Authoring Integrated, Dynamic Statistical Documents

S.E. Buttrey — Naval Postgraduate
D. Nolan — UC Berkeley
D. Temple Lang — Bell Labs

April, 2001

http://www.stat.berkeley.edu/∼statdocs

The Web

- Fantastic growth, acceptance, pervasive
- Simplicity
  - HTML
- Interactivity
  - Java and JavaScript
  - Multimedia plug-ins
- Applications
  - Web-based reporting
  - Electronic journals
  - On-line courses

Publishing Challenges

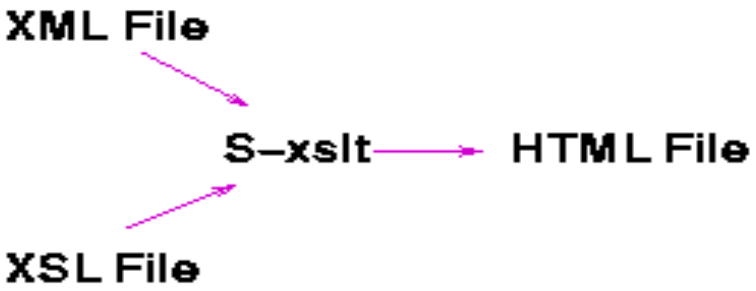
- Lack of statistical libraries and software
- Integration
  - text
  - numeric output
  - figures
  - code
- Reproducible results
  - track the analysis process
- Maintenance of templates
  - periodic or repeated reports

Application: Indoor Air Environment

- Goal– Public information on radon in homes
- Data – EPA survey of homes in the US
- Question – Should I remediate my home?
- Report
  - State report
  - General recommendations
  - State specific input – map, data, cost
- Interactive additions
  - Home owner – action level, family size
  - Risk analyst – cost function, parameters

Report Generation

- Author creates report template in XML:
  - contains text and code fragments.
  - state independent
- Transform to HTML document via XSL
  - state specific text, numbers and graphics
- XSL transformer calls embedded R directly.



Why XML and XSL?

- XML - eXtensible Markup Language
  - Generalization of HTML
  - Extensible - define new tags
  - Tags describe content
- XSL - eXtensible Stylesheet Language
  - Filter: rules for transforming XML tags
  - Apply multiple XSL templates
  - Different style sheets give different views
- XSL Translator - Embedded R interpreter (Omegahat)

Author writes XML

```
 <h1> <state/> </h1>

...of the
<code lang="S">nrow(statdata)</code> counties

<countymap/>

<histogram variable="gmean"/>
```

XSL rules

```
<xsl:template match="state">
 <xsl:value-of select="$state" />
</xsl:template>

<xsl:template match="code[@lang='S']">
 <xsl:value-of select="r:eval(string(.))" />
</xsl:template>

<xsl:template match="histogram">
<xsl:element name="img">
 <xsl:attribute name="src">
  <xsl:value-of select="r:histogram(@variable)"/
 </xsl:attribute>
</xsl:element>
</xsl:template>
```

Sources of Interactivity

- Forms – content computed for specific county
- Java – slider for changing action level
- JavaScript – handles events
- S plug-in – statistical computations (Omegahat)
- S graphics device – display plots (Omegahat)
- Other plug-ins – Tcl/Tk, Flash

Advantages of this approach

- Integrate text with numeric content
- Easy to tailor application – modular
- Reformat text w/out reformatting statistical content
- Link pieces of document
  sliders, plots, tables
- Decomposable programs –Not black box applets
- Commands are function calls, not catenated strings

Result – Authors

- Developers program in *their* language of choice
  Users invoke those functions in *their* language of choice.
- *Develop once, invoke anywhere.*
- Use the appropriate tool for the job
  access functionality in others.
- At times better for other applications to be in control.
- Part of general inter-system interface project of Omegahat.

Result – Readers

- Large audience of non-statisticians and researchers.
- Many using low-quality statistical methodology.
- Too complex to
  - learn other language
  - switch between applications synchronize data.
  - minor part of overall task.
- Allow *us* to our use tools in other applications.

Application: Chip Manufacturing

- Goal– Provide engineers with information to easily monitor manufacturing process.
- Data – Multiple lots per day.
  ≈ 50 wafers per lot.
  Multiple chips on each wafer.
- Question – What is the yield or proportion of acceptable chips?
- Report – Daily report on process
- Interactive additions
  - Different failure types for chips.
  - Spatial patterns of failures.
  - Generate plots/reports as they are needed

Application: Teaching the CLT

- Goal– Students learn basic CLT in familiar browser
  Teacher adapts according to level of student
- Data – Simulated on the fly
- Question – How big does n need to be?
- Interactive
  - Vary the distribution sampled
  - Vary the statistic computed
- Animation
  - Uncover the process
  - Use the same components

Embedded Graphics

- Use <**EMBED**> for graphics devices.

```
<EMBED TYPE="app/x-sgraphics"
        WIDTH=300 HEIGHT=300 NAME="distPlot">
```

- Treat device as JavaScript object
  self-activating.
- Provides plotting methods as JavaScript methods.

```
distPlot.call("showPopulation", args);
```

Authoring Tools

- Add dynamic components to document
- Connect components to each other
- Connect components to statistical system
- Add dynamic feedback to document
- Add GUI components to document

Authoring Process

- Write text
  - Netscape composer
  - psgml mode for emacs
  - Frame Maker
- Edit text - tree of XML nodes
- R commands
  - XML representation of R commands
  - Edit an R session
- R output
  - XML representation of R output
  - Merge R with document

Plans

- Develop XSL templates
  - Consultants
  - Research Papers
  - Education
- Build GUI component library based on S
- Provide prototype documents
- Design Authoring tools
- Direct Manipulation graphics