**GreenFIE: A Green Form-Based**

**Information-Extraction System for Historical Documents**

Tae Woo Kim, Computer Science Department, Brigham Young University

David W. Embley, Advanced Technology Research, FamilySearch

Stephen W. Liddle, Information Systems Department, Brigham Young University

**1. Introduction**

In his 2012 Keynote Address at the Family History Technology Workshop, George Nagy advocated the use of “Green” systems—systems that improve themselves as they are used [Nagy12a]. GreenFIE is a **Green** **F**orm-based **I**nformation-**E**xtraction system for semi-automatic form-filling of genealogical information extracted from family history documents.

GreenFIE is built on top of COMET, a tool that allows forms to be filled in with the information of interest from pages of OCRed historical documents. Starting from scratch, users fill in form fields as they click on items of interest. More typically, users check and correct forms that have already been filled in automatically by an ensemble of information-extraction tools. In our family-history application, forms are simple and are more like records—one for birth and death information, one for marriage information, and one for parent-child relationships. COMET maintains a collection of these record-like forms for each page. GreenFIE observes users as they work. If a user corrects a record or adds a new one, GreenFIE creates a new extraction rule that would have correctly extracted the record, generalizes it, executes it, and adds any new records it finds to the collection. GreenFIE can also execute the new extraction rule on subsequent pages in the user’s task, prepopulating additional form records in advance.

We are not aware of any other system like GreenFIE—a system designed to allow users to annotate OCRed historical documents by filling in forms while it simultaneously learns from user annotation work and shifts the burden of annotation as much as possible to itself. We are, however, aware of work on various aspects of GreenFIE: (1) systems that learn as users process documents (e.g. see [BloN12] and [Nagy 12b]); (2) semi-supervised information extraction, particularly OLERA [ChaK04] in which users enclose an information block of interest and then specify relevant information slots for each field in the record; and (3) extraction-rule monitoring and revision (e.g. see [LLC+10] and [SaJS10]).

**2. GreenFIE**

Figure 1 shows the COMET interface with some record fields on the left filled in for Page 31 of a transcript of the Kilbarchan Parish Record [Gr1912] on the right. As users fill-in form fields, they have the option of clicking on a *Regex* button on the right of each record (hidden in Figure 1, but accessible as a user moves to the right with the slider-bar at the bottom of the screen). When a user clicks on the *Regex* button, GreenFIE generates and generalizes a regular-expression extraction rule for the record and creates additional filled-in records.

In Figure 1, a user has entered the highlighted information into the first record and clicked on the *Regex* button. As a result GreenFIE created and filled in the following 14 records.

