

# 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


Cyndi's List - Cowboys, Ranchers and The Wild West - Windows Internet Explorer


http://www.cyndislist.com/cowboys.htm

Google


Search web...

Cyndi's List - Cowboys, Ranchers and The Wild West

  
[Planting Your Family Tree Online](#)

  
[Netting Your Ancestors](#)

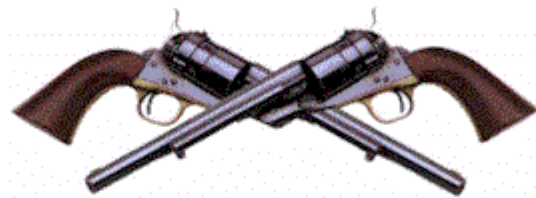
Advertisements

  
[Cyndi's List Genealogy](#)

- [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 & 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 a message board where researchers can exchange information.
- [OK Lawmen & Outlaws](#)  
Lawmen and outlaws from Oklahoma Territory, Indian Territory or the State of Oklahoma
- [Outlaw JAMES Gang](#)
- [Outlaw Women - Who Really Tamed The Wild West!](#)
- [Outlaws and Lawmen of the Old West](#)
- [Scribe's Tribute to Billy the Kid](#)

# KANSAS GUNFIGHTERS

*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](#)

[Sam Bass](#) // [William Bonney--Billy the Kid](#) // [William "Billy" L. Brooks](#) // [Henry Brown](#) // [Henderson Brunley](#) // [William F. Cody](#) // [Dalton Gang](#) // [William "Bill" M. Doolin](#) // [Wyatt Earp](#) // [Patrick "Pat" Floyd Garrett](#) // [John Wesley Hardin](#) // [Wild Bill Hickok--James Butler Hickok](#) // [John Henry "Doc" Holliday](#) // [Tom Horn](#) // [Jesse James Gang](#) // [William Bartholomew "Bat" Masterson](#) // [George Newcomb](#) // [Ed O'Kelly](#) // [James "Jim" Riley](#) // [Luke Short](#) // [Ben Thompson](#) // [Henry Clay White](#) // [Younger Gang](#)

## What's Inside


Old West Kansas

## Kansas Heritage


### Articles and Books

- [Cutler's History](#) of Kansas, 1883
- [Kansas Collection](#)
- [Connelley's History](#) of Kansas, 1918
- [Wild West Show!](#)
- [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)

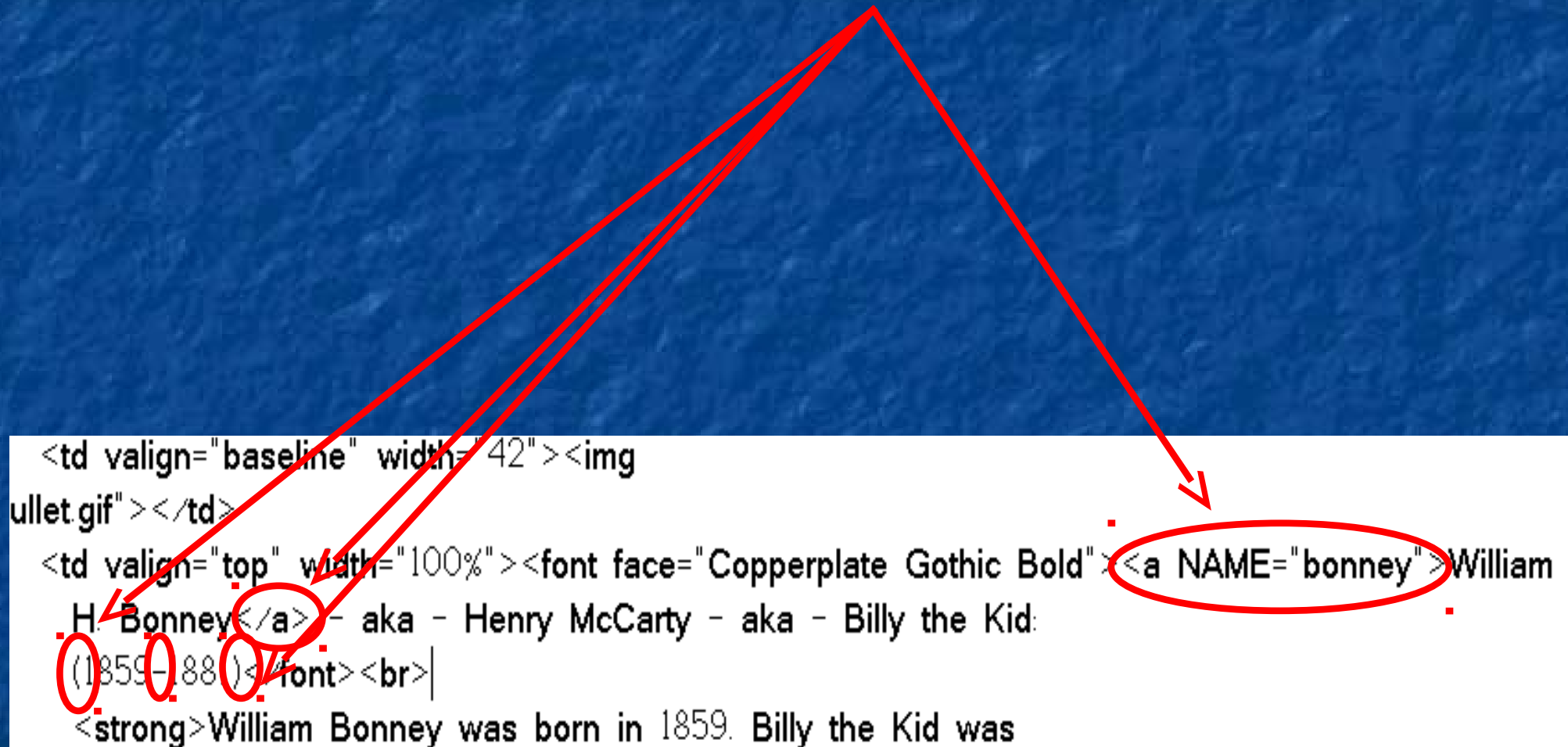
William Bonney was born in 1859. Billy the Kid was a lad with buck teeth who could do remarkable things with a .44-40 pistol. His career began in Silver City, New Mexico Territory. 14 Jul 1881 Billy the Kid was fatally shot by his old friend, Pat Garrett, in the bedroom of Pete Maxwell at Fort Sumner 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



