

Virginia News Archive

or

*building a resource with almost no
staff*



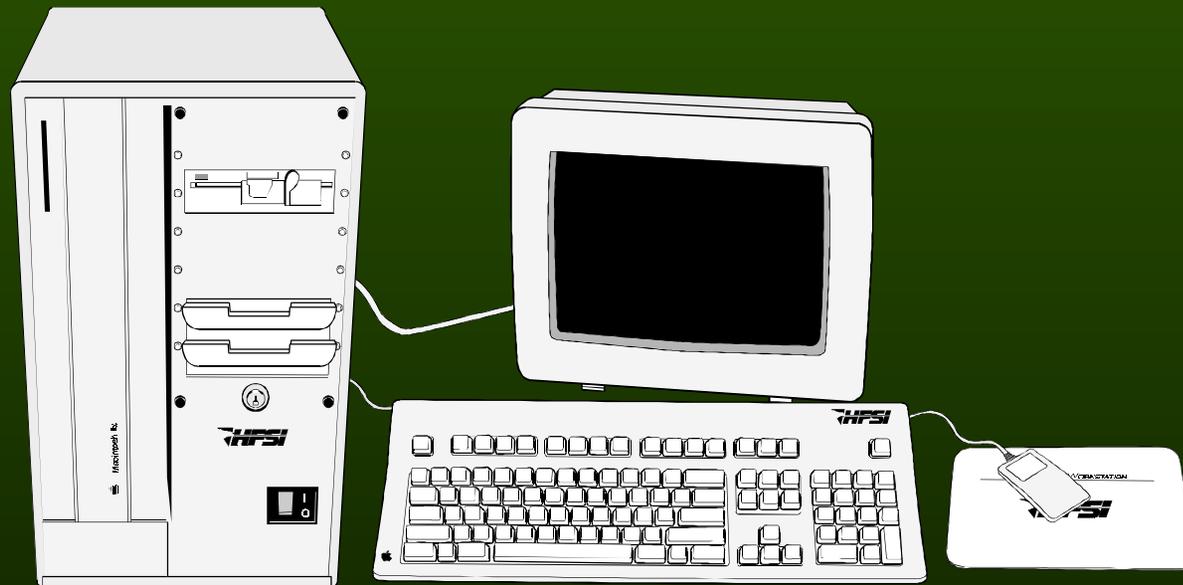
SCP Servers and Data

| Scholar | Scholar2 | Scholar3 | Scholar4 |
|---------------------------------------------------------------------|-----------------------------------------------------------------------|-----------------------------------------------------------------------|-----------------------------------------------------------------------|
| Electronic Journals/ Research Data | Digital Images | Electronic Newspapers (VA-Pilot) | Electronic Newspapers (ROA Times and Spectrum) |
| Library and VT Publications | Special Collections Projects | WDBJ-7 script/ photo archive | Library Publications |
| Electronic Theses and Dissertations | Art and Architecture Projects | Project backup server | |
| Main Project home page | | | |
| http://scholar.lib.vt.edu/ | http://scholar2.lib.vt.edu/ | http://scholar3.lib.vt.edu/ | http://scholar4.lib.vt.edu/ |



Why so many servers?

- Improved performance
- Better security
- Less chance of the entire project going down



Scholarly Communications Project



Electronic Newspapers

- Full text searchable/browsable
- HTML format - compatible with all web browsers
- Updated nightly
- Available 24 hours a day
- No access restrictions





What newspapers do we publish?

- Roanoke Times



- Virginian-Pilot



- Virginia Tech Spectrum





How do we get an issue?

