WebCheck

Mark Balbes, Ph.D.

Osiris Consultants

and

Computerized Medical Systems, Inc.

Symantec Café 1.5.1

- JDK 1.02
- Good debugger
- Bad editor. Doesn’t do enough automatic formatting.
- Inconsistent or bad code generation.
- Overall, a reasonable environment

What does WebCheck do?

- Check for invalid active links on a web site
- Check for invalid links in a local file or directory
- Recursive file or link checking
- Pause checking
- Give real-time feedback
- Bookmark frequently checked sites
- Generate a report.

Recursive WebChecking

Files

- Check all subdirectories.
- Don’t follow links.

Batch mode

•Still to come

Example

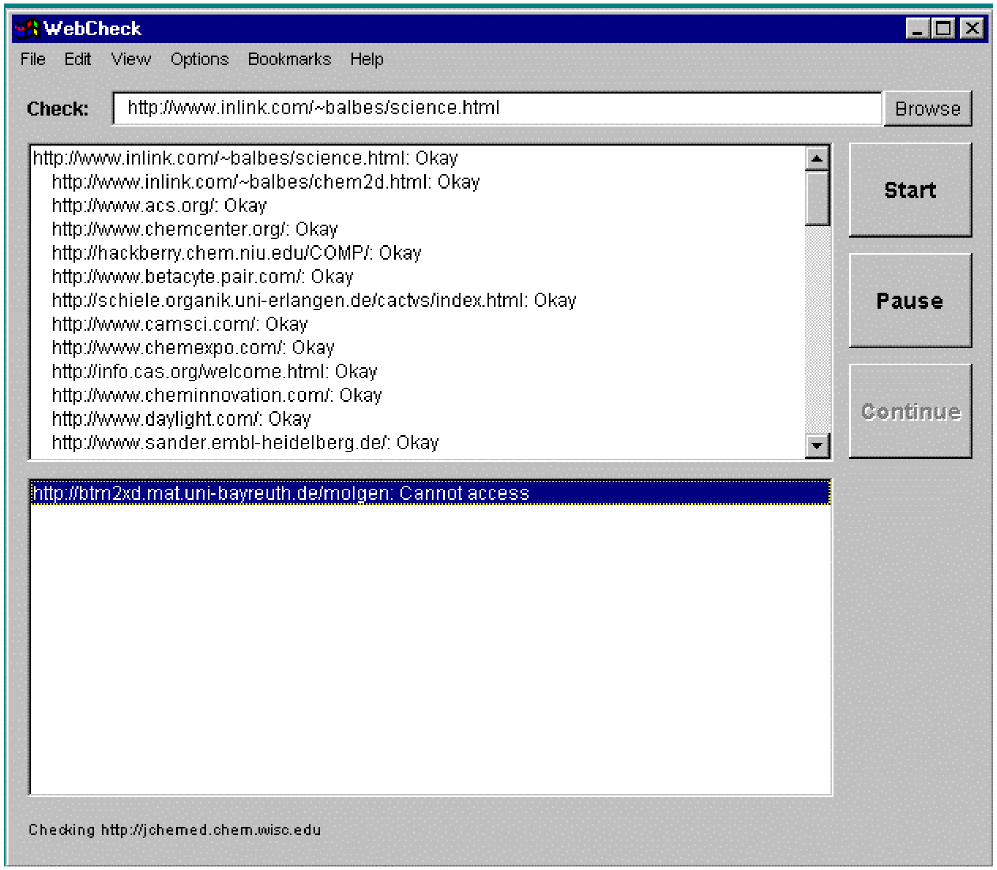
http://www.cms-stl.com
http://www.NIH.com
product/Focus.html
handbook.html

http://www.nytimes.com

Live Sites

- Follow live links if:
  - relative link
  - absolute link that contains all but the filename of the initial URL.
  - Must be careful not to “breakout” of the site being checked
  - Must be careful not to recheck URL’s that were already checked

The GUI in Windows 95



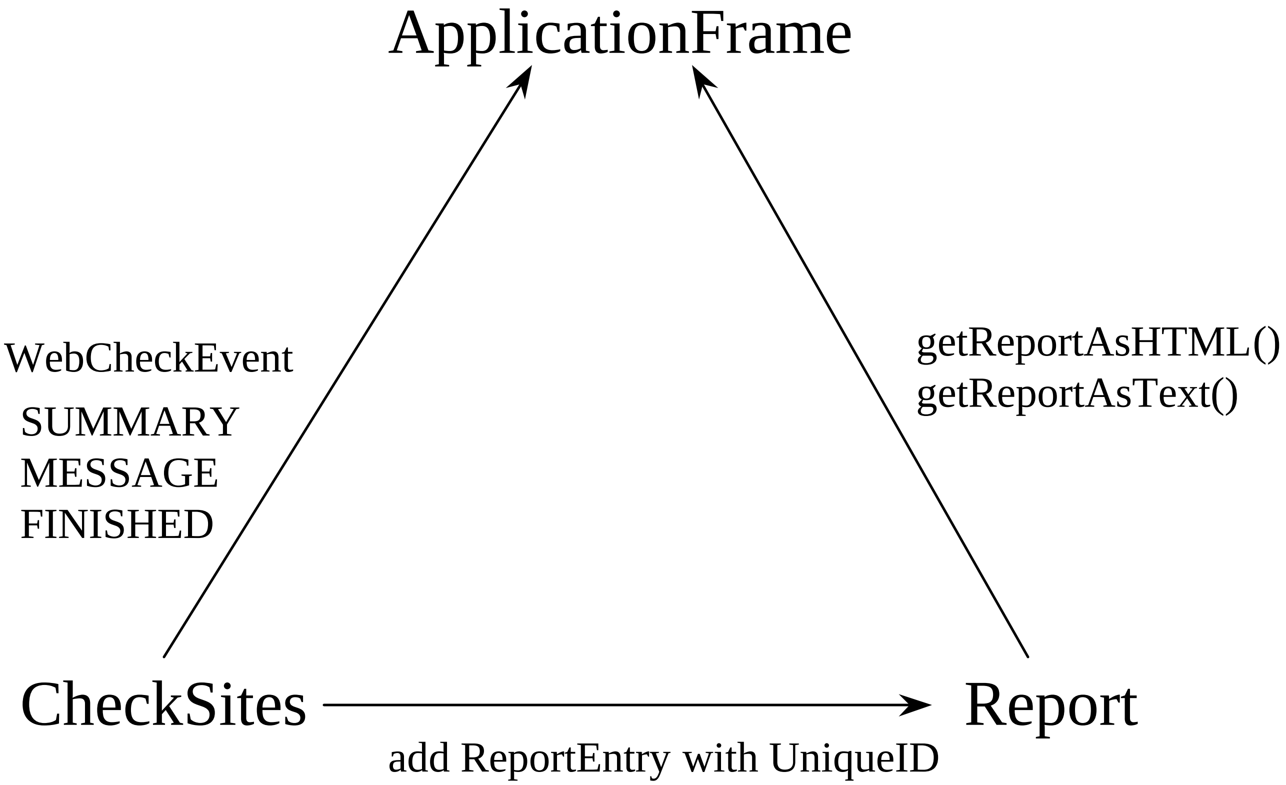
Check: http://www.inlink.com/~balbes/science.html

```
http://www.inlink.com/~balbes/science.html: Okay
   http://www.inlink.com/~balbes/chem2d.html: Okay
   http://www.acs.org/: Okay
   http://www.chemcenter.org/: Okay
   http://hackberry.chem.niu.edu/COMP/: Okay
   http://www.betacyte.pair.com/: Okay
   http://schiele.organik.uni-erlangen.de/cactvs/index.html: Okay
   http://www.camsci.com/: Okay
   http://www.chemexpo.com/: Okay
   http://info.cas.org/welcome.html: Okay
   http://www.cheminnovation.com/: Okay
   http://www.daylight.com/: Okay
   http://www.sander.embl-heidelberg.de/: Okay
```

```
http://btm2xd.mat.uni-bayreuth.de/molgen: Cannot access
```