# Structural Patterns



```
<td valign="baseline" width="42"><img  
ullet.gif"></td>  
<td valign="top" width="100%"><font face="Copperplate Gothic Bold"><a NAME="bonney">William  
H. Bonney</a> - aka - Henry McCarty - aka - Billy the Kid:  
(1859-1881)</font><br>  
<strong>William Bonney was born in 1859. Billy the Kid was
```



```
<td valign="baseline" width="42"><img  
'bullet.gif"></td>  
<td valign="top" width="100%"><font face="Copperplate Gothic Bold"><a NAME="bonney">William  
H. Bonney</a> - aka - Henry McCarty - aka - Billy the Kid:  
(1859-1881)</font><br>  
<strong>William Bonney was born in 1859. Billy the Kid was  
a lad with buck teeth who could do remarkable things with  
a .44  
Terr  
his o  
at Fo  
at ag  
a vid  
Linco  
</em>  
<!--r  
</td>  
</tr>  
<!--msim  
<tr>  
<!--ms  
<td va  
'bullet.gif  
<td va  
&quot;  
<stro  
a to  
buffe  
(whic
```

### Find/Replace

Find: `<a\sNAME="[^"]+">([^\<]+)`

Replace With:

Direction: ☒ Forward ☐ Backward

Scope: ☒ All ☐ Selected Lines

Options: ☒ Case Sensitive ☒ Wrap Search  
☐ Whole Word ☐ Incremental  
☒ Regular expressions

Find Replace/Find  
Replace Replace All  
Close

```
ME="brooks">William
```

et.gif"></td>  
<td valign="top" width="100%"><font face="Copperplate Gothic Bold"><a NAME="brown">Henry  
Newton Brown</a>: (1857 - 30 April 1884)</font><br>

<strong>Born at Cold Spring Twp, Missouri in 1857, had one

sister

an un

at se

entat

Billy

up in

marsh

appo

They

new

regar

Apri

terr

havin

the

Terr

they

by th

same

Brown

The

</em>

**Find/Replace**

Find:

Replace With:

**Direction**

☒ Forward

☐ Backward

**Scope**

☒ All

☐ Selected Lines

**Options**

☒ Case Sensitive ☒ Wrap Search

☐ Whole Word ☐ Incremental

☒ Regular expressions

Find Replace/Find Replace All Close



```
<td valign="baseline" width="42"><img
allet.gif"></td>
<td valign="top" width="100%"><font face="Copperplate Gothic Bold"><a NAME="brumley">Henderson
Brumley</a>:</font><br>
<em><strong>A member of the Rube Burrow train robbing gang
in TV for a short time. (has
</s
<!--
</td>
</tr>
<!--msi
<tr>
<!--m
<td v
allet.gi
<td v
Fre
<em
F.
</s
<hr
<!--
</td>
</tr>
<!--msi
<tr>
<!--m
```

**Find/Replace**

Find: `<a\sNAME="[^"]+">([^\<]+)`

Replace With:

**Direction**

☒ Forward

☐ Backward

**Scope**

☒ All

☐ Selected Lines

**Options**

☒ Case Sensitive ☒ Wrap Search

☐ Whole Word ☐ Incremental

☒ Regular expressions

Find Replace/Find

Replace Replace All

Close

AME="cody">William

# Programmer Defined Regular Expressions

## Regular Expression A



```
<td valign="top" width="100%"><font face="Copperplate Gothic Bold"><a NAME="bonney">William  
H. Bonney</a> - aka - Henry McCarty - aka - Billy the Kid:  
(1859-1881)</font><br>  
<strong>William Bonney was born in 1859. Billy the Kid was  
a lad with buck teeth who could do remarkable things with
```

