An Ontology-based HTML to XML Conversion Using Intelligent Agents

Thomas E. Potok, Mark T. Elmore, Joel W. Reed, and Nagiza F. Samatova
Oak Ridge National Laboratory
Post Office Box 2008, Mail Stop 6414
Computer Science and Mathematics Division
Oak Ridge, Tennessee 37831-6414
Phone: 865-574-0834
Fax: 865-241-6211
potokte@ornl.gov

Abstract

How to organize and classify large amounts of heterogeneous information accessible over the Internet is a major problem faced by industry, government, and military organizations. XML is clearly a potential solution to this problem, [1,2] however, a significant challenge is how to automatically convert information currently expressed in a standard HTML format to an XML format. Within the Virtual Information Processing Agent Research (VIPAR)[1] *project, we have developed a process using Internet ontologies and intelligent software agents to perform automatic HTML to XML conversion for Internet newspapers. The VIPAR software is based on a number of significant research breakthroughs. Most notably, the ability for intelligent agents to use a flexible RDF ontology to transform HTML documents to XML tagged documents. The VIPAR system is currently deployed at the US Pacific Command, Camp Smith, HI, traversing up to 17 Internet newspapers daily.*

1. Introduction

One of the key successes of information technology is the ability to bring more information to people faster than ever before. This however, has spawned the new challenge of what to do with this massive amount of new information. Most of this information is tagged with HTML tags, which provide the ability to view this information, but not to organize it. For example, there is no way to automatically determine the author of a document, the abstract, or the conclusion. XML has the ability to solve this problem by providing a structure to information contained in an HTML document. Unfortunately, there is no simple way of converting an HTML marked document to XML. Until this challenge can be solved, the expensive task of manually rewriting HTML documents to XML will be required.

This paper describes the challenge faced by the US Pacific Command's Virtual Information Center (VIC). The VIC is responsible for reading open source information, typically from the Internet, organizing this information, and summarizing the results. They normally read twenty to thirty Internet newspapers a day. From this information, a daily summary is produced that allows the military staff to quickly absorb the day's news and developments. Manually reading and organizing this volume of information requires a significant investment of time. We were asked to look at ways that this information could be automatically gathered and organized so that analysts could then proceed with summarizing this information, rather than spending time collecting it.

To automatically gather and organize articles from Internet newspapers involves a number of steps. Of

[1] The Virtual Information Processing Agent Research (VIPAR) is a software research and development project funded by the Office of Naval Research (ONR) via Interagency Agreement 2302-Q326-A1 with the Department of Energy's Oak Ridge National Laboratory (ORNL). The submitted manuscript has been authored by a contractor of the U.S. Government under contract No. DE-AC05-96OR22464. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government Purposes.

relevance to this paper is the process for converting HTML information to XML. This process is a hard problem to solve based on several issues. There is no single, uniform structure existing across all or most Internet newspapers, which would allow a simple conversion from the HTML format to the desired XML format. Not surprisingly, the various Internet newspapers have no common structure in the organization of the site's articles, i.e., what subdirectories contain current news articles, or in the application of the HTML formatting, i.e., what tags, if any, are used to wrap the title of a document. Somewhat surprisingly, even within a single newspaper's site, the structure may not be consistent. The main issue becomes defining a common description that allows disparate HTML pages to be converted to XML in a consistent way. There are five basic elements of this description: 1) *article metadata* - metadata about retrieved articles used to capture information about what has been retrieved, 2) *traversal directives* - site-specific actions on how a site should be traversed, 3) *traversal maps* - a map of an Internet newspaper site that contains pages of interest, 4) *article delimiters* - markers to delimit the text of an article from other information on the web page, and 5) *article structuring rules* - rules for structuring the article text as XML. This description becomes the ontology that we use in performing the HTML to XML conversion.

The HTML to XML conversion is based on ontologies, one for each newspaper site. These ontologies can be represented as directed graphs characterizing the five basic elements enumerated above. The ontological representation is described using the Resource Description Framework (RDF). In the VIPAR system, literally thousands of news articles are converted from HTML to XML on a daily basis. It is this powerful ontological approach that allows VIPAR to robustly perform HTML to XML conversions on the varied Internet newspaper sites.