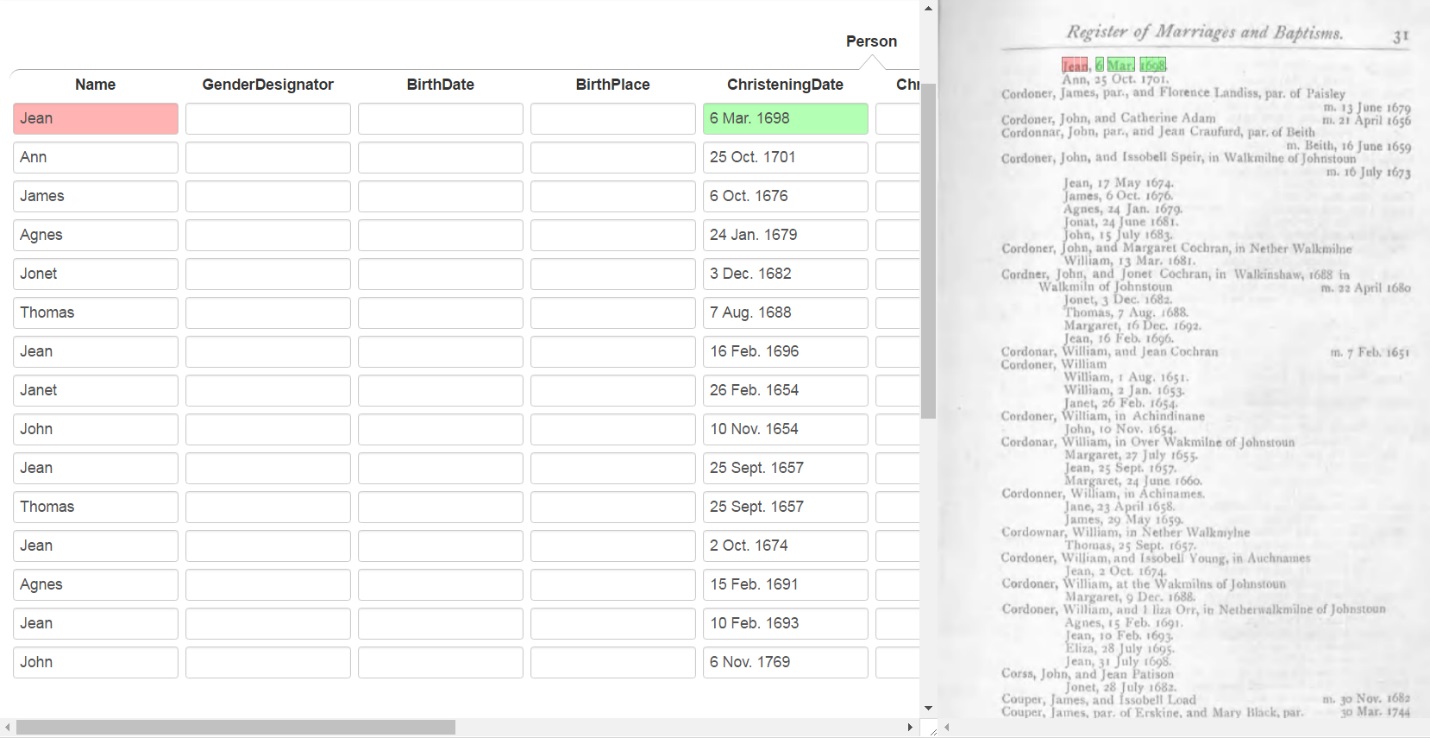


Figure 1. Display of Results after Annotation of “Jean, 6 Mar. 1698.” and Execution of GreenFIE’s Basic Generated Regular-Expression Extraction Rule.

For the highlighted text in Figure 1 (“Jean, 6 Mar. 1698.”) a regular expression matching the information is:

\n([A-Z]{1}[a-z]{3}),\s(\d{1}\s[A-Z]{1}[a-z]{2}\.\s\d{4})\.

When a regular expression matches a text string, the part of the matched string that corresponds to a parenthesized sub-regular expression is captured and associated with a capture-group number—numbered by counting opening parentheses reading left to right. GreenFIE then associates the capture-group number with the field name in the form, which after some internal processing causes the captured text to be displayed in the named form field. Capture-group 1, ([A-Z}{1}[a-z]{3}), causes “Jean” to be placed in the **Name** field in the record and Capture-group 2, (\d{1}\s[A-Z]{1}[a-z]{2}\.\s\d{4}), causes “6 Mar. 1698” to be placed in the **ChristeningDate** field.

To capture additional christening records on the page, GreenFIE generalizes the regular expression and executes it. GreenFIE’s basic regex generalizer for fields increases the span of quantifiers of [a-z] and \d by the ceiling of plus and minus 40%. Thus, the generalized rule that extracts the additional 14 records is:

\n([A-Z]{1}[a-z]{1,5}),\s(\d{0,2}\s[A-Z]{1}[a-z]{1,3}\.\s\d{2,6})\.

Because GreenFIE knows the type of each form field, it can take advantage of this knowledge to better generalize capture-group expressions. Generalizing given names and day-month-year dates to:

\n([A-Z]{1}[a-z]{1,10}),\s(\d{1,2}\s[JFMASOND][a-z]{1,5}\.?\s\d{4})\.

extracts all 32 of the christening records on Page 31. Instead of generalizing person names, dates, and place names with respect to the type-forms found in a document, it could draw type-specific regular expressions from a library (e.g. <http://regexlib.com>) or create them, relying, as needed, on extensive name dictionaries and gazetteers.

Not all pages in the Kilbarchan Parish record are as regular as Page 31. Page 32, for example, has some birth dates marked with the word “born” in place of christening dates. Other pages include twins on the same line with one birth or christening date. OCR errors appear occasionally; some are regular such as commas appearing as periods and vice versa, but others are unique, requiring human correction and rendering them useless as patterns for extraction rules.

When GreenFIE applies the above generalized regular-expression extraction rule to Page 32, it immediately extracts 24 of the 31 records. When a user annotates the first non-captured record, “Robert, born 29 July 1756.”, GreenFIE generates:

\n([A-Z]{1}[a-z]{1,10}),\sborn\s(\d{1,2}\s[JFMASOND][a-z]{1,5}\.?\s\d{4})\.

which captures the remaining seven records.