Both the Roanoke Times and Virginian Pilot maintain an in-house full text database called VU/TEXT. VU/TEXT is not networkable but is accessible via dial-up.

```

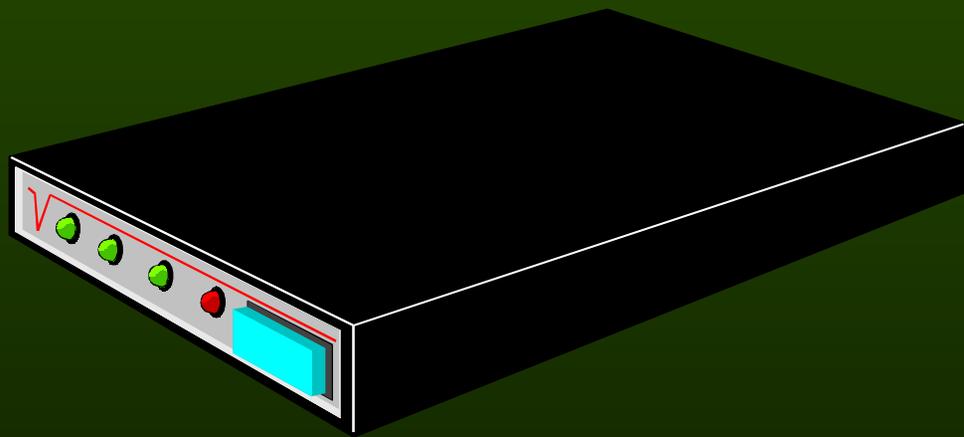
A. headline.1
DOC  DATE  FREQ LINES  DB  HEADLINE
=====
 1 08/11/96  0   161 cur  RAMBLING THE BACK ROADS OF UNSCRUBBED NEW MEXICO
 2 08/11/96  0   139 cur  NAPA VALLEY WINEMAKERS DRINK IN SUCCESS
 3 08/11/96  0    83 cur  ATMS ARE ABOUT TO GET A LOT SMARTER
 4 08/11/96  0    83 cur  EMERGENCY BABY SITTEERS SAVE DAY FOR EMPLOYEES
 5 08/11/96  0    73 cur  RICHMOND PACKAGING PLANT KEEPS TURNING OUT LITTLE
 6 08/11/96  0    79 cur  EXPERIENCE IS CRUCIAL WHEN LOOKING FOR A HIGHER S
 7 08/11/96  0   129 cur  LONG-AIRDOX NAMES PLANT MANAGER
 8 08/11/96  0   476 cur  GUIDE TO APARTMENT LIVING
 9 08/11/96  0    25 cur  UNHERALDED HEROES
10 08/11/96  0    67 cur  SISTER'S CALL TO TEACH A GODSEND FOR MANY
11 08/11/96  0    68 cur  ROANOKE COULDN'T BE BUILT WITHOUT INSPECTOR SMITH
12 08/11/96  0    72 cur  MR. MARKET BUILDING IS THE LUNCH RUSH
13 08/11/96  0    67 cur  GIVE HER AN A FOR MAKING SURE KIDS REACH SCHOOL
14 08/11/96  0    65 cur  PHARMACIST DISPENSES MORE THAN MEDICINE
15 08/11/96  0    65 cur  PLANNER'S JOB IS KIDS' PLAY (AND ADULTS', TOO)
16 08/11/96  0    67 cur  THEY WROTE THE BOOK ON INFORMATION GATHERING
=====
60 Docs                                     Pg 1 of 4

Type first letter of feature OR type help for list of commands
FIND S-DB DB OPT SS WRD QUIT

```

Automated Dial-up

- Initially, searches were performed by a person who dialed up their service, logged in and captured an issue each day.
- Now, automated scripts dial up during off-peak hours to retrieve an issue.



Retrieving an issue

- First, the script dials the modem, and logs in
- Next, it performs a search for all articles published on a certain day using the VU/TEXT *find* (dd/mm/yy) command
- All articles are then “printed” to the screen and simultaneously captured in a single text file.



Tagging an Issue

- As with the retrieval process, we started out using real people who tagged an issue manually.
- This is a labor-intensive process as the typical issue is 60-80 articles per day!
- Header information proved to be the key to automating markup.