The paper is organized as follows, a background description of the research in the area, followed by the approach taken to address this problem, then the results of our work is presented, followed by a discussion, and finally some thoughts on future directions and conclusion.

2. Background

Conversion from HTML to XML is a challenging problem that has seen much attention from industry and academia. Generally, the notion of a "wrapper" is used to address this challenge. Conceptually, the idea is to create a software layer, a wrapper, which provides a mapping between an existing HTML page, and a desired XML interface. This wrapper is typically software that is written with prior knowledge of the HTML page [3,4]. Adding to these ideas, a parse tree using HTML tags as nodes can be used for identifying parts of a web page and converting those parts into XML [5]. A limitation with these approaches is that the HTML tagging scheme does not always map cleanly to the desired XML mapping scheme, and that the pages are not converted to XML, but mapped through an interface. For subsequent textual analysis, we need the XML documents to be very rapidly processed which dictates a conversion-based solution. Further, new, previously unseen articles are posted each day at Internet newspapers, and writing individual wrappers, a priori, for each new page is not a viable option. Additionally, the VIC needs to be able to add additional newspapers into the VIPAR system without writing specialty software each time. Thus, a wrapper approach was not suitable for our needs.

Another means of performing this conversion is to use a generic parsing engine that can be driven by an ontology. There has been a great deal of work on ontologies, however, of particular relevance is the current effort in defining the Semantic Web concepts [6,7,8]. The Semantic Web is intended to extend the current Internet by giving well-defined meaning to information to allow better cooperation among computers and people. This is the goal of XML also, but XML’s capabilities are bounded by syntax-based interoperability and the Semantic Web wishes to step up to a higher level of abstraction by using semantic-based interoperability [9]. The Semantic Web is based on the Resource Description Framework (RDF) [10] and the DARPA Agent Markup Language (DAML) [11]. These two technologies are tightly connected, with RDF being foundational to DAML. RDF is a mechanism for using XML to describe an Internet resource as a directed graph. DAML extends RDF's capability with a richer set of concepts that describe more complex relationships than described by RDF alone [12]. In the case of the VIPAR project, the RDF foundation of DAML sufficiently describes the relationships we wished to capture. DAML, as a superset of RDF, suggests our solution may be classified as either an RDF solution or a DAML solution.

3. Approach

We have selected an ontology-based approach to converting HTML documents to XML documents. The concept is that we can use RDF to describe the five key elements (mentioned above and discussed below) of an Internet newspaper site. Each Internet newspaper in the VIPAR system has a human written RDF file associated with it.

We then use software retrieval agents to retrieve information from Internet sites. These agents behave based on two features. The first is the generic ability to parse Internet pages, and the second is the ability to interpret an RDF ontology, allowing the agents to automatically traverse a site, retrieve relevant articles, and convert them to XML. Each of these information agents monitors the newspaper's Internet site watching for new articles that have not yet been processed. Any time a new article is found, the retrieval agent captures the article, formats it, and then posts it to the VIPAR system for further processing.

The ontological description of the site includes the five key elements of information mentioned above:

1) Article metadata - Metadata about the retrieved articles. This meta-information includes the newspaper's name and the collection under which VIPAR classifies the newspaper. A collection is a grouping of newspapers based on geographical region.

2) Traversal directives - site-specific actions for traversing the site. This includes the search depth limit (how many hops) from the root URL, and the number of minutes to wait between rechecking the site for new articles.

3) Traversal maps - maps of an Internet newspaper site containing the pages of interest. The map starts with the root URL from which the agent is to begin a traversal of the site, and from which the agent can resolve relative URLs found at the site. A rule-based map of the pages of interest on the site is based on the URL structure of the site and is encoded via regular expressions.

4) Article delimiters - markers to delimit the text of an article from other information on a given web page. The map of the Internet site includes information used by the retrieval agent to delimit the text of an article from the myriad of other information on the page (boilerplate, banners, advertisements, etc).

5) Article structuring rules - rules for structuring the article text as XML. Again, regular expressions can be used to reduce the various structural characteristics of an article, such as the title, author, and paragraphs.