Checking http://jchemed.chem.wisc.edu

WebCheck Design



ApplicationFrame

WebCheckEvent

SUMMARY
MESSAGE
FINISHED

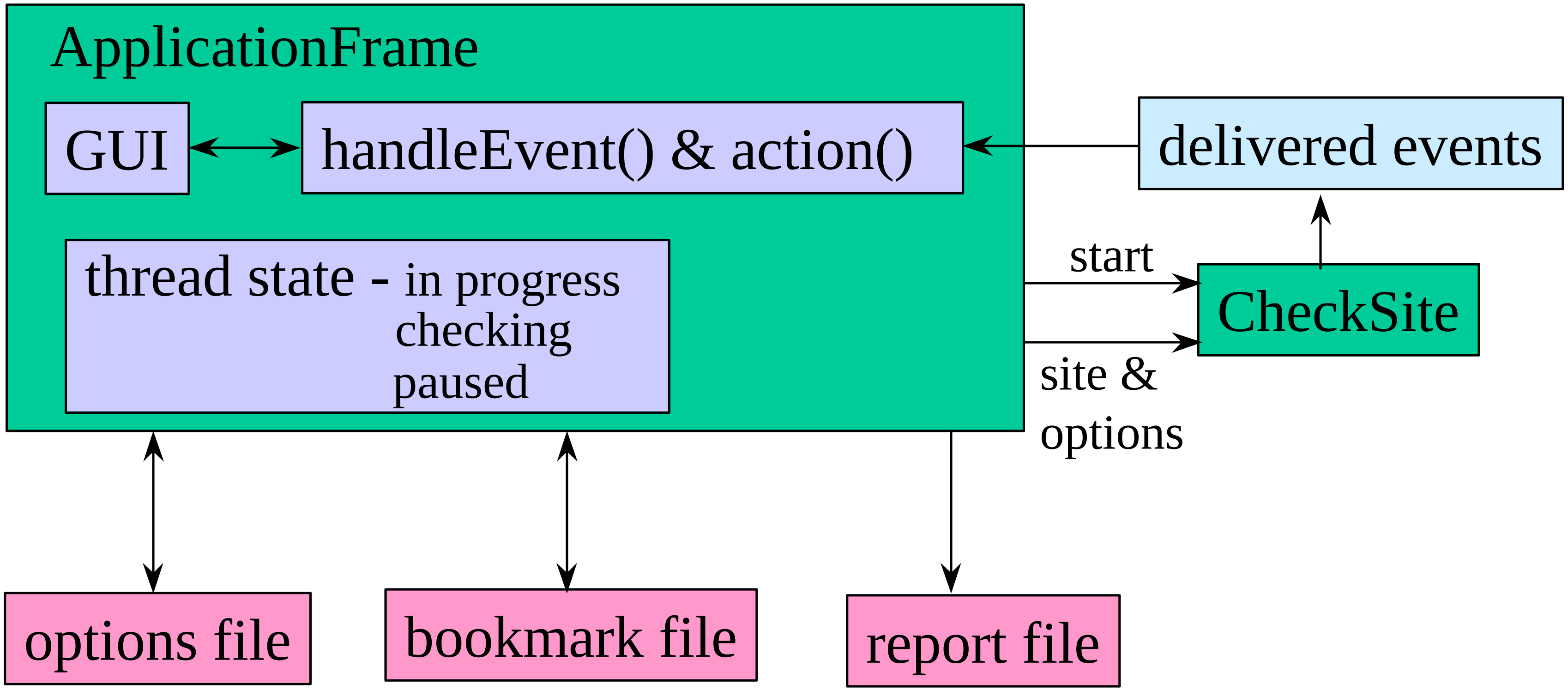
getReportAsHTML()
getReportAsText()

