list (TEI Consortium 2017, sec. 5.2).[2]

10 We created a TEI file named `chars.xml` consisting of a character declaration (`<charDecl>`) element which contains a `<glyph>` element for each non-conventional form (kyūjitai or hyōgaiji) to describe it and its conventional equivalent. Within each `<glyph>` element, we use a `<mapping>` element with a specific value for the `@type` attribute for each variant of the glyph. In the body of the data file, we use a `<g>` element to encode the kanji with an `@xml:id` attribute which points to the appropriate `<glyph>` element in the `chars.xml` file. This approach allows us to capture the three classes of pairs of non-conventional and conventional forms consistently, as shown in the following three examples (note that some characters may not display properly on some user agents).

11 Example of kyūjitai with shinjitai counterpart and in Unicode Standardized Variation Sequences:

12 In `chars.xml`:

```
<charDecl>
 <glyph xml:id="u793E">
  <mapping type="kyūjitai">社</mapping>
  <mapping type="shinjitai">社</mapping>
  <mapping type="uniStdVar">&#x793E;&#xFE00;</mapping>
 </glyph>
</charDecl>
```

In `data.xml`:

```
<body> ... <g ref="chars.xml#u793E">社</g> ... </body>
```

13 Example of kyūjitai with shinjitai counterpart, but not in Unicode Standardized Variation Sequences:

14 In `chars.xml`:

```
<charDecl>
 <glyph xml:id="u6703">
  <mapping type="kyūjitai">會</mapping>
  <mapping type="shinjitai">会</mapping>
 </glyph>
</charDecl>
```

In `data.xml`:

```
<body> ... <g ref="chars.xml#u6703">會</g> ... </body>
```

15 Example of hyōgaiji that does not appear in the kyūjitai-shinjitai list nor in Standardized Variation Sequences:

16 In `chars.xml`:

```
<charDecl>
 <glyph xml:id="u5834">
  <mapping type="hyōgaiji">塲</mapping>
  <mapping type="regularization">場</mapping>
 </glyph>
</charDecl>
```

In `data.xml`:

```
<body> ... <g ref="chars.xml#u5834">塲</g> ... </body>
```

17 The values we used for the `@type` attribute (`"kyūjitai"`, `"shinjitai"`, and `"hyōgaij"`) reflect our circumstances; for anyone not already familiar with the twentieth-century history of kanji, their meanings would be explained by a simple search for those terms in Wikipedia. The specific values we have used for the `@type` attribute may not be semantically accurate for other languages or other

eras of Japanese. However, the utility of the approach does not depend on those specific values, so it could easily be implemented using more appropriate values for the `@type` attribute tailored to the specific circumstances.

4. Training Encoders of Texts Containing Disappeared Characters

18 Having established a data model, we then turned to the job of applying that model to the relevant documents. There are three stages involved in this kind of markup: identify the kanji that are instances of kyūjitai or hyōgaiji, determine if the non-conventional form appears in the Standardized Variation Sequence, and associate the non-conventional form with a conventional form if possible. Within the context of a TEI encoding project, the required skill sets are knowledge of and facility with (1) what is to some degree arcane Japanese, especially for second-language users and those outside Japan; (2) the Unicode standard, especially Standardized Variants; and (3) TEI XML and specifically the elements described above.

19 An important aspect of the project is engagement with the Japanese-Canadian community and providing that community with a sense of editorial input, if not authorship, of the material. Clearly the most critical skill set is facility with the non-conventional kanji forms. In general, it is usually better to start with someone with subject matter expertise and train them in the technical and workflow skills. In our circumstance, and after substantial consultations with colleagues at our partner Japanese-Canadian museum, we concluded that the most suitable candidate to do the volume of work we required to an adequate level of competence would be a student who is reasonably fluent in Japanese, knowledgeable about the history, and technically competent. That person would focus on improving their facility with the various forms of kanji within the documents. This approach has proven workable given that our project's primary scholarly focus is not on the evolution of kanji, though it has approximately doubled the amount of time required to encode the document.