Based on this RDF ontology, a retrieval agent checks each page link found at an Internet newspaper site against the traversal map to determine if the article page is of interest. Once found, the agent checks with the VIPAR system to verify that the article has not already been incorporated into the VIPAR system. If the article is indeed new, the agent retrieves the page, discerning the actual article text from the article delimiters, and cleaning it of extraneous information on the page. The agent then marks up the clean text using XML, tagging the parts of the article (title, author, date, location, paragraphs, etc) depending on the site's article structuring rules. The article is then posted to the VIPAR system where the article metadata is used to determine what collection it should be added to. The agent continues to monitor the site based on the traversal directives, and posting new information of interest as it becomes available.

4. Results

The ontology-based HTML to XML conversion process is operational within the VIPAR system at the US Pacific Command's Virtual Information Center. The current version is retrieving and processing articles from 13 different newspaper sites selected by the VIC staff. These newspapers are:

1. Asahi Shimbun
2. Asia Times
3. BBC
4. Japan Times Online
5. Japan Update
6. Korea Times
7. Manila Times
8. Pacific Islands Report
9. Sydney Morning Herald
10. Taipei Times
11. The Hindu
12. The Star
13. Times of India

Each newspaper has a unique RDF ontology which is assigned to a generic retrieval agent.

We have described the five key elements of a web page that we need to convert it to XML. Now we look in detail at how these elements are expressed in RDF ontologies. To describe the results of converting HTML pages to XML, we will go through a detailed example of the process. We begin this by describing the layout of a site's ontology as represented in an RDF file. We are going to use the ontology defined for the Pacific Islands Report[2] (PIR), a Hawai‘i-based newspaper focusing on news from the Pacific Islands. The PIR has a clear-cut Internet site that is basic, clean, and well organized, as are its individual Web pages. Thus, it provides a good illustration of our basic approach in practice. We begin by giving an overview of the site.

[2] Used with permission

1.1 PIR overview

Figure 1 shows the overall layout of the PIR site. This site has two levels of interest, the root URL that forms a “table of contents” for the site, and the individual article pages. There are also a number of links that are not of interest, and are thus excluded from processing. For example, pages that do not to conform the URL pattern of “http://pidp.ewc.Hawaii.edu/pireport/…” are excluded from processing, as will be described below. The root is at the URL http://pidp.ewc.hawaii.edu/pireport/. From this, a number of articles are linked, using the date in the path names of the articles, for example, the URL for the first article is http://pidp.ewc.hawaii.edu/pireport/2001/June/06-05-01.htm, where the last number, 01, represents the article number for that day. On this day, there were 26 articles. On other sites we have used, it is quite likely to have several tables of contents of articles. For example, one may contain local news, while another contains state news, and yet another contains national news. Next, we look at the HTML for a typical news article from this newspaper. Again, these pages are quite straightforward as compared to some of the other more complex sites, such as the BBC. The HTML in Figure 2 shows HTML tags for formatting the page, then the text of the article itself, followed by more formatting tags. The HTML tags do not provide any structuring of the article text; it merely changes the display of the text. Without understanding the content of the page, there is no way to automatically determine what the title of the article is or who wrote it.