CheckSites

Report

add ReportEntry with UniqueID

ApplicationFrame class


ApplicationFrame

GUI

handleEvent() & action()

delivered events

thread state - in progress
checking
paused

start

CheckSite

site &
options

options file

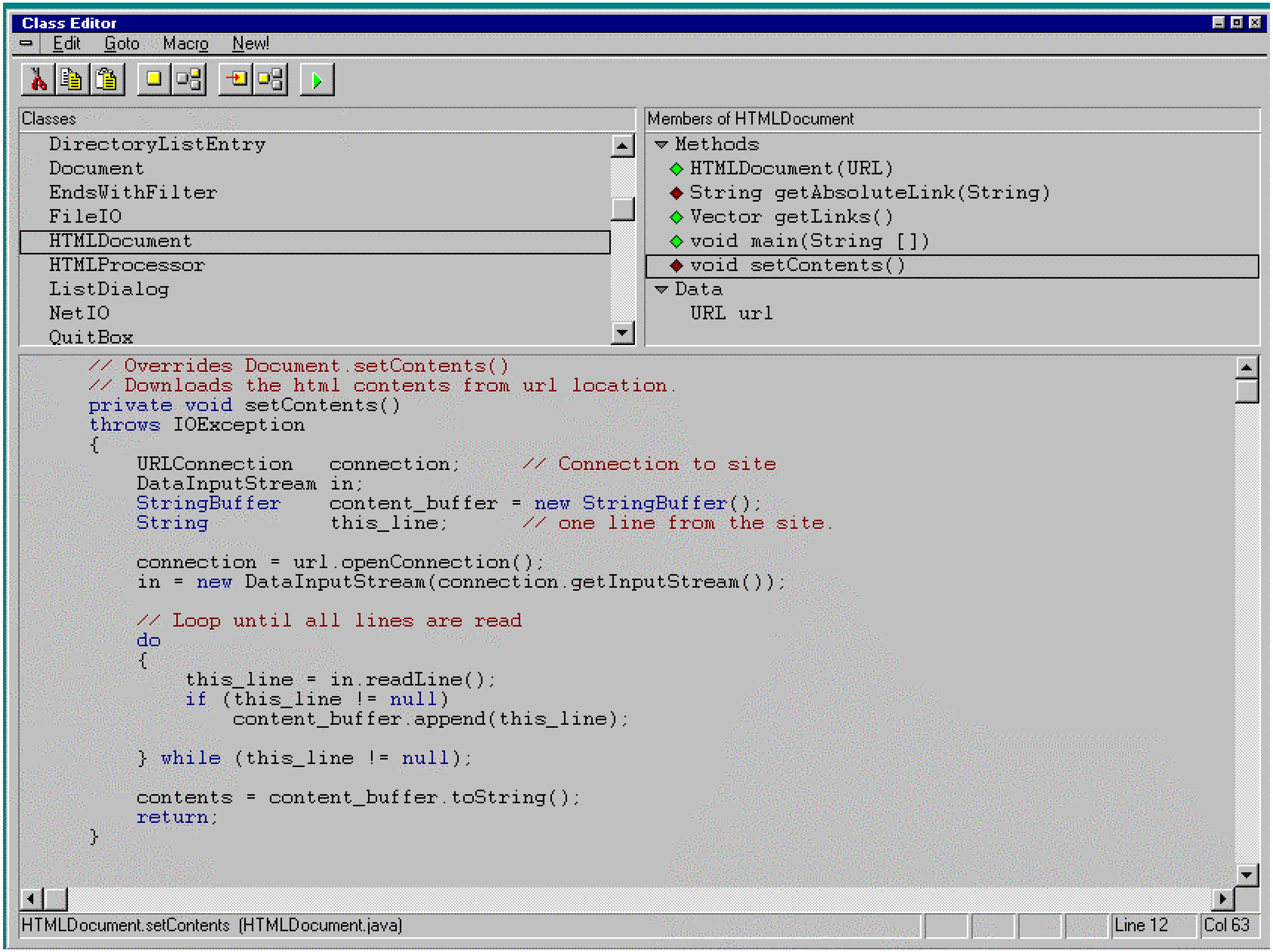
bookmark file

report file

ApplicationFrame.startCheck()

```
 // Start up the new checkSite thread.
        thread_state = THREADINPROGRESS;
        reportToUser("Checking " + EnterURL.getText());
        ContinueButton.disable();
        report = new Report();
        report_site = EnterURL.getText();
        checkSites  = new CheckSites(this, report, filter,
recurse_item.getState());
        try {
            // See if the location is an existing file or makes a valid URL.
            setFirstLocationToCheck();
        }
        catch (FileNotFoundException e1) {
            /* Couldn't set the first location to check. The
               error message was already generated inside
               setFirstLocationToCheck() */
            thread_state = THREADNOTCHECKING;
            return;
        }
        checkSites.start();  // Start the thread.
        PauseButton.enable();
```

HTMLDocument class



```java
    // Overrides Document.setContents()
    // Downloads the html contents from url location.
    private void setContents()
    throws IOException
    {
        URLConnection   connection;     // Connection to site
        DataInputStream in;
        StringBuffer    content_buffer = new StringBuffer();
        String          this_line;      // one line from the site.

        connection = url.openConnection();
        in = new DataInputStream(connection.getInputStream());

        // Loop until all lines are read
        do
        {
            this_line = in.readLine();
            if (this_line != null)
                content_buffer.append(this_line);

        } while (this_line != null);

        contents = content_buffer.toString();
        return;
    }
```

Parsing HTML

Example:

<A HREF = “http://www.cms-stl.com”> CMS </A>

```
public Vector getLinks()
  {
      // Go through the html code looking for the start_string
      loop: while ( (start_index=html_lower.indexOf(start_string, start_index)) != -1)
      {
          // Increment start_index to just after start_string
          start_index += start_string.length();

          // Find the next </a>
          end_index = html_lower.indexOf(end_string, start_index);
          if (end_index == -1) break loop;  // break out of the main loop

          // Create a new String from the indices
          substring = contents.substring(start_index, end_index);

          // Tokenize the string
          tokenizer = new StringTokenizer(substring);
```

```
// Go through the tokens of the string
        try
        {
           /* The first element must be the tag and can only be bounded
             by spaces and an equals sign */
           if (tag.compareTo(tokenizer.nextToken(" \t\n\r=").toLowerCase())!=0)
              continue loop; // This can't be a link

           // Get the URL address. It must be in quotes.
           new_site  = tokenizer.nextToken(" =\"");
        }
        catch (NoSuchElementException e1)
        {
           // This can't be a valid link so go to next attempt
           continue loop;
        }

        try
        {
           // This is just to get past the " we are currently sitting on
           tokenizer.nextToken("><");
           // Get the label.
           new_label = tokenizer.nextToken("><");
        }
        catch (NoSuchElementException e1)
        {
           // It's okay if there is no label;
           new_label = null;
        }
```

```
// Filter unwanted links and add the information
            if (!new_site.toLowerCase().startsWith("gopher://") &&
                !new_site.toLowerCase().startsWith("mailto:") &&
                !new_site.toLowerCase().startsWith("ftp://") &&
                !new_site.toLowerCase().startsWith("news:") &&
                !new_site.toLowerCase().startsWith("newsrc:") &&
                !new_site.toLowerCase().startsWith("nntp://") &&
                !new_site.toLowerCase().startsWith("telnet://") &&
                !new_site.toLowerCase().startsWith("wais://")
               )
            {
                // Get the absolute link for new_site
                new_absolute_site = getAbsoluteLink(new_site);
                this_URL = new URLInfo(new_site, new_absolute_site, new_label);
                extracted_links.addElement(this_URL);
            }
        }

    return extracted_links;
    }
```

Getting absolute links

- The ideal world

  <A HREF = “http://www.cms-stl.com”> CMS </A>

  <A HREF = “development.html”>Development</A>

  Can use URL(URL, String) to create the absolute link.

- The not-so-ideal world

  <A HREF = “http://www.cms-stl.com/eng”> CMS </A>

  <A HREF = “development.html”>Development</A>

  URL(URL, String) creates

  http://www.cms-stl.com/development.html

  but we wanted

  http://www.cms-stl.com/eng/development.html

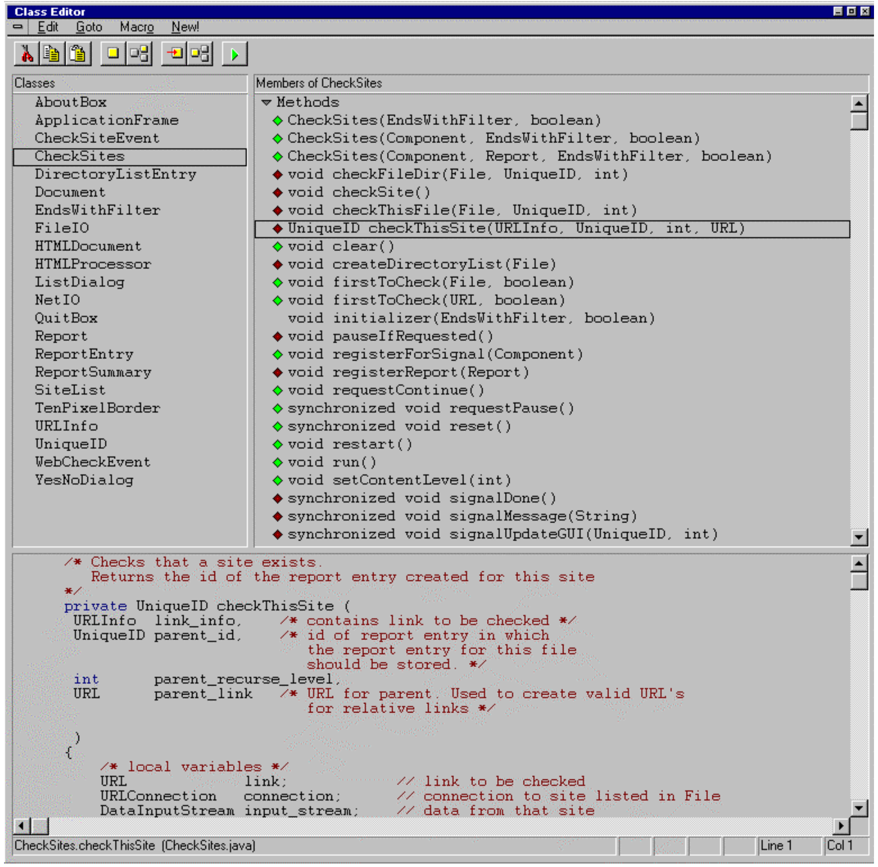
Getting absolute links (continued)

```
private String getAbsoluteLink(String link) {
        String parent_link = url.toExternalForm(); // String version of parent
        StringBuffer new_link; // full form of relative link

        // Check to see if link is already a full path
        if (link.toLowerCase().startsWith("gopher://") ||
            link.toLowerCase().startsWith("mailto:")   ||
            link.toLowerCase().startsWith("ftp://")    ||
            link.toLowerCase().startsWith("news:")     ||
            link.toLowerCase().startsWith("newsrc:")   ||
            link.toLowerCase().startsWith("nntp://")   ||
            link.toLowerCase().startsWith("telnet://") ||
            link.toLowerCase().startsWith("wais://")   ||
            link.toLowerCase().startsWith("file:/")    ||
            link.toLowerCase().startsWith("http://")
           )
             return link;
```

Getting absolute links (continued)

```
// Determine if the parent link ends with a filename.
      if (parent_link.toLowerCase().endsWith(".htm")   || parent_link.toLowerCase().endsWith(".htm/")  ||
         parent_link.toLowerCase().endsWith(".html")   || parent_link.toLowerCase().endsWith(".html/")
        ) {
         /* The parent link ends with a file so remove the last character to get rid of a "/" if it's there
            and search for the last occurrence of "/" */
         int last_index = parent_link.lastIndexOf("/", parent_link.length()-2);

         // Form a new string using the substring up to the "/"
         return parent_link.substring(0, last_index+1) + link;
      }  else  {
         /* The link ends with a directory so append the
            relative link to the end of the parent URL
         */
         if (parent_link.endsWith("/") && link.startsWith("/"))
              return parent_link + link.substring(1);
         else if (parent_link.endsWith("/") || link.startsWith("/"))
            return parent_link+link;
         else
            return parent_link + "/" + link;
      }
   }
```