# Programmer Defined Regular Expressions

Regular Expression B

```
bullet.gif"></td>  
<td valign="top" width="100%"><font face="Copperplate Gothic Bold"><a NAME="bonney">William  
H. Bonney</a> - aka - Henry McCarty - aka - Billy the Kid:  
(1859-1881)</font><br>  
<strong>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

Regular Expression C



```
bullet.gif"></td>  
<td valign="top" width="100%"><font face="Copperplate Gothic Bold"><a NAME="bonney">William  
H. Bonney</a> - aka - Henry McCarty - aka - Billy the Kid:  
(1859-1881)</font><br>  
<strong>William Bonney was born in 1859. Billy the Kid was  
a lad with buck teeth who could do remarkable things with
```

# 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.

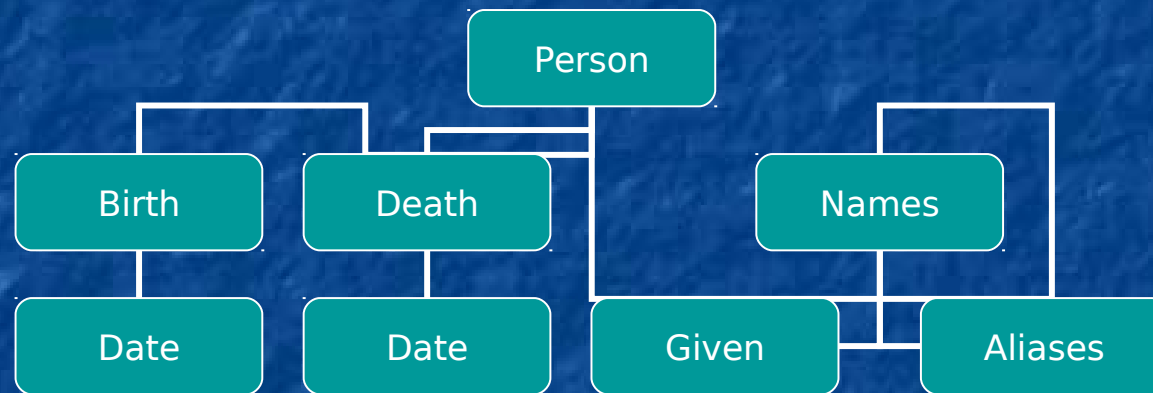
 William H. Bonney- aka - Henry McCarty - aka - Billy the Kid: (1859 - 1881)

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

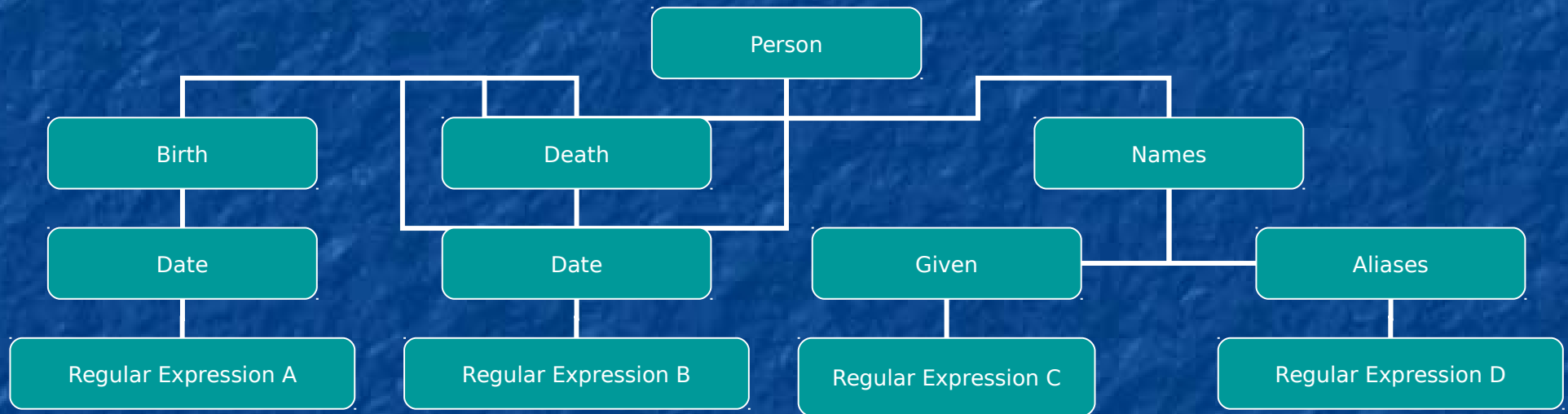
• [Edw](#)  
County  
• [W. I](#)  
• [Bat](#)  
• [Qua](#)  
• [Fre](#)



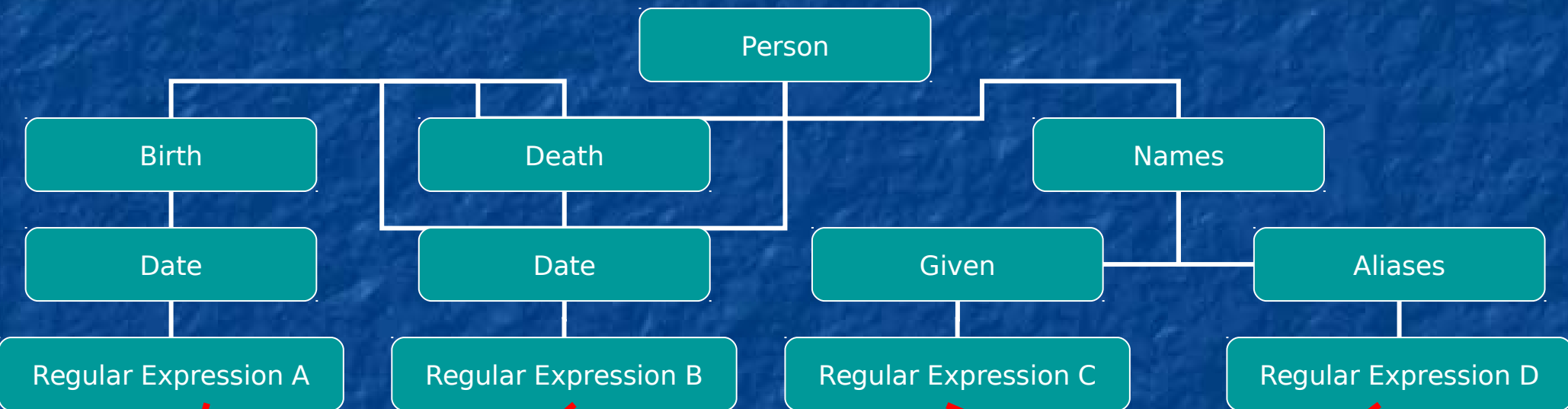
# Simple Schema Represents Relationships



# Combine Schema and Regular Expressions

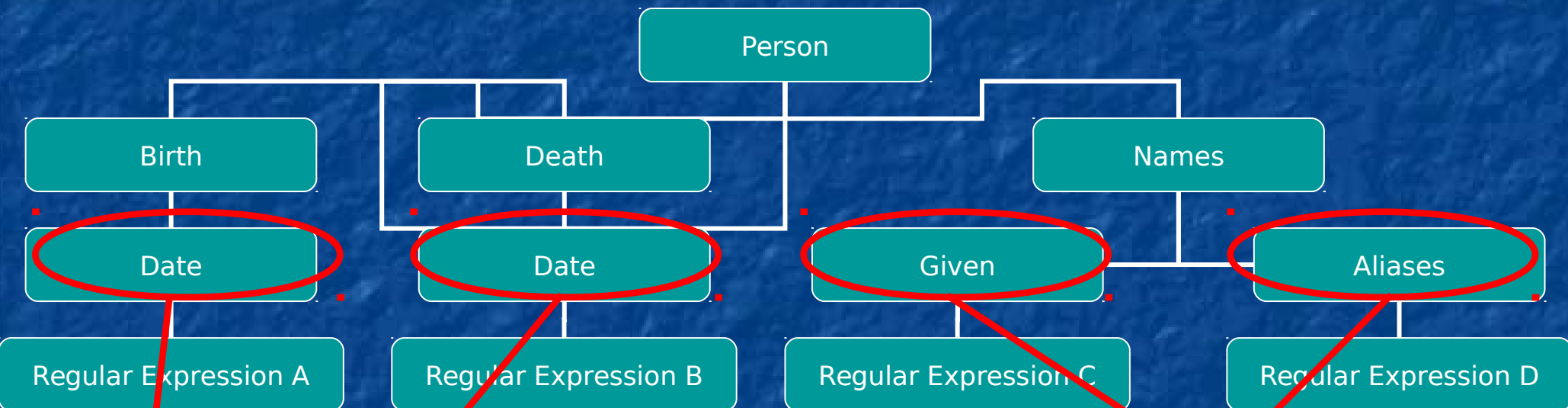


Tree Represented by XML =  
PatML