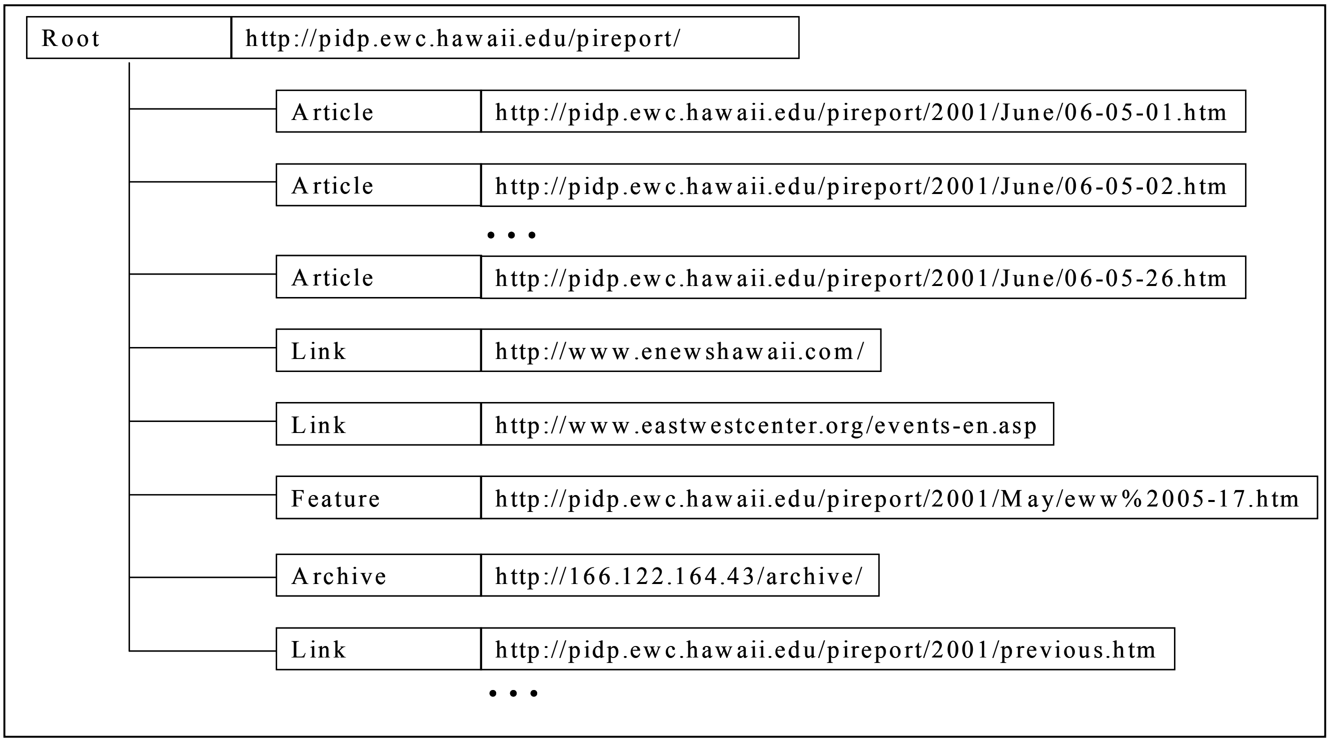
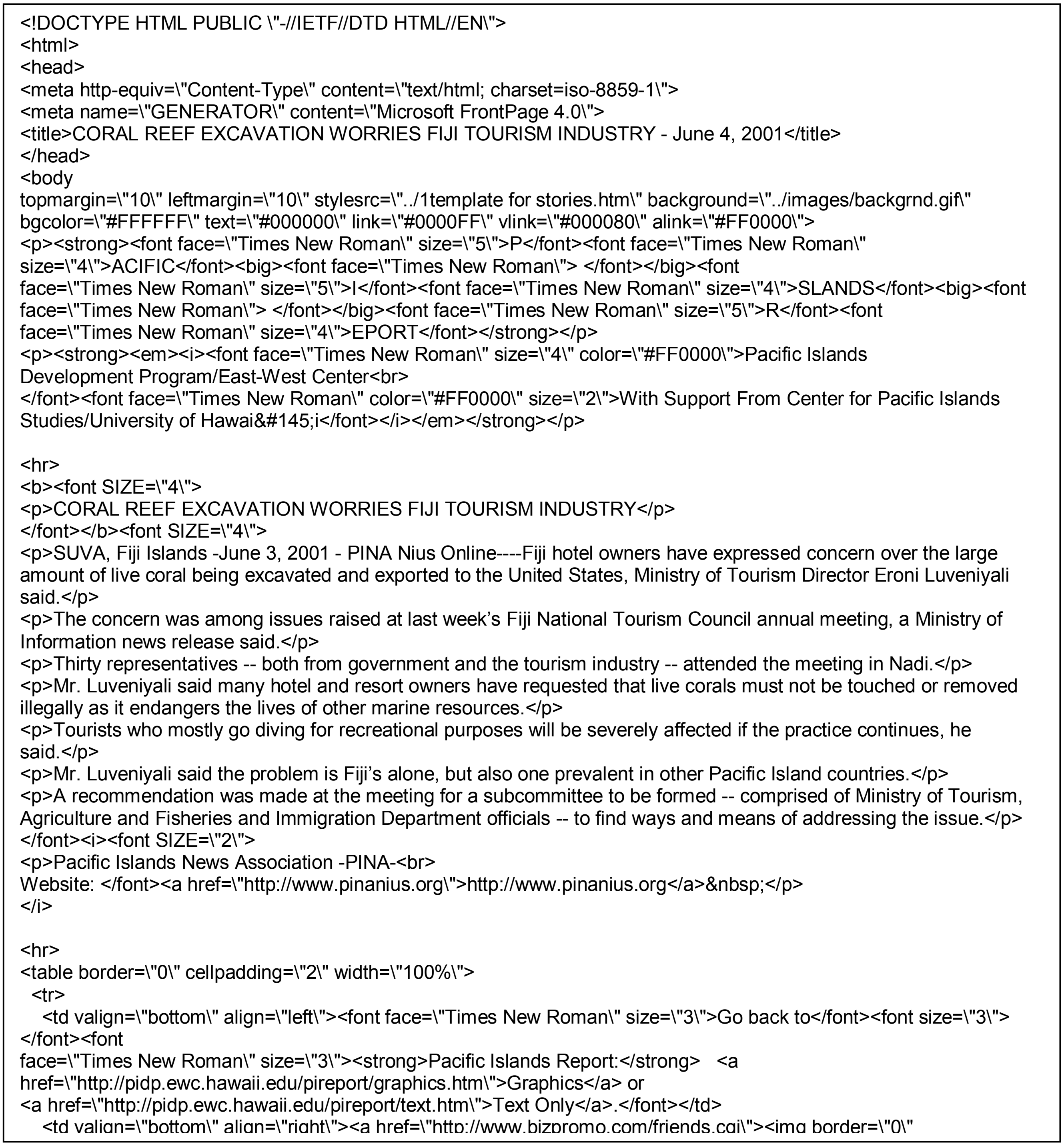