VU/TEXT Article Header

- The article header can vary but the three fields essential for automated processing (section, tag, date) are always present:

ROANOKE TIMES
Copyright (c) 1996, Roanoke Times

DATE: Sunday, August 4, 1996
SECTION: CURRENT
COLUMN: Claws & Paws
SOURCE: JILL BOWEN

TAG: 9608050001
PAGE: NRV15 EDITION: NEW RIVER



Raw text to HTML: rt_txt2html

- We developed a Perl script which recognizes and “decodes” article headers and splits an issue capture file into individual articles:

```
Printing ...
Printing ...
Press [RETURN] to continue or type q to return to Menu:
cur  WHITE-COLLAR WORKERS TRY THE SELLING LIFE ON FOR SIZE    02/11/96
=====
                ROANOKE TIMES
            Copyright (c) 1996, Roanoke Times

DATE: Sunday, February 11, 1996      TAG: 9602090022
SECTION: BUSINESS                    PAGE: G1  EDITION: METRO
SOURCE: TRIP GABRIEL THE NEW YORK TIMES

                WHITE-COLLAR WORKERS TRY THE SELLING LIFE ON FOR SIZE

To sell vitamins and shampoo to friends and neighbors, Sharon Killion
```

Building an Issue

- As each article is extracted, a line is added to Section index file (e.g. SPORT.html for the Sports Section).
- Each new section is added to the issue index file: index.html.
- Each article is written to a file named after the TAG field, which is a unique identifier in VU/TEXT.



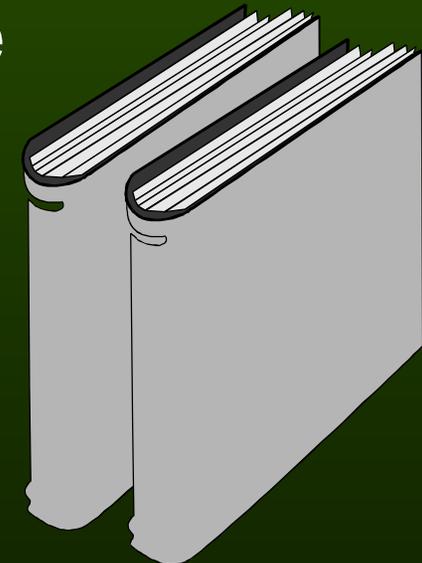
Tagging an Article

- The script retains the header as a single block and tags it with the HTML preformatted text tag `<PRE>`
- The script identifies the article title by its proximity to the header and a blank line between it and the article body. It makes the title a level 1 heading `<H1>`
- Finally, the script assumes each indented line it encounters represents a new paragraph tags it with a paragraph tag `<P>`



Archiving the Results

- The completed index.html, section index files and article files are placed into a single directory
- This directory is moved to the issues/19xx subdirectory under the web server and the issue is then publicly available





Indexing

- Depending on the paper, the archive is indexed with WAIS or Excite
- We are migrating all our newspapers to Excite because
 - the entire archive can be indexed (WAIS forces us to index each year separately)
 - the search interface is more flexible
 - many experts consider Excite to be one of the best search engines currently available
 - Commercial quality software at a great price: free

Reindexing

- WAIS is quite slow. It takes 28-32 hours to reindex a year of the Virginian-Pilot. 
- Excite can reindex FIVE YEARS in 5 hours.

This is because:

- Excite is faster
- Its index is smaller
- The server is a dual-processor Pentium