```

<td valign="baseline" width="42"><img
ullet.gif"></td>
<td valign="top" width="100%"><font face="Copperplate Gothic Bold"><a NAME="bonney">William
H. Bonney</a> - aka - Henry McCarty - aka - Billy the Kid:
(1859-1881)</font><br>
<strong>William Bonney was born in 1859. Billy the Kid was
  
```



```

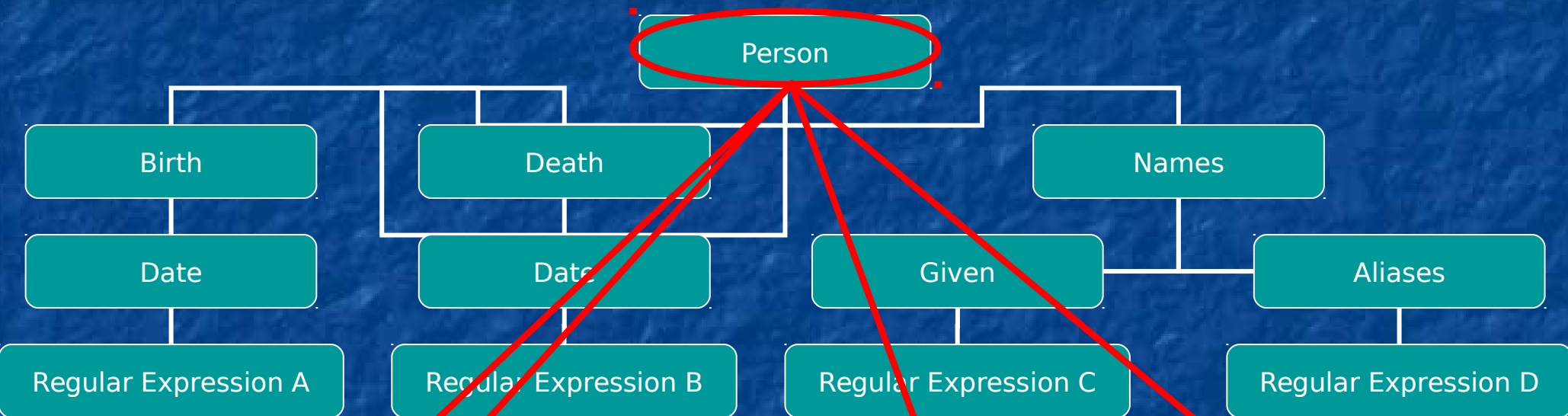
<td valign="baseline" width="42"><img
ullet.gif"></td>
<td valign="top" width="100%"><font face="Copperplate Gothic Bold"><a NAME="bonney">William
H. Bonney</a> - aka - Henry McCarty - aka - Billy the Kid:
(1859-1881)</font><br>
<strong>William Bonney was born in 1859. Billy the Kid was
  
```



```

<td valign="baseline" width="42"><img
ullet.gif"></td>
<td valign="top" width="100%"><font face="Copperplate Gothic Bold"><a NAME="bonney">William
H. Bonney</a> - aka - Henry McCarty - aka - Billy the Kid:
(1859-1881)</font><br>
<strong>William Bonney was born in 1859. Billy the Kid was
  
```

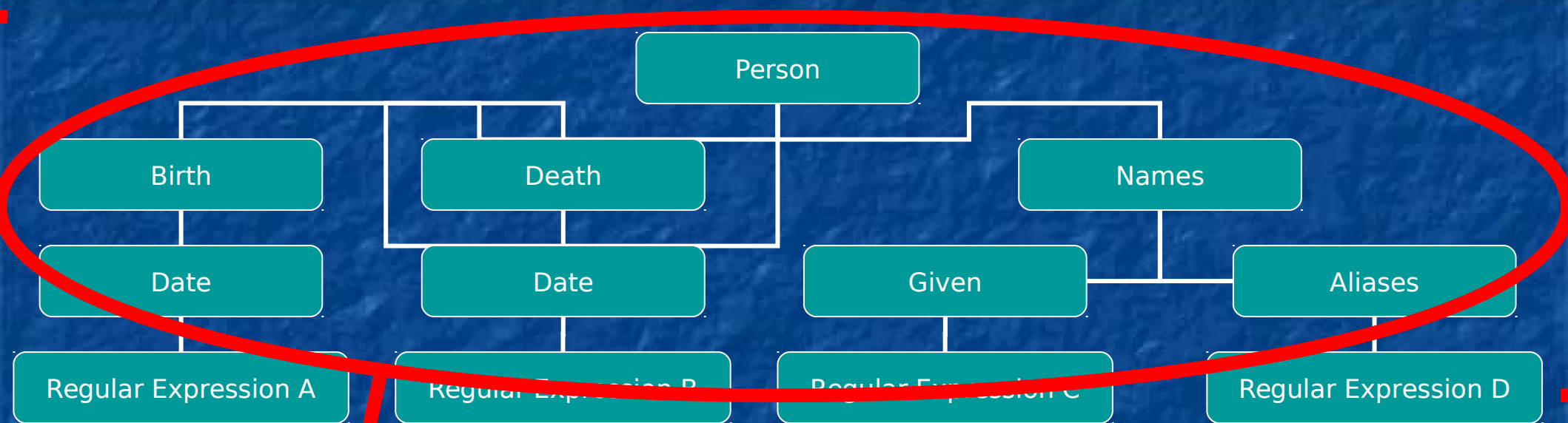