Root
http://pidp.ewc.hawaii.edu/pireport/
Article
http://pidp.ewc.hawaii.edu/pireport/2001/June/06-05-01.htm
Article
http://pidp.ewc.hawaii.edu/pireport/2001/June/06-05-02.htm
• • •
Article
http://pidp.ewc.hawaii.edu/pireport/2001/June/06-05-26.htm
Link
http://www.enewshawaii.com/
Link
http://www.eastwestcenter.org/events-en.asp
Feature
http://pidp.ewc.hawaii.edu/pireport/2001/May/eww%2005-17.htm
Archive
http://166.122.164.43/archive/
Link
http://pidp.ewc.hawaii.edu/pireport/2001/previous.htm
• • •

Figure 1. The layout of the Pacific Islands Report newspaper

The generated XML document is shown in Figure 3, the file contains a significant amount of information beyond that merely stored within the article text. For example, the time stamp of when the article was retrieved, the ontology metadata information, the raw HTML, the clean text, as well as the actual text of the article marked up in XML.

The key point is that the unstructured HTML document has been automatically converted to a structured XML document that contains a variety of useful information. Software programs and agents can then readily process this information. The XML representation in Figure 3 can be used to display the article contents within a Web browser using style sheets. Likewise, the article is structured, so that queries and searches can be performed over the XML tags. Now we look in detail at the RDF ontology.



```
<!DOCTYPE HTML PUBLIC \"-//IETF//DTD HTML//EN\">
<html>
<head>
<meta http-equiv=\"Content-Type\" content=\"text/html; charset=iso-8859-1\">
<meta name=\"GENERATOR\" content=\"Microsoft FrontPage 4.0\">
<title>CORAL REEF EXCAVATION WORRIES FIJI TOURISM INDUSTRY - June 4, 2001</title>
</head>
<body
topmargin=\"10\" leftmargin=\"10\" stylesrc=\"../1template for stories.htm\" background=\"../images/backgrnd.gif\"
bgcolor=\"#FFFFFF\" text=\"#000000\" link=\"#0000FF\" vlink=\"#000080\" alink=\"#FF0000\">
<p><strong><font face=\"Times New Roman\" size=\"5\">P</font><font face=\"Times New Roman\"
size=\"4\">ACIFIC</font><big><font face=\"Times New Roman\"> </font></big><font
face=\"Times New Roman\" size=\"5\">I</font><font face=\"Times New Roman\" size=\"4\">SLANDS</font><big><font
face=\"Times New Roman\"> </font></big><font face=\"Times New Roman\" size=\"5\">R</font><font
face=\"Times New Roman\" size=\"4\">EPORT</font></strong></p>
<p><strong><em><i><font face=\"Times New Roman\" size=\"4\" color=\"#FF0000\">Pacific Islands
Development Program/East-West Center<br>
</font><font face=\"Times New Roman\" color=\"#FF0000\" size=\"2\">With Support From Center for Pacific Islands
Studies/University of Hawai&#145;i</font></i></em></strong></p>

<hr>
<b><font SIZE=\"4\">
<p>CORAL REEF EXCAVATION WORRIES FIJI TOURISM INDUSTRY</p>
</font></b><font SIZE=\"4\">
<p>SUVA, Fiji Islands -June 3, 2001 - PINA Nius Online----Fiji hotel owners have expressed concern over the large
amount of live coral being excavated and exported to the United States, Ministry of Tourism Director Eroni Luveniyali
said.</p>
<p>The concern was among issues raised at last week’s Fiji National Tourism Council annual meeting, a Ministry of
Information news release said.</p>
<p>Thirty representatives -- both from government and the tourism industry -- attended the meeting in Nadi.</p>
<p>Mr. Luveniyali said many hotel and resort owners have requested that live corals must not be touched or removed
illegally as it endangers the lives of other marine resources.</p>
<p>Tourists who mostly go diving for recreational purposes will be severely affected if the practice continues, he
