Pattern Markup-Language
A tool for simplifying data extraction from semi-structured sources

Jonathan Baker, Hilton Campbell
Jordan Crabtree, David W. Embley
Many Sites with Genealogical Data

- FrontierTimes - Outlaws, Gunfighters-Gunslingers and Lawmen of the West
- Genealogy.com: Ancestry of Jesse JAMES
- Genealogy.com: Ancestry of Wild Bill HICKOK
- Genealogy.com: Ancestry of Wyatt Earp - Wild West Personalities I Bang-Up Pedigree
- By Myra Vanderpool Gormley, CG.
- The Gunfighter Zone
- Jesse James
- John Wesley Hardin
- Judge Roy Bean
- Kansas Gunfighters
- Lawmen & Outlaws - Oklahoma Lawmen & Outlaws Including O. T. & O.K. Corral Famous Gunfight Site, Tombstone AZ
- OKLAHOMBRES Online!
  Oklahombres is an association for the preservation of lawman and outlaw history in Oklahoma. Has a searchable archive of past issues of our quarterly publication and message board where researchers can exchange information.
- OK Lawmen & Outlaws
  Lawmen and outlaws from Oklahoma Territory, Indian Territory or the State of Okla
- Outlaw JAMES Gang
- Outlaw Women - Who Really Tamed The Wild West!
- Outlaws and Lawmen of the Old West
- Scriba’s Tribute to Billy the Kid
Kansas Gunfighters, KS Outlaws and KS Lawmen

Table of Contents

- Hide Park Gunfight at Newton, Kansas
- Gunfight at the OK Corral, Tombstone, Arizona
- Benjamin Cardozo Meets Gunslinger Bat Masterson
- General Gunfighters History
- Kansas Gunfighters Sources

Articles and Books
- Cutler's History of Kansas, 1883
- Kansas Collection
- Connelley's History of Kansas, 1918
- Wild West Showl
- Dodge City History
- Dodge City, The Cowboy Capital, by Robert M. Wright
- The Rath Trail
- Native American Bibliographies
- Old West Gunfighter Books a powells.com bookstore
- US History Museums
- U.S. Marshals History
- Old West Museums, WWW-VL: American West
dollars in shiny new twenty-dollar gold coins from the San Francisco mint. The passengers of the train turned over an additional $400 cash and gold watches.

William H. Bonney - aka - Henry McCarty - aka - Billy the Kid: (1859 - 1881)

William Bonney was born in 1859. Billy the Kid was a lad with buck teeth who could do remarkable things with a .44-40 pistol. His career began in Silver City, New Mexico Territory. 14 Jul 1881 Billy the Kid was fatally shot by his old friend, Pat Garrett, in the bedroom of Pete Maxwell at Fort Sumner in New Mexico Territory. Billy the Kid died at age 21, having killed 21 men during his gunslinger career, a victim of circumstances, and many claim the dupe of the Lincoln County War.

William "Billy" L. Brooks (Abt. 1849 - 1874)

By 1870 he already had the reputation as a tough character. He was also supposed to have been a noted buffalo hunter and was to have been dubbed 'Buffalo Bill' (which confuses him with William F. Cody, the best known, or William Mathewson, the original Kansas 'Buffalo Bill' who was known as 'Buffalo Bill' as early as the 1860's). Brooks had appeared in Wichita in 1870, he was employed as a driver by the Southwestern Stage Company, the stage company switched routes to Newton, Brooks found that in Newton the cattle trade was in full swing and was in bad need of
William Bonney was born in 1859. Billy the Kid was Henry McCarty - aka - Billy the Kid.
William Bonney - aka - Henry McCarty - aka - Billy the Kid: (1859-1881)

William Bonney was born in 1859. Billy the Kid was a lad with buck teeth who could do remarkable things with a Ter...
H. Bonney – aka – Henry McCarty – aka – Billy the Kid:
(1859–1881)

William Bonney was born in 1859. Billy the Kid was a lad with buck teeth who could do remarkable things with a .44 single-action six-shooter. He was known for his wild, Territory days, when he and his gang terrorized the town at Fort Sumner. They were involved in a violent standoff with a vigilante group. Billy the Kid was killed by a Federal Marshal at Lincoln, New Mexico.
Newton Brown was born at Cold Spring, Twp, Missouri in 1857, had one sister, and entered law in 1884.
A member of the Rube Burrow train robbing gang in TV for a brief time, due to

Find/Replace dialog box showing search pattern and options.
Programmer Defined Regular Expressions

Regular Expression A

William H. Bonney - aka - Henry McCarty - aka - Billy the Kid (1859-1881)

William Bonney was born in 1859. Billy the Kid was a lad with buck teeth who could do remarkable things with
William H. Bonney" aka Henry McCarty aka Billy the Kid (1859-1881)

*William Bonney was born in 1859. Billy the Kid was a lad with buck teeth who could do remarkable things with a .44-40 pistol. His career began in Silver City, New Mexico.*
**Programmer Defined Regular Expressions**

<table>
<thead>
<tr>
<th>Regular Expression C</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>William Bonney</td>
<td>- aka - Henry McCarty - aka - Billy the Kid: (1859-1881)</td>
</tr>
<tr>
<td></td>
<td>&lt;strong&gt;William Bonney was born in 1859. Billy the Kid was a lad with buck teeth who could do remarkable things with</td>
</tr>
</tbody>
</table>
Which Relationships Found?