```

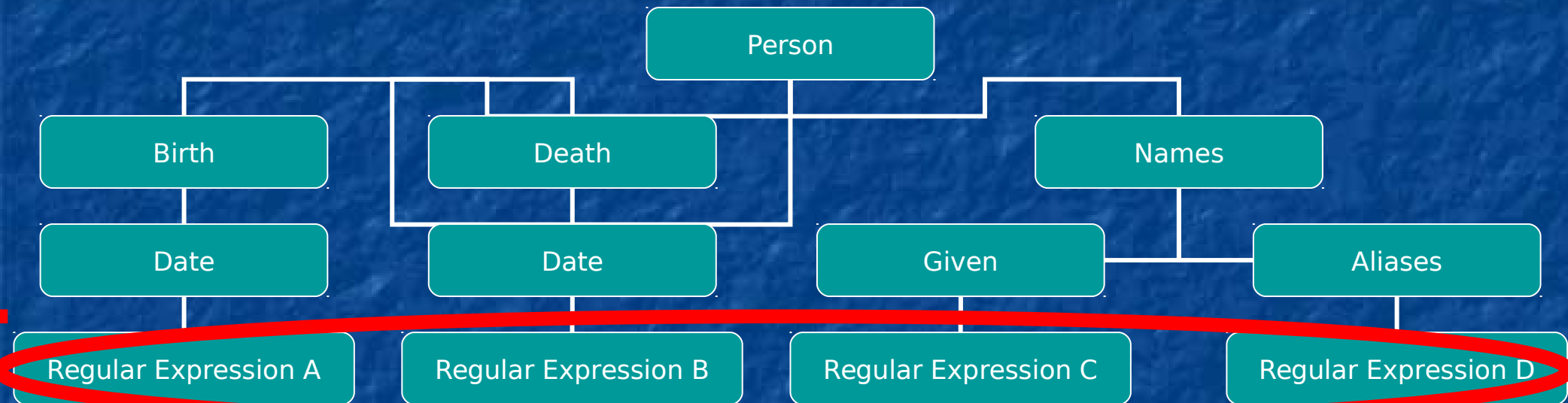
<td valign="baseline" width="42"><img
ullet.gif"></td>
<td valign="top" width="100%"><font face="Copperplate Gothic Bold"><a NAME="bonney">William
H. Bonney</a> - aka - Henry McCarty - aka - Billy the Kid:
(1859-1881)</font><br>
<strong>William Bonney was born in 1859. Billy the Kid was
  
```

# PatML Generation Tools



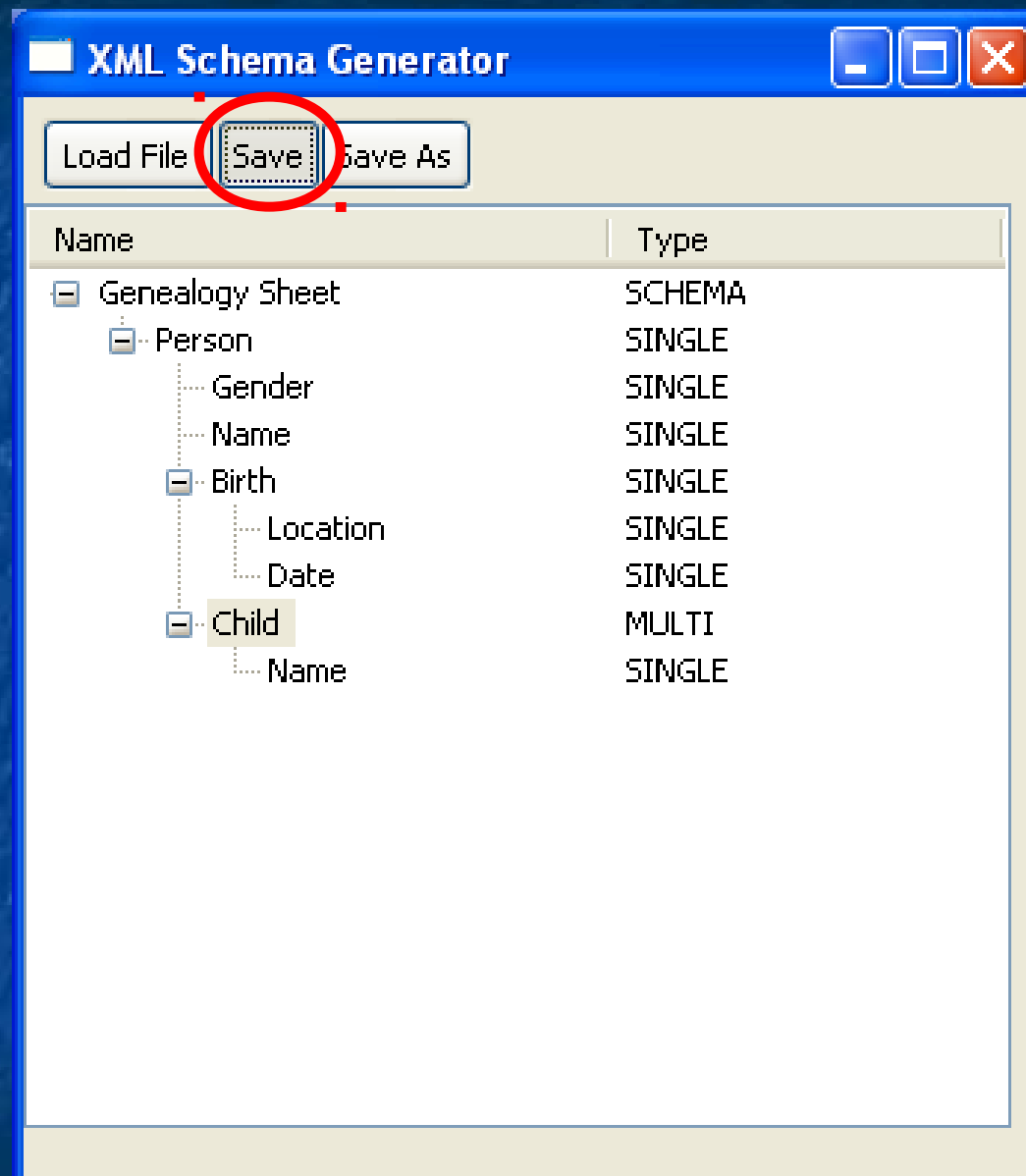
Schema Generator  
Establishes relationships

# PatML Generation Tools



## PatML Editor

Helps write the regular expressions and establish which facts they match



# Using PatML Editor

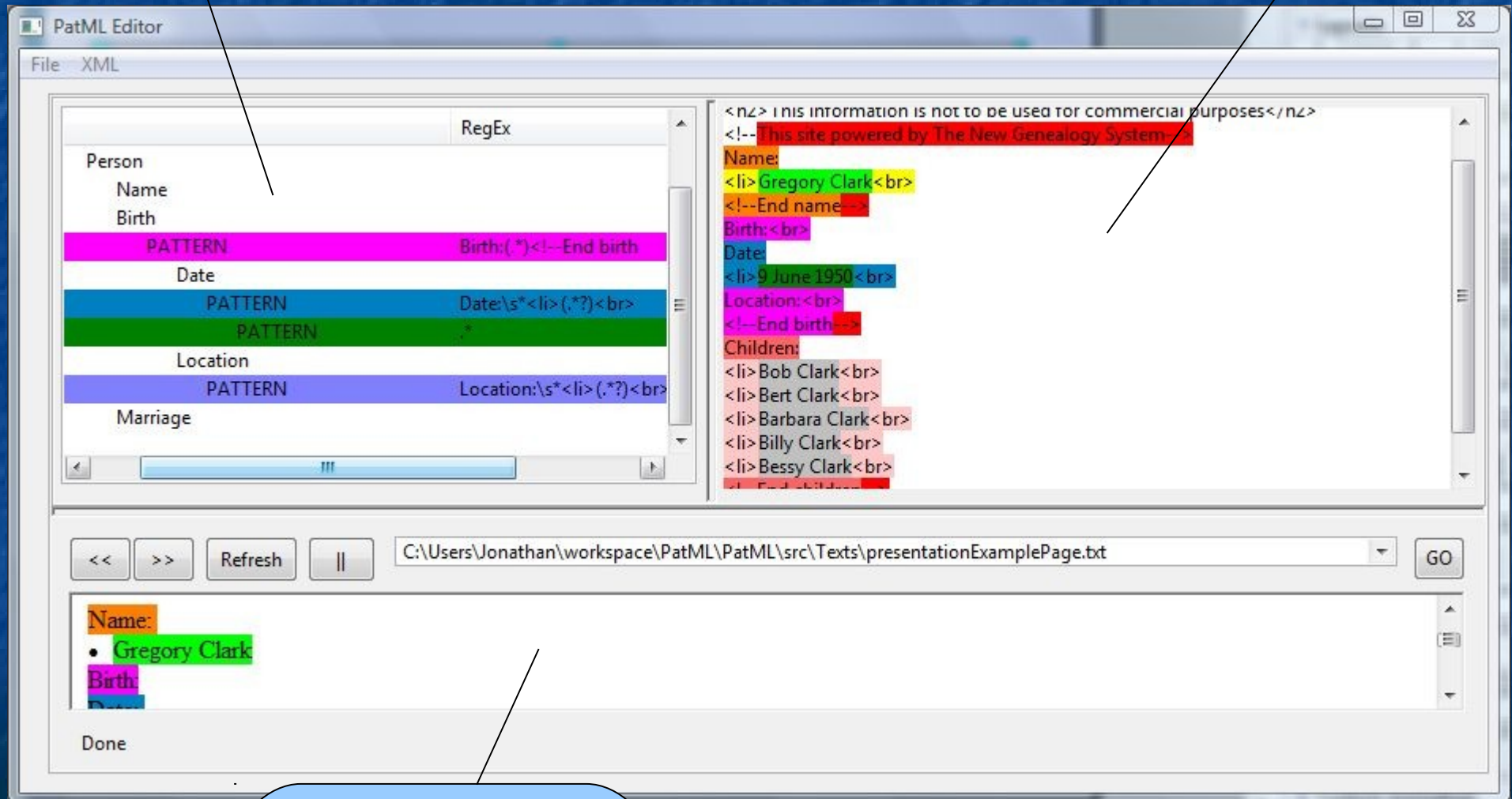
- Get your schema file
- Browse for sample page
- Add nodes
- Add expressions
- See the highlights in source
- Adjust



# PatML Editor Interface

Tree representing  
PatML structure

Text area with  
sample  
page source



Browser with  
rendered  
sample page

Markup Language

Type	RegEx
Genealogy Sheet	
PATTERN	This site powered by The N.
Person	
Name	
PATTERN	Name:(.*)<!--End name-->
PATTERN	<li>(.) 
PATTERN	,*
Birth	
PATTERN	Birth:(.*)<!--End birth-->
Date	
PATTERN	Date:.*<li>(.) .*<!--End date-->
PATTERN	,*
Location	
PATTERN	Location:.*<li>(.) .*<!--End location-->