said.</p>
<p>Mr. Luveniyali said the problem is Fiji’s alone, but also one prevalent in other Pacific Island countries.</p>
<p>A recommendation was made at the meeting for a subcommittee to be formed -- comprised of Ministry of Tourism,
Agriculture and Fisheries and Immigration Department officials -- to find ways and means of addressing the issue.</p>
</font><i><font SIZE=\"2\">
<p>Pacific Islands News Association -PINA-<br>
Website: </font><a href=\"http://www.pinanius.org\">http://www.pinanius.org</a>&nbsp;</p>
</i>

<hr>
<table border=\"0\" cellpadding=\"2\" width=\"100%\">
  <tr>
    <td valign=\"bottom\" align=\"left\"><font face=\"Times New Roman\" size=\"3\">Go back to</font><font size=\"3\">
</font><font
face=\"Times New Roman\" size=\"3\"><strong>Pacific Islands Report:</strong>   <a
href=\"http://pidp.ewc.hawaii.edu/pireport/graphics.htm\">Graphics</a> or
<a href=\"http://pidp.ewc.hawaii.edu/pireport/text.htm\">Text Only</a>.</font></td>
    <td valign=\"bottom\" align=\"right\"><a href=\"http://www.bizpromo.com/friends.cgi\"><img border=\"0\"
```

Figure 2. The unaltered HTML for an article at the Pacific Island Report web site

```
<article>
  <fileBuildTimeMilliSec>
    991680761171
  </fileBuildTimeMilliSec>
  <downloadDate>
    <year> 2001 </year>
    <month> Jun </month>
    <day> 4 </day>
  </downloadDate>
  <articleURL>  http://pidp.ewc.hawaii.edu/pireport/2001/June/06-04-05.htm  </articleURL>
  <collection>   Pacific  </collection>
  <newspaperName>  Pacific Islands Report  </newspaperName>
  <articleParentURL>  http://pidp.ewc.hawaii.edu/pireport/graphics.htm  </articleParentURL>
  <articleRootURL>  http://pidp.ewc.hawaii.edu/pireport/   </articleRootURL>
  <articleDepthFromRoot>  2   </articleDepthFromRoot>
  <articleContentEncoding>    null   </articleContentEncoding>
  <articleContentType>   text/html   </articleContentType>
  <articleDate>   991680957000  </articleDate>
  <articleExpiration>   0  </articleExpiration>
  <articleLastMod>   991628284000  </articleLastMod>
  <articleRawHTML>

            … (omitted for the figure)

  </articleRawHTML>
  <rdfFileName>
    C:\Program Files\ORNL\VIPAR Server V3.0\VIPARServer\DownloadAgent\Rdf\pireport.rdf
  </rdfFileName>
  <articleCleanText>

            … (omitted for the figure)

  </articleCleanText>
  <xmlMarkedUpText>
    <newspaperName>  Pacific Islands Report   </newspaperName>
    <url>   http://pidp.ewc.hawaii.edu/pireport/2001/June/06-04-05.htm  </url>
    <title>  CORAL REEF EXCAVATION WORRIES FIJI TOURISM INDUSTRY   </title>
    <city>  SUVA, Fiji Islands  </city>
    <date>  June 3, 2001  </date>
    <newsService> - PINA Nius Online  </newsService>
    <paragraph number=”1”>
      Fiji hotel owners have expressed concern over the large amount of live coral being excavated and exported
to the United States, Ministry of Tourism Director Eroni Luveniyali said.
    </paragraph>
    <paragraph number=”2”>
      The concern was among issues raised at last week s Fiji National Tourism Council annual meeting, a
Ministry of Information news release said.
    </paragraph>
          …
    <paragraph number=”7”>
      A recommendation was made at the meeting for a subcommittee to be formed -- comprised of Ministry of
Tourism, Agriculture and Fisheries and Immigration Department officials -- to find ways and means of addressing
the issue.
    </paragraph>
    <paragraph number=”8”>
      Pacific Islands News Association -PINA-
 Website: http://www.pinanius.org
    </paragraph>
  </xmlMarkedUpText>
```