**3. Experimental Results**

Starting from scratch, we ran GreenFIE on Pages 31 and 32 using GreenFIE’s naïve, basic rule for generalizing regular expressions. The graph in Figure 2 shows the results.

Figure 2. Graph of Running GreenFIE with Naïve Generalization

for Person Records on Kilbarchan, Pages 31 and 32.

The shape of the graph is typical for GreenFIE execution. Iteratively, generating rules from user annotations and executing them eventually captures all records on a page—in Figure 2, after five iterations. Execution of the rules generated for the first page on the second page may already capture many of its records, and subsequent iterations eventually capture them all.

How quickly all records are captured depends on the regularity of the records and on GreenFIE’s ability to generalize well. As described in Section 2, when GreenFIE generalizes names and dates with respect to types, 100% recall is achieved on Page 31 after just one iteration and is achieved for both Pages 31 and 32 after a second iteration.

The marriage records on Page 31 in Figure 1 are not nearly as uniform as the birth and christening records. Figure 3 shows the results of successively annotating, generating, and executing extraction rules for couple records, which each contain a person and spouse name and optionally a date of the marriage or proclamation of the marriage and place when it is not in the Kilbarchan Parish.

Figure 3. Graph of Running GreenFIE for Couple Records on Kilbarchan, Pages 31 and 32.

To complete the extraction of the 16 couple records, a user must annotate 13 of them. The variation in the records is considerable, despite all of them being generally formatted as “<husband> … and <wife> … <optional place> <date>”. As a basic variation, the optional place and date are marked either by “m.” (marriage date) or by “p.” (proclamation of banns date). The most prominent variation is the commentary denoted here by “…”, which is whatever the parish priest wanted to write, most typically indicating the parish of residence when it’s not Kilbarchan. In addition, OCR errors and punctuation typesetting inconsistencies also account for a fair amount of variation. Observe, however, that executing the 13 extraction rules obtained from Page 31 on Page 32 immediately extracts many of the couple records on Page 32—15 of 23 to be exact. The high variability continues, requiring 7 new annotations to capture the remaining 8 records.

**4. Conclusion**

GreenFIE is indeed “green”; it self-improves with use. When records in documents are highly regular (as are the Kilbarchan christening and birth records), GreenFIE can provide a many-fold increase in productivity. When the records are highly variable (as are the Kilbarchan marriage records), GreenFIE can still provide an increase in productivity.

For highly variable records, attempts to have GreenFIE generalize to account for more of the variability have so far led to a loss of precision—extracted records with incorrect information. In these cases, COMET users can correct or altogether discard these incorrectly extracted records. As future work, we intend to continue to explore ways to better generalize extraction rules, while at the same time to avoid over-generalizations as much as possible and to have GreenFIE correct or retract over-generalized extraction rules as it observes users rejecting or correcting erroneously extracted records.

**References**

[BloN12] D. Blostein and G. Nagy, Asymptotic Cost in Document Conversion, *Proceedings of the 19th Conference on Document Recognition and Retrieval* (DRR’12), San Francisco, California, USA, 24–26 January, 2012.

[ChaK04] C.-H. Chang and S.-C. Kuo, OLERA: Semisupervised Web-Data Extraction with Visual Support, *IEEE Intelligent Systems*, 19(6):56–64, November/December, 2004.

[LLC+10] B.Liu, L. Chiticariu, V. Chu, H.V. Jagadish, and F. R. Reiss, Automatic Rule Reﬁnement for Information Extraction, *Proceedings of the VLDB Endowment* (PVLDB), 588–597, September, 2010.

[Nagy12a] G. Nagy, Keynote Address: Back to the Future, *Family History Technology Workshop*, Salt Lake City, Utah, February, 2012.

[Nagy12b] G. Nagy, Estimation, Learning, and Adaptation: Systems that Improve with Use, *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, Hiroshima, Japan, 1–10, November, 2012.

[Gr1912] F.J. Grant (ed.), *The Register of Marriages and Baptisms in the PARISH OF KILBARCHAN* *1649–1772*, Edinburgh, Scotland, J. Skinner & Company, LTD., 1912.

[SaJS10] A.D. Sarma , A. Jain , D. Srivastava, I4E: Interactive Investigation of Iterative Information Extraction, *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, Indianapolis, Indiana, 795–806, June 6-10, 2010.