```

<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: <br>
Date: <br>
<li>9 Jun 1950<br>
<!--End date-->
Location: <br>
<!--End birth-->
Children: <br>
<li>Bob Clark<br>
<li>Bert Clark<br>
<li>Billy Clark<br>
<li>BarbaraClark<br>
<li>Bessy Clark<br>
<!--End children-->
</html>

```

<< >> Refresh ||  GO

**This information is not to be used for commercial purposes**

Name:

- Gregory Clark

Birth:

Date:

- 9 Jun 1950

Location:

Children:

- Bob Clark
- Bert Clark
- Billy Clark
- BarbaraClark
- Bessy Clark

Done

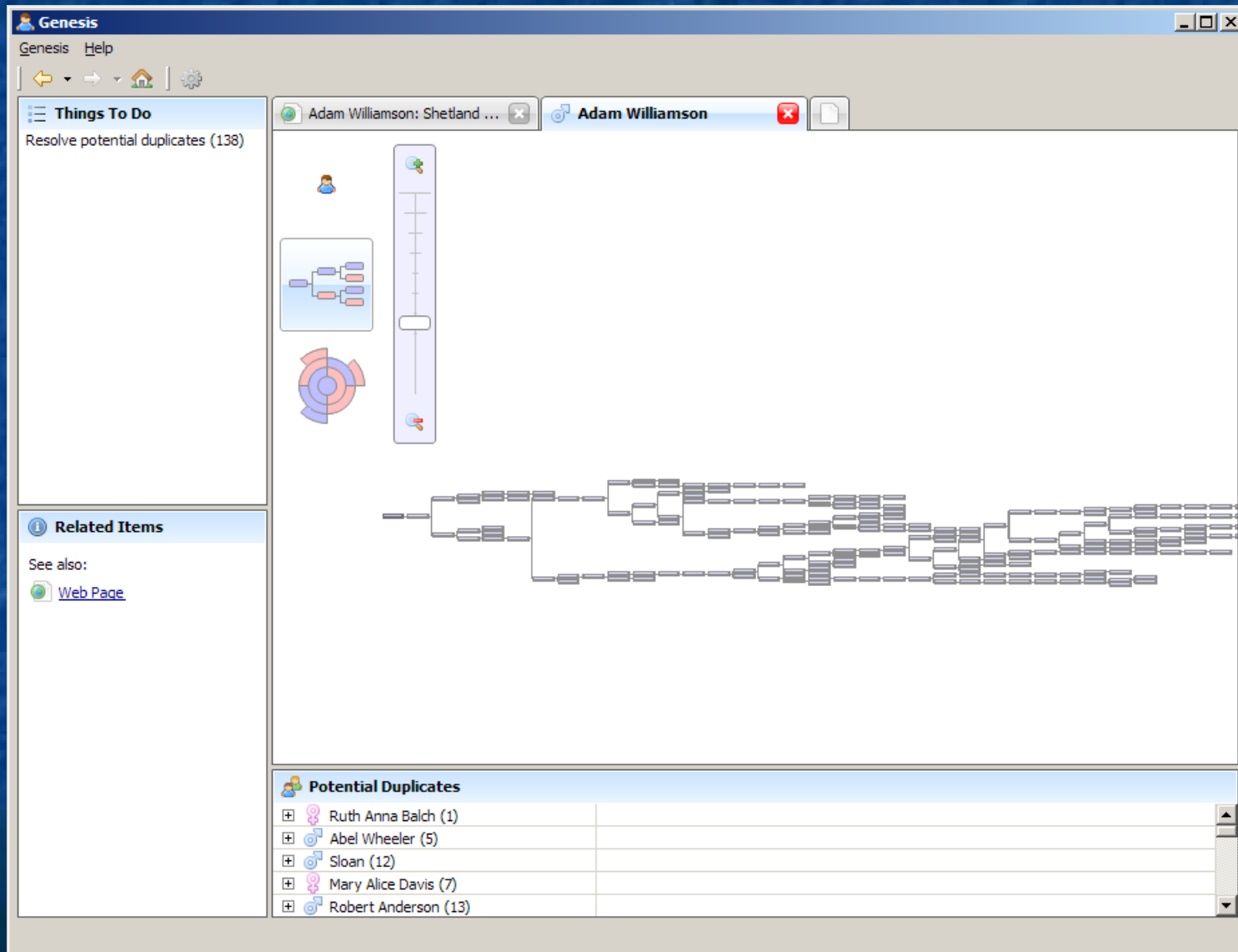
# Fast and Versatile

- 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
- ~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