- Death Date
- Birth Date
- Given Name
- Aliases

Dollars in shiny new twenty dollar gold coins from the San Francisco mint. The passengers of the train turned over an additional $400 cash and gold watches.

William H. Bonney - aka - Henry McCarty - aka - Billy the Kid: (1859 - 1881)

William Bonney was born in 1859. Billy the Kid was a lad with buck teeth who could do remarkable things with a .44-40 pistol. His
Simple Schema
Represents Relationships

Person

Birth
  Date

Death
  Date

Names
  Given
  Aliases
Combine Schema and Regular Expressions

Tree Represented by XML = PatML
William H. Bonney - aka - Henry McCarty - aka - Billy the Kid: (1859-1881)

William Bonney was born in 1859. Billy the Kid was
William Bonney was born in 1859. Billy the Kid was H. Bonney/a - aka - Henry McCarty - aka - Billy the Kid.
William H. Bonney - aka - Henry McCarty - aka - Billy the Kid: (1859-1881)
William H. Bonney - aka - Henry McCarty - aka - Billy the Kid (1859-1881)

William Bonney was born in 1859. Billy the Kid was
PatML Generation Tools

Schema Generator
Establishes relationships
PatML Generation Tools

PatML Editor
Helps write the regular expressions and establish which facts they match
XML Schema Generator

Load File  Save  Save As

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genealogy Sheet</td>
<td>SCHEMA</td>
</tr>
<tr>
<td>Person</td>
<td>SINGLE</td>
</tr>
<tr>
<td></td>
<td>SINGLE</td>
</tr>
<tr>
<td>Gender</td>
<td>SINGLE</td>
</tr>
<tr>
<td>Name</td>
<td>SINGLE</td>
</tr>
<tr>
<td>Birth</td>
<td>SINGLE</td>
</tr>
<tr>
<td></td>
<td>SINGLE</td>
</tr>
<tr>
<td>Location</td>
<td>SINGLE</td>
</tr>
<tr>
<td>Date</td>
<td>SINGLE</td>
</tr>
<tr>
<td>Child</td>
<td>MULTI</td>
</tr>
<tr>
<td></td>
<td>SINGLE</td>
</tr>
<tr>
<td>Name</td>
<td>SINGLE</td>
</tr>
</tbody>
</table>
Using PatML Editor

- Get your schema file
- Browse for sample page
- Add nodes
- Add expressions
- See the highlights in source
- Adjust
PatML Editor Interface

- Tree representing PatML structure
- Text area with sample page source
- Browser with rendered sample page
<table>
<thead>
<tr>
<th>Type</th>
<th>RegEx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genealogy Sheet</td>
<td>This site powered by The N.</td>
</tr>
<tr>
<td>Person</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td></td>
</tr>
</tbody>
</table>
| PATTERN         | Name:(.*)<--End name-->
| PATTERN         | <li>(.*)<br>
| PATTERN         | ,*
| Birth           | Birth:(.*)<--End birth
| Date            | Date:*<li>(.*)<br>,*<--End date-->
| Location        | Location:ls*(.*)<--End loc

<html>
<h2>This information is not to be used for commercial purposes</h2>
<!--This site powered by The New Genealogy System-->
Name:
<li>Gregory Clark<br>
<!--End name-->
Birth:
<li>9 Jun 1950<br>
<!--End date-->
Children:
<li>Bob Clark<br>
<li>Bert Clark<br>
<li>Billy Clark<br>
<li>Barbara Clark<br>
<li>Bessy Clark<br>
<!--End children-->
</html>
Fast and Versatile

- Regular sites can be integrated in hours
- Adaptable to any type of information
Implementation to Date

- Genesis uses PatML files to search a variety of sites
  - Searches TNG, Retrospect-GDS, Family Search, GedCom and Kansas Gunslingers
  - Standardizes information for a common datamodel
  - Simultaneously searches other sites (in different formats) for people with similar information
Results
Results

- Produced PatML that correctly extracts data from TNG, RGDS, GedCom Sites, and Kansas Gunslingers
- User Interface allows for improved debugging environment
- \( \sim 1/10 \) coding time with PatML generation tools compared to similarly functioning hand coded parsers
Limitations

- Sites must be recognizable with regular expressions
  - Even regular sites have page to page HTML variations
- Programmer error with regular expressions
- Regular expression operations can be slow
Future work

- Automatic regular expression generation
- Parsing links to extract data on connected pages
- Use in other applications and fields
- XPath approaches