FAST



Virginia Tech Spectrum

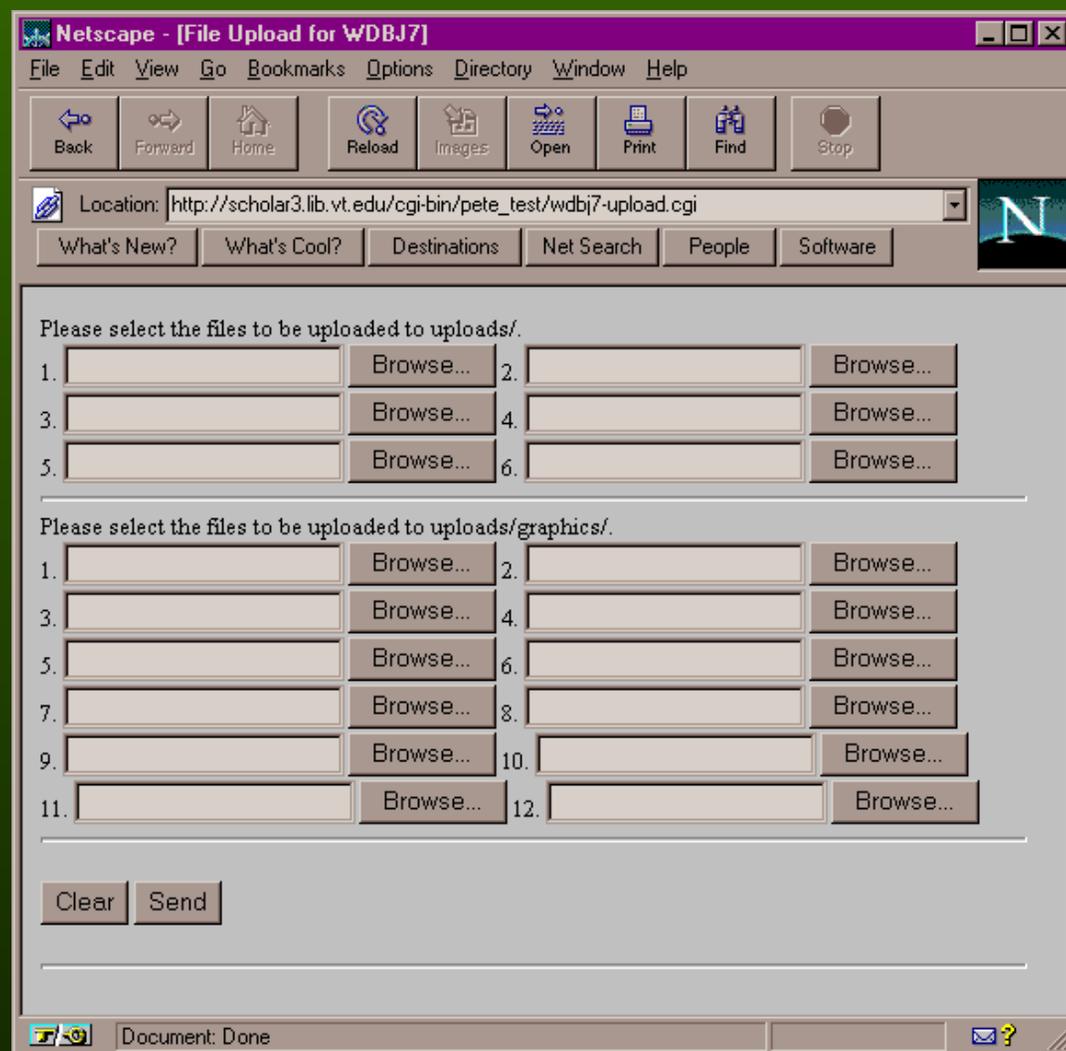
- The Spectrum requires more human intervention but some steps have been automated.
- Issues are delivered via Appleshare as RTF files.
- Mac rtftohtml is used to convert each article to HTML.
- An index file pointing to all articles is created manually.

WDBJ-7 Script/Image Archive

- Completely automated
Upload, Cleanup, Markup, Archival, WAIS
- Full search and browsing capability
- Low to no maintenance
- Intuitive interface
- Ease of use by WDBJ-7 staff

Upload and cleanup of files

- This page is accessed using Netscape 2.0 or better
- The selected files are uploaded
- The cleanup process begins (DOS to UNIX)
- NULL spaces
- Files are marked up using a C program



The screenshot shows a Netscape browser window titled "Netscape - [File Upload for WDBJ7]". The address bar contains the URL "http://scholar3.lib.vt.edu/cgi-bin/pete_test/wdbj7-upload.cgi". The browser interface includes a menu bar (File, Edit, View, Go, Bookmarks, Options, Directory, Window, Help) and a toolbar with buttons for Back, Forward, Home, Reload, Images, Open, Print, Find, and Stop. Below the toolbar is a search bar with "Location:" and the same URL. There are also buttons for "What's New?", "What's Cool?", "Destinations", "Net Search", "People", and "Software".