Figure 3. The XML translation of the Pacific Islands Report article

```
<? xml version="1.0" ?>
<rdf:RDF xmlns:ORNL = "http://csm.ornl.gov/VIPAR">

 <rdf:Description about = "http://pidp.ewc.hawaii.edu/pireport/">

      <ORNL:newspaperName>
           Pacific Islands Report
      </ORNL:newspaperName>

      <ORNL:rootURLStr>
           http://pidp.ewc.hawaii.edu/pireport/
      </ORNL:rootURLStr>

      <ORNL:collection>
           Pacific
      </ORNL:collection>
      <rdf:Description ID="agentDirective">
             <ORNL:searchDepthLimit>
                 2
             </ORNL:searchDepthLimit>
             <ORNL:minutesWaitBetweenDownloadSessions>
                 60
             </ORNL:minutesWaitBetweenDownloadSessions>
      </rdf:Description>
```

Continued on Figure 5

Figure 4. First half, RDF for Pacific Islands Report

1.2 Ontology

Perhaps the best definition of an ontology is an explicit specification of a conceptualization [13]. In our work, the conceptualization is the information that is contained within newspaper Internet sites. Our ontology seeks to specify relationships, metadata, and access information for an agent to be able to process this newspaper information conceptualization. We use the RDF specification to represent this ontology in an attempt to be compliant with a semantic web representation. As an example, the RDF ontology for PIR is presented across Figures 4 and 5. Of the five key elements of this ontological information, 1) article metadata, 2) traversal directives, 3) traversal maps, 4) article delimiters, and 5) article structuring rules, Figure 4, captures the first two elements.

The article metadata includes the <ORNL:newspaperName> tag that contains the name of the newspaper. In this example, it is the "Pacific Islands Report." The <ORNL:rootURLStr> tag contains the root URL of the newspaper site. This is the page from which the agent will begin its traversal of site's contents and is also the base URL used to resolve relative links found within the site. <ORNL:collection> is the tag that describes the VIPAR collection (based on region of the world) to which the articles will be added.

The traversal directives are contained within the <rdf:Description ID="agentDirective"> tag set. These directives include the <ORNL:searchDepthLimit> tag that defines how many nesting levels deep the search is to go. Although this can be used in filtering articles, its main function is as a failsafe measure in the event a search goes awry. For example, it prevents the agent from traversing into an archive, where thousands of old articles may be stored. How often an agent will revisit a given site to check for new articles is controlled by the <ORNL:minutesWaitBetweenDownloadSessions> tag.

The portion of the RDF in Figure 5 captures the third and fourth key elements of information, the traversal map and the article delimiters.