5. Conclusions

20 Our goal is to encode documents containing non-conventional forms of kanji so that all forms are available for processing and for use by human users. A potential solution based on Unicode Standardized Variation Sequences did not cover enough of the instances we encountered. Of the problematic forms in our data, the proportion of kyūjitai-shinjitai pairs was much lower than we expected, and the proportion of hyōgaiji much higher. We therefore decided to encode the variant glyphs explicitly, using the features provided in the gaiji module in TEI. This allowed us to specify type attributes to describe different classes of kanji forms and the Unicode Standardized Variant in our encoding of the document. It was difficult to find people with all the necessary skills to do this encoding. The best solution for us was to train an otherwise competent encoder of Japanese to recognize and accurately encode the non-conventional kanji.

21 We now have a robust and consistent encoding which covers all the instances in our data. The next phase of the project will focus on processing the TEI to represent the characters in output products for use by researchers and by the public. The project will produce not only web-based outputs, but also print-based and museum installations, and for these we will need to make careful editorial decisions about which kanji to use to balance our wish to honor the names (as they were at the time) of the people who suffered the injustices presented by the project, and our wish to ensure that those names (and the people they represent) do not disappear to modern readers.

BIBLIOGRAPHY

Agency for Cultural Affairs, Government of Japan. 2010. "Academic Index of Kanji Table." Accessed May 8, 2019. http://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kijun/naikaku/kanji/joyokanjisakuin/.

Jenkins, John H., Richard Cook, and Ken Lunde. 2018. "Unicode Han Database (UniHan)" (report). Unicode Standard Annex #38, revision 27, February 15. Mountain View: Unicode Consortium. https://www.unicode.org/reports/tr38/.

Lunde, Ken. 2015. "Exploring Typekit's New Dynamic Kits." *CJK Type Blog*, June 16. https://blogs.adobe.com/CCJKType/2015/06/typekit-dynamic-kits.html.

Paterson, Duncan. 2018. "I Just Want to Be Normal: Character Normalization between Unicode and TEI." Panel paper presented at the TEI Members' Meeting, Tokyo, Japan, September 11.

TEI Consortium. 2017. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.2.0. Last updated July 10*. N.p.: TEI Consortium. https://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/.

Unicode Consortium. 2018a. "Chapter 23: Special Areas and Format Characters." In *The Unicode® Standard Version 11.0 – Core Specification*, 855–90. Mountain View: Unicode Consortium. http://www.unicode.org/versions/Unicode11.0.0/ch23.pdf.

———. 2018b. "Specification of the Variation Sequences That Are Defined in the Unicode Standard," v. 11.0.0. February 2. https://unicode.org/Public/11.0.0/ucd/StandardizedVariants.txt.

NOTES

1 https://landscapesofinjustice.com.

2 Ken Lunde has pointed out that while it is straightforward to provide this kind of mapping in TEI, in fact the Unicode Consortium, through its Unihan Database, already has a mechanism for mapping equivalences such as these, and it would be worthwhile to propose updates to the Unihan Database for any mappings it does not yet handle. Coincidentally, at the TEI 2018 conference in Tokyo, Duncan Paterson proposed a new `<uniHan>` element for TEI, which would be a child of `<charProp>`, and whose content would be one of the Unihan Database properties (Jenkins, Cook, and Lunde 2018). So by using existing properties and proposing new ones where necessary, then capturing those properties through the `<uniHan>` element, these relationships could be efficiently encoded.

AUTHOR

STEWART ARNEIL

Stewart Arneil is a programmer/consultant at the Humanities Computing and Media Centre at the University of Victoria, Canada. He holds an MA in computational theory and certifications in instructional design and in project management. He has thirty years of experience in the private and public sectors managing academic projects and developing software, databases, and websites for research and educational purposes in collaboration with language and subject matter experts.