The main content area contains two sections for file selection:

Please select the files to be uploaded to uploads/.

| | | | | | |
|----|----------------------|-----------|----|----------------------|-----------|
| 1. | <input type="text"/> | Browse... | 2. | <input type="text"/> | Browse... |
| 3. | <input type="text"/> | Browse... | 4. | <input type="text"/> | Browse... |
| 5. | <input type="text"/> | Browse... | 6. | <input type="text"/> | Browse... |

Please select the files to be uploaded to uploads/graphics/.

| | | | | | |
|-----|----------------------|-----------|-----|----------------------|-----------|
| 1. | <input type="text"/> | Browse... | 2. | <input type="text"/> | Browse... |
| 3. | <input type="text"/> | Browse... | 4. | <input type="text"/> | Browse... |
| 5. | <input type="text"/> | Browse... | 6. | <input type="text"/> | Browse... |
| 7. | <input type="text"/> | Browse... | 8. | <input type="text"/> | Browse... |
| 9. | <input type="text"/> | Browse... | 10. | <input type="text"/> | Browse... |
| 11. | <input type="text"/> | Browse... | 12. | <input type="text"/> | Browse... |

At the bottom of the form are "Clear" and "Send" buttons.



Markup and archival

- Cleaned files and “magic”
- Using words and strings of characters to decide markup
- Adding the files to the archive
- Building the list of files in the archive
- Building the up to date HTML index pages
- Building months, years



Intuitive Interface

YOUR HOMETOWN STATION
WDBJ-7

'95 '96

'96

| | | | | | |
|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Jan | Feb | Mar | Apr | May | Jun |
| Jul | Aug | Sep | Oct | Nov | Dec |

February 1996

| Sun | Mon | Tue | Wed | Thu | Fri | Sat |
|--------------------------------------------------|--------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------|
| | | | | 1 Morn 5pm AM 6pm Noon 11pm | 2 Morn 5pm AM 6pm Noon 11pm | 3 6pm 11pm |
| 4 6pm 11pm | 5 Morn 5pm AM 6pm Noon 11pm | 6 Morn 5pm AM 6pm Noon 11pm | 7 Morn 5pm AM 6pm Noon 11pm | 8 Morn 5pm AM 6pm Noon 11pm | 9 Morn 5pm AM 6pm Noon 11pm | 10 6pm 11pm |

How to read a WDBJ-7 news script

There are six records filed each weekday:

- [News-7 Mornin'.....Airs 6:00 a.m. to 7:00 a.m.](#)
- [AM...5-minute news segments that air at 7:25 a.m., 7:55 a.m. and 8:25 a.m.](#)
- [News-7 at Noon.....30-minute newscast](#)
- [News-7 at 5.....30-minute newscast](#)
- [News-7 at Six.....30-minute newscast](#)
- [News-7 at Eleven.....35-minute newscast](#)

The text versions of newscasts include scripting instructions that help in the production of the newscast. Here's a list of what these symbols mean:

- **SLUG**= This is the title of the story that appears at the top. It allows the producers to keep track of each day's stories. Beyond that, the slug has no significance and is not part of the story.
- **ANCHOR**= The first name of the anchor for that story.
- **WRITER**= The computer initials of the reporter, anchor or producer who wrote the story.
- **TAPE#**= If the story has videotape with it, this tells the producer where it is.
- **CD**...**DTTC**...

Ease of use for WDBJ-7 staff

- Form upload and symbolic links
- Just drop the files and go
- No need to ask for SCP “staff” to do anything
- Continuous update of scripts
- Can add files to archive at any time
- Can not remove files accidentally
- No need for anything other than a WWW browser