CheckSites class



```
AboutBox
ApplicationFrame
CheckSiteEvent
CheckSites
DirectoryListEntry
Document
EndsWithFilter
FileIO
HTMLDocument
HTMLProcessor
ListDialog
NetIO
QuitBox
Report
ReportEntry
ReportSummary
SiteList
TenPixelBorder
URLInfo
UniqueID
WebCheckEvent
YesNoDialog
```

```
▽ Methods
  ◇ CheckSites(EndsWithFilter, boolean)
  ◇ CheckSites(Component, EndsWithFilter, boolean)
  ◇ CheckSites(Component, Report, EndsWithFilter, boolean)
  ◆ void checkFileDir(File, UniqueID, int)
  ◆ void checkSite()
  ◆ void checkThisFile(File, UniqueID, int)
  ◆ UniqueID checkThisSite(URLInfo, UniqueID, int, URL)
  ◇ void clear()
  ◆ void createDirectoryList(File)
  ◇ void firstToCheck(File, boolean)
  ◇ void firstToCheck(URL, boolean)
    void initializer(EndsWithFilter, boolean)
  ◆ void pauseIfRequested()
  ◇ void registerForSignal(Component)
  ◆ void registerReport(Report)
  ◇ void requestContinue()
  ◇ synchronized void requestPause()
  ◇ synchronized void reset()
  ◇ void restart()
  ◇ void run()
  ◇ void setContentLevel(int)
  ◆ synchronized void signalDone()
  ◆ synchronized void signalMessage(String)
  ◆ synchronized void signalUpdateGUI(UniqueID, int)
```

```
    /* Checks that a site exists.
       Returns the id of the report entry created for this site
    */
    private UniqueID checkThisSite (
     URLInfo  link_info,    /* contains link to be checked */
     UniqueID parent_id,    /* id of report entry in which
                               the report entry for this file
                               should be stored. */
     int      parent_recurse_level,
     URL      parent_link   /* URL for parent. Used to create valid URL's
                               for relative links */

     )
    {
        /* local variables */
        URL             link;             // link to be checked
        URLConnection   connection;       // connection to site listed in File
        DataInputStream input_stream;     // data from that site
```

CheckSites.checkThisSite (CheckSites.java)   Line 1   Col 1

CheckSite()
requestPause()
requestContinue()
pauseIfRequested()
checkThisSite()
checkThisFile()
checkFileDir()
firstToCheck(URL)
firstToCheck(File)
signalUpdateGUI()
registerForSignal()
run()

CheckSite.checkThisSite()

```
// First, pause if we need to
        pauseIfRequested();

        // Tell the registered Components what we are doing
        signalMessage("Checking " + link_info.getLink());

        // Create the link
        try {
            link = new URL(parent_link, link_info.getAbsoluteLink());
        }
        catch (MalformedURLException e1) {
            // Create error report
            id = report.add(parent_id, ReportEntry.ERROR, link_info.getLink(),
                            link_info.getLabel(), "Invalid link", null);
            // Now signal the GUI to update
            signalUpdateGUI(id, recurse_level);
            return id;
        }
```

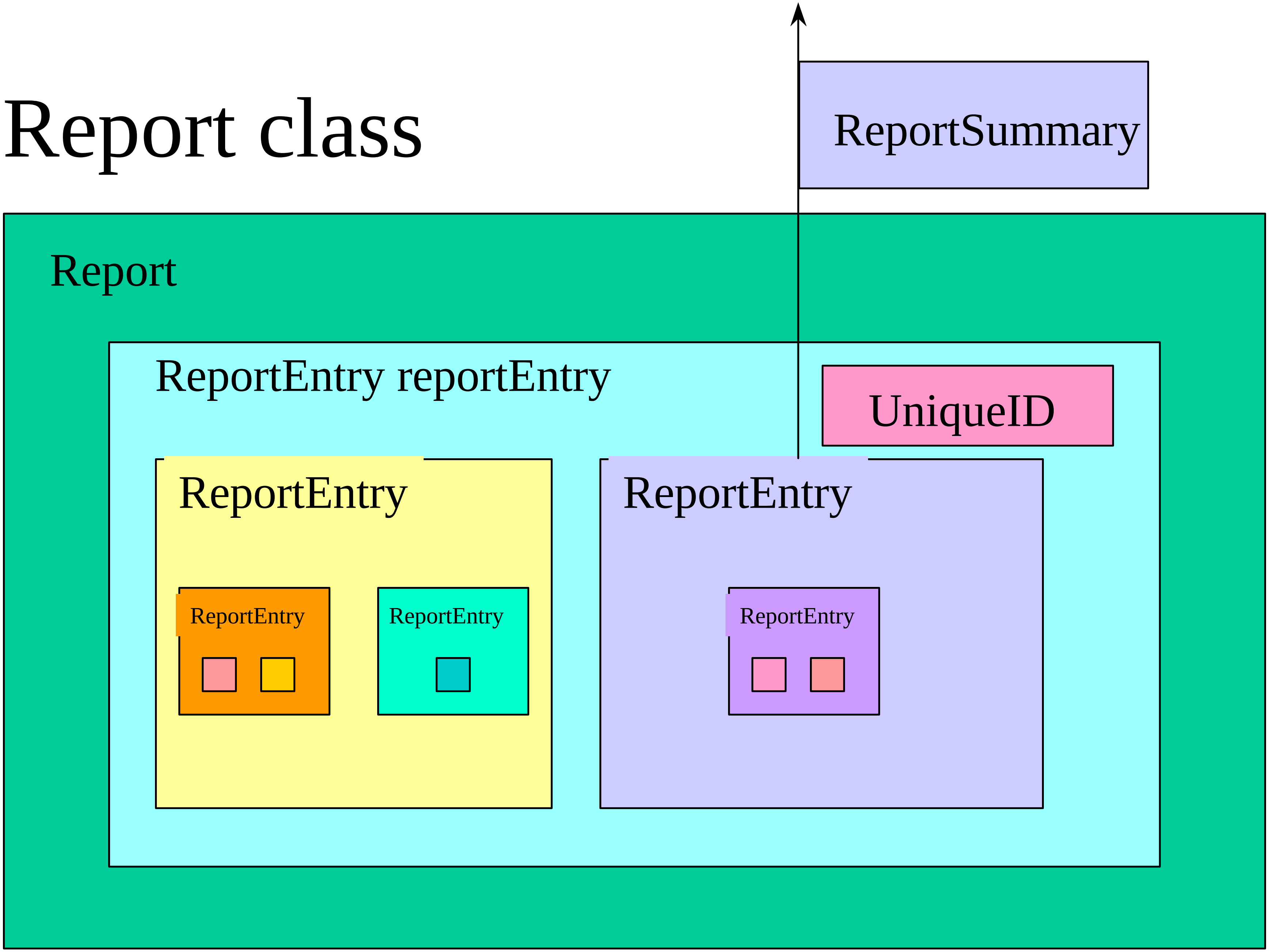
```
try {
      connection = link.openConnection();  // Open connection
      input_stream = new DataInputStream(connection.getInputStream());
      contents[0] = "Last modified: " + connection.getLastModified();
      contents[1] = "Content length: "+ connection.getContentLength();
      // read first lines
      for(int i=0; i<contentLevel; i++) {
        contents[i+2] = input_stream.readLine();
      }

      // Add to the report
      id = report.add(parent_id, ReportEntry.LOG, link, link_info.getLabel(),
              "Okay", contents);

      // Now signal the GUI to update
      signalUpdateGUI(id, recurse_level);
      }
    catch (IOException e2) {
      // Create error report
      id = report.add(parent_id, ReportEntry.ERROR, link, link_info.getLabel(),
              "Cannot access", contents);
      // Now signal the GUI to update
      signalUpdateGUI(id, recurse_level);
      return id;
    }
```

```
public void registerForSignal(Component newComponent)
  {
    componentsToSignal.addElement(newComponent);
    return;
  }

private synchronized void signalUpdateGUI(UniqueID id, int indent_level)
  {
    Component tempComponent; // For looping and typecasting.
    CheckSiteEvent newEvent =
         new CheckSiteEvent(this,CheckSiteEvent.SUMMARY, id,
                                      indent_level);
    for (int i=0; i<componentsToSignal.size(); i++) {
      tempComponent = (Component)componentsToSignal.elementAt(i);
      tempComponent.deliverEvent(newEvent);
    }
    yield();  // Allows Macintosh to update
    return;
  }
```

