**GreenFIE: A Green Form-Based**

**Information-Extraction System for Historical Documents**

**(Extended Abstract)**

Tae Woo Kim, Computer Science Department, Brigham Young University

David W. Embley, Advanced Technology Research, FamilySearch

**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 [Nagy12]. 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is built on top of COMET, a tool that allows forms to be filled in with the information of interest from pages of OCRed historical documents.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 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.

**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). 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, the 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. 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})\.

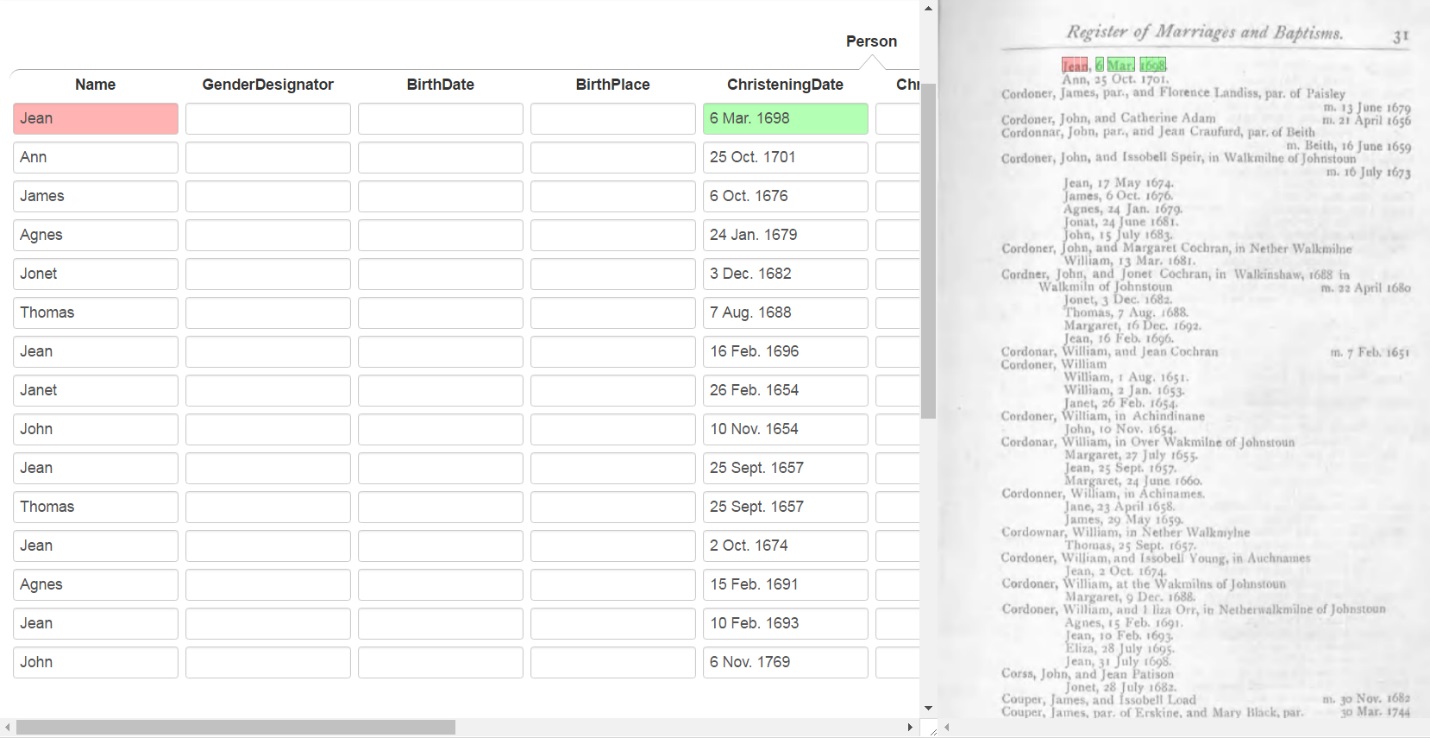
where Capture-group 1 is the **Name** in the record and Capture-group 2 is the **ChristeningDate**. GreenFIE’s basic regex generalizer for fields increases the span of quantifiers of [a-z] and \d by the ceiling of plus and minus 40%. Thus, the generalized rule that extracts the additional 14 records 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})\.

Knowing the type of field allows us to better generalize extraction expressions. Generalizing given names and day-month-year dates to:

\n([A-Z]{1}[a-z]{1,15}),\s(\d{1,2}\s(January|Jan\.?| … |Dec.\?)\s\d{4})\.

extracts all 32 of the christening records on Page 31. 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. And 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.



**3. Experimental Results**

Starting from scratch, we independently processed Pages 31 and 32 in the Kilbarchan Parish Record by iteratively filling in the fields in a record with the information for the first not-yet-extracted name and christening or birth date. Table 1 shows the results.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1. Recall and Precision by GreenFIE Number of Extraction Rules Generated (%) | | | | | | | |
| Rule Generation Nr. | | 1 | 2 | 3 | 4 | 5 | 6 |
| Page 31 | Recall | 47 | 72 | 91 | 97 | 100 |  |
|  | Precision | 100 | 100 | 100 | 100 | 100 |  |
| Page 32 | Recall | 6 | 45 | 58 | 74 | 90 | 100 |
|  | Precision | 100 | 100 | 100 | 100 | 100 | 100 |

The rate of recall coverage depends on the regularity of the page and GreenFIE’s generalization of regex extraction rules. Observe that precision remains constant at 100%, which means that GreenFIE generates no rules that extract incorrectly. Over-generalization can lead to a reduction of precision.

**4. Conclusion**

A demonstration version of GreenFIE operates with basic generalizations on single pages. Current efforts are focused on initialization of subsequent pages with the collection of extraction rules accumulated by processing previous pages. Application-specific generalizations are on the backlog. Although there is still much to do, the experimental result look promising.

**References**

[Nagy12] G. Nagy, Keynote Address: Back to the Future, *Family History Technology Workshop*, Salt Lake City, Utah, February, 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.