UniqueID class

```
class UniqueID
    {
    public boolean isEqual(UniqueID otherID)
    {
        if (this.this_counter == otherID.this_counter)
            return true;
        else
            return false;
    }

    private static int counter = 0;
    private final int this_counter = counter;
    // Constructor
    public UniqueID ()
    {
        counter++;
    }
    }
```

Report class



ReportSummary

Report

ReportEntry reportEntry

UniqueID

ReportEntry

ReportEntry

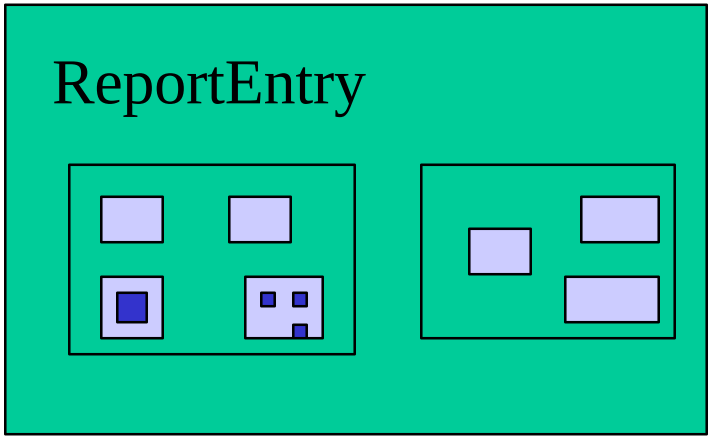
ReportEntry

ReportEntry

ReportEntry

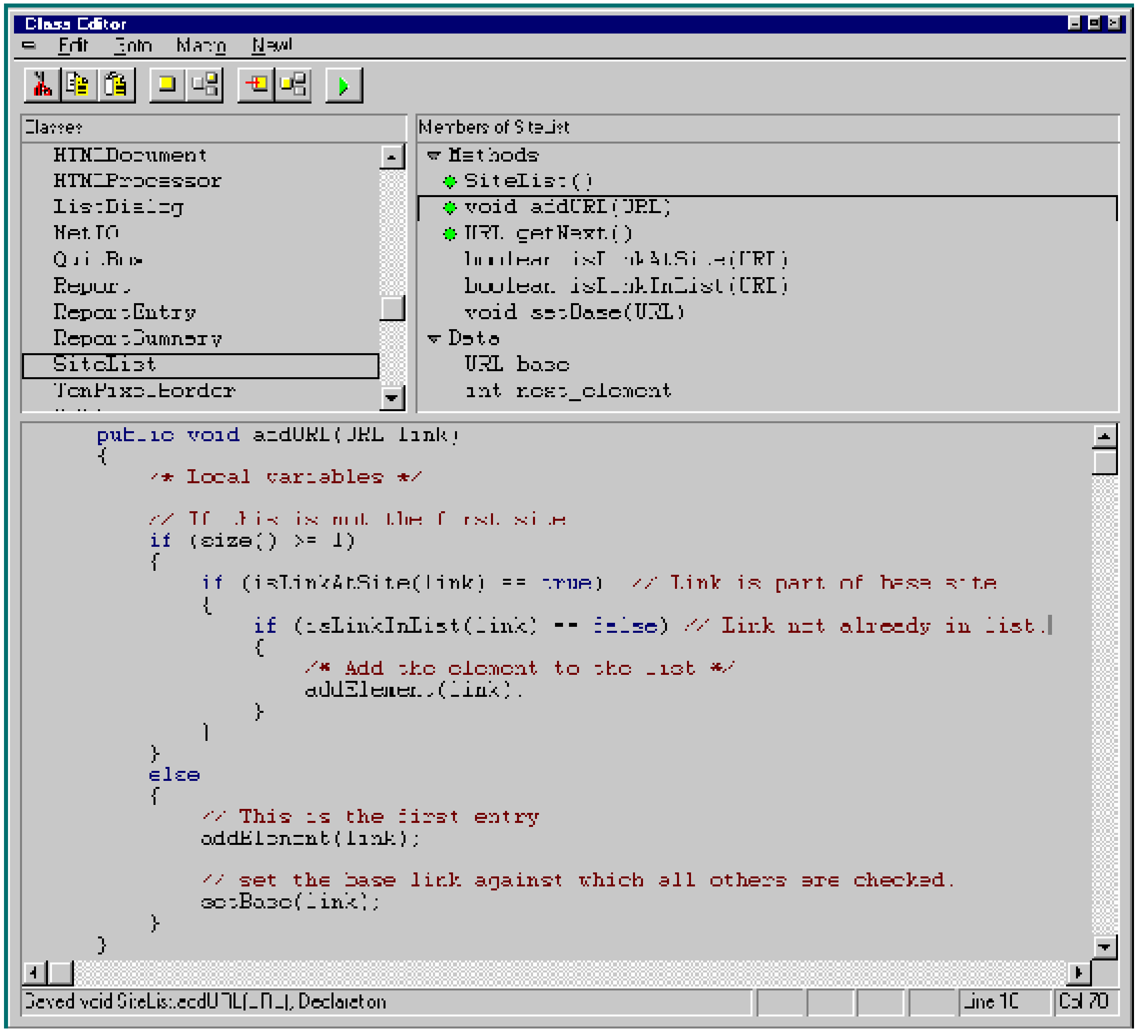
ReportEntry class

```
public ReportEntry GetEntry(UniqueID id) {
    ReportEntry reportInList;    // the next subreport in the list.
    ReportEntry requestedReport; // Report with Unique id matching id.
    Enumeration list;            // Enumeration of all subreports

    // First check if this entry is the one
    if (id.isEqual(this.getId()))   return this;

    // Check all subReports
    list = subReports.elements();
    while(list.hasMoreElements()) {
        reportInList = (ReportEntry)list.nextElement();
        requestedReport = reportInList.GetEntry(id);
        if (requestedReport != null)  return requestedReport;
    }
    // id must not be in this ReportEntry
    return (ReportEntry)null;
  }
```



SiteList class

extends Vector



```
    public void addURL(URL link)
    {
        /* Local variables */

        // If this is not the first site
        if (size() >= 1)
        {
            if (isLinkAtSite(link) == true)  // Link is part of base site
            {
                if (isLinkInList(link) == false) // Link not already in list.
                {
                    /* Add the element to the list */
                    addElement(link);
                }
            }
        }
        else
        {
            // This is the first entry
            addElement(link);

            // set the base link against which all others are checked.
            setBase(link);
        }
    }
```

Saved void SiteList.addURL(URL), Declaration | Line 10 | Col 70

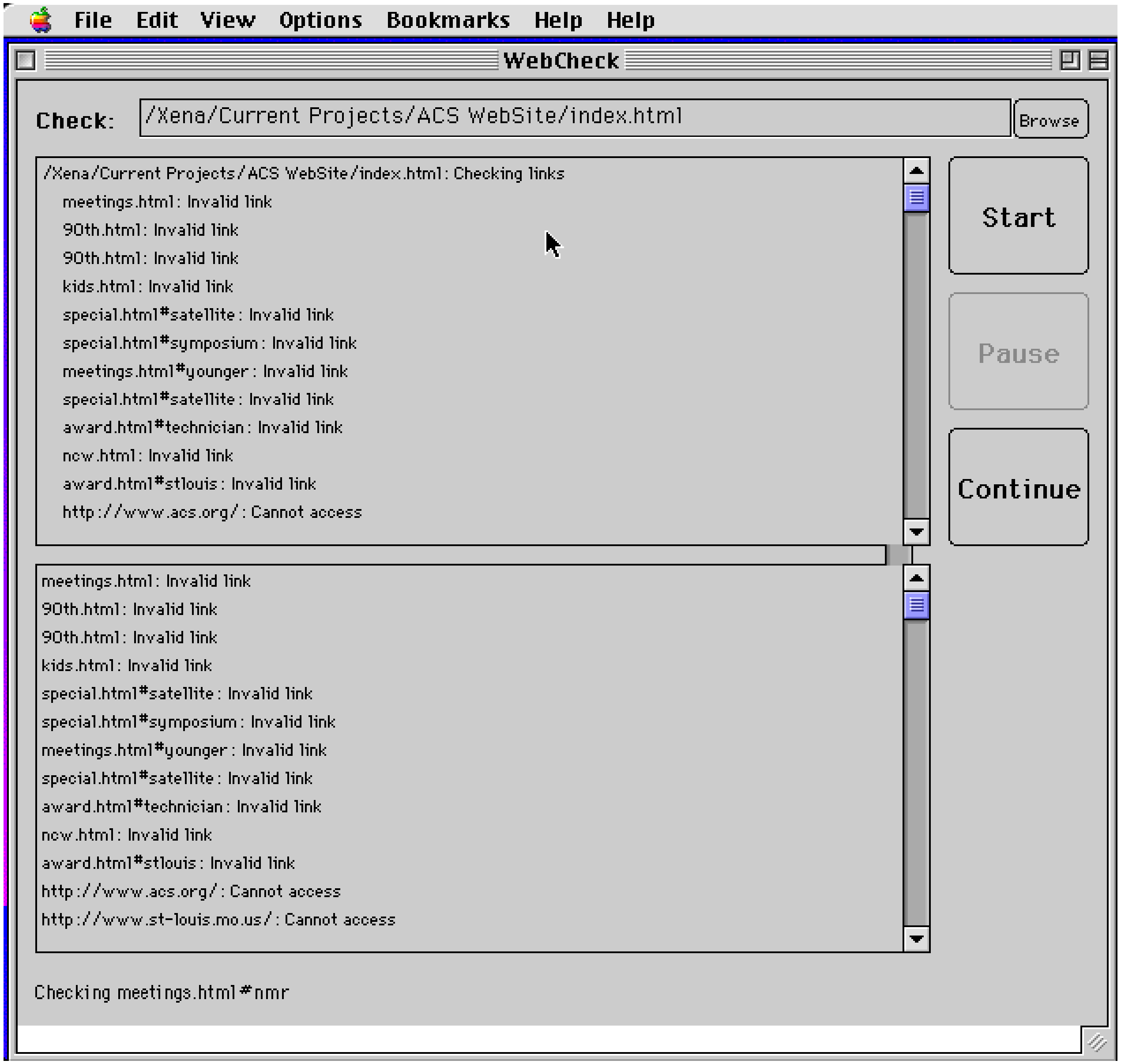
URL base

setBase()
isLinkInList()
getNext()
addURL()

SiteList.AddURL

```
public void addURL(URL link) {
    // If this is not the first site
    if (size() >= 1) {
        if (isLinkAtSite(link) == true) { // Link is part of base site.
            if (isLinkInList(link) == false) // Link not already in list.
                addElement(link); /* Add the element to the list */
        }
    } else {
        addElement(link); // This is the first entry
        // set the base link against which all others are checked.
        setBase(link);
    }
 }
```

The GUI on the Macintosh



Check: /Xena/Current Projects/ACS WebSite/index.html

```
/Xena/Current Projects/ACS WebSite/index.html: Checking links
   meetings.html: Invalid link
   90th.html: Invalid link
   90th.html: Invalid link
   kids.html: Invalid link
   special.html#satellite: Invalid link
   special.html#symposium: Invalid link
   meetings.html#younger: Invalid link
   special.html#satellite: Invalid link
   award.html#technician: Invalid link
   new.html: Invalid link
   award.html#stlouis: Invalid link
   http://www.acs.org/: Cannot access
```

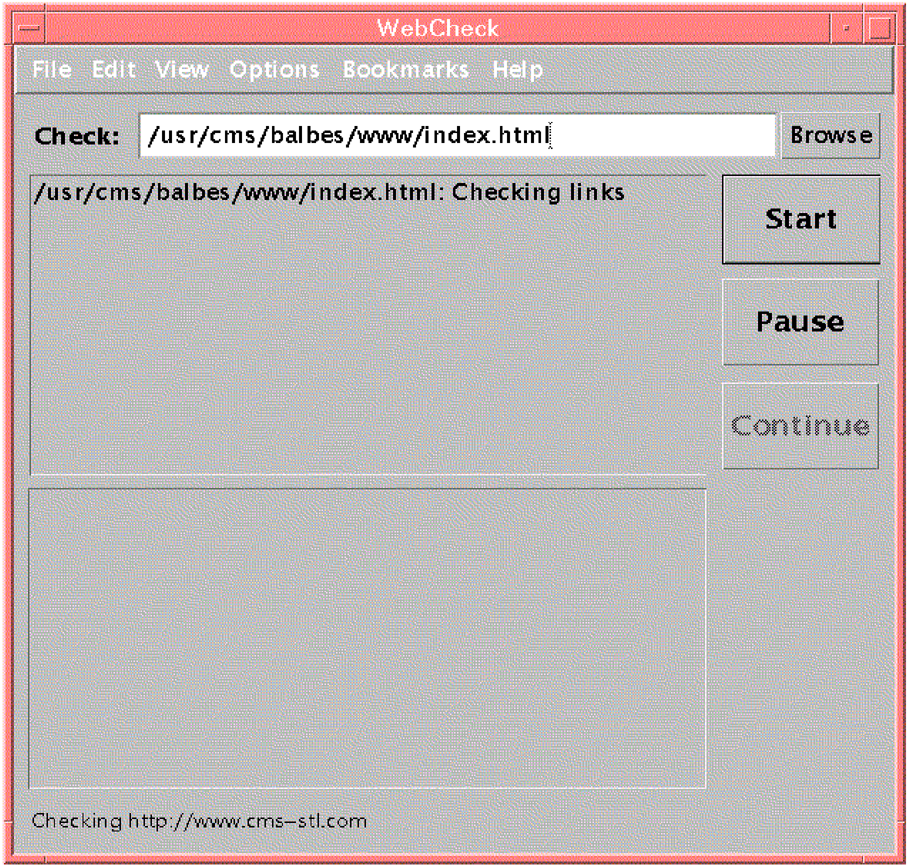
```
meetings.html: Invalid link
90th.html: Invalid link
90th.html: Invalid link
kids.html: Invalid link
special.html#satellite: Invalid link
special.html#symposium: Invalid link
meetings.html#younger: Invalid link
special.html#satellite: Invalid link
award.html#technician: Invalid link
new.html: Invalid link
award.html#stlouis: Invalid link
http://www.acs.org/: Cannot access
http://www.st-louis.mo.us/: Cannot access
```

Checking meetings.html#nmr

Porting to the Macintosh

- cooperative vs. pre-emptive threads
- Macintosh SDK behaves differently than Visual Café.

The GUI on the HP B132



Check: /usr/cms/balbes/www/index.html

/usr/cms/balbes/www/index.html: Checking links

Checking http://www.cms-stl.com

Porting to the HP

- Getting errors from X.

  Warning:

  Name: slist

  Class: XmList

  Invalid item(s) to delete

- Can’t get out of company firewall because of the proxy server.

Current problems

- Cut and Paste
- Proxy servers
- Is the log file useful?
- Displaying HTML in a web page w/o having it parsed
- Detecting when the server generates a page that explains that the link is not available or moved instead of generating an error.
- Relative links for files. Creating the absolute link is system dependent.