The traversal map represents pages on the site that are of interest. For example, in the case of VIPAR, current news articles of interest are represented in the site map, while classified ads are explicitly blocked. The map is represented by a series of regular expressions that are used to classify the links found on the site into one of three categories. In the first category, a link is to a page that contains links of interest. Such a page may be thought of as a table of contents page. In the second category, a link is to an article of interest, while in third category, a link is to a page of no interest. The key aspect here is that only the pages of relevance are considered.

Continued from Figure 4

```
       <rdf:Description ID = "tocMetaData">
          <rdf:Bag>
           <ORNL:urlRegEx>
              http://pidp.ewc.hawaii.edu/pireport/graphics.htm
           </ORNL:urlRegEx>
        </rdf:Bag>
     </rdf:Description>

     <rdf:Description ID="articleMetaData">
        <rdf:Bag>
           <rdf:Description ID="article">
              <ORNL:urlRegEx>
                   http://pidp\.ewc\.hawaii\.edu/pireport/[0-9]{4}/
                   (January|February|March|April|May|June|July|August
                   |September|October|November|December)/[0-9]{2}-[0-
                   9]{2}-[0-9]{2}\.htm
              </ORNL:urlRegEx>
              <ORNL:startOfTextStr>
                   <b><font SIZE="4">
              </ORNL:startOfTextStr>
              <ORNL:endOfTextStr>
                   <font face="Times New Roman" size="3">
              </ORNL:endOfTextStr>
           </rdf:Description>
        </rdf:Bag>
     </rdf:Description>
 </rdf:Description>
</rdf:RDF>
```

Figure 5. Second half, RDF for Pacific Islands Report

Continuing in Figure 5, the rdf:Description="tocMetaData"> tag contains one or more table of contents (toc) regular expressions. These are an unordered list, and thus wrapped in the <rdf:Bag> container tags. The <ORNL:urlRegEx> tag contains a regular expression to categorize the link. Those links that match the regular expression are considered to be table of contents pages, and are recursively scoured for links to pages of interest. For PIR, there was only one type of table of contents to describe, thus there is only one description within the <rdf:Bag> container tags. The <rdf:Description="articleMetaData"> tag contains one or more unordered article descriptions. The <rdf:Description ID="article"> tag contains information for one type of article of interest found at a site; this tag set contains an association of three sub-tags, <ORNL:urlRegEx>, <ORNL:startOfTextStr>, and <ORNL:endOfTextStr>. The <ORNL:urlRegEx> tag contains a regular expression with which the retrieval agent tests links found on the site. Those links that pass this regular expression test are considered to be article pages. In this example, the regular expression:

```
http://pidp\.ewc\.hawaii\.edu/pireport/[0-
9]{4}/(January|February|March|April|May|June
|July|August|September|October|November|Dece
mber)/[0-9]{2}-[0-9]{2}-[0-9]{2}\.htm
```

is used to test the links for articles.

The fourth key element of information, article delimiters, is also contained within the `<rdf:Description ID="article">` tag. Article delimiters are only needed for pages that contain articles. Note, however, that a page may be both an article and a table of contents, that is, the page contains both article text and links of other pages of interest. In such a case, a regular expression for such a page would appear in both the `<rdf:Description ID="article">` tag and in the `<rdf:Description="tocMetaData">` tag.

The `<ORNL:startOfTextStr>` tag contains a character string that delimits the beginning of the article text, and the `<ORNL:endOfTextStr>` tag contains a character string that delimits the end of the article text. The goal is to be able to find a consistent combination of characters that delimit the article text for all articles matching the regular expression contained in the associated `<ORNL:urlRegEx>` tag. Note that these delimiting character strings must match the HTML found at the newspaper's web site, whether or not the HTML is well-formed. So far, we have not found a site where this cannot be done. Note that in this PIR example, these characters are HTML tags, but that is not the case with all sites.

The fifth key element of information, article structuring rules, have been added to the text processing software of the VIPAR system, and works very well for converting the raw article text to XML. As of the writing of this paper, we have yet to fully implemented this element in our RDF ontology. The implementation would be very similar to the article delimiters, where the consistent structure of an article would be identified throughout the pages of a site.

To reiterate, the key point is that an XML document has been automatically generated from an unstructured HTML document using an RDF ontology. This conversion process now allows the full power of XML to be exploited on the converted pages.

5. Discussion

This approach to converting HTML documents to XML is automatic once an RDF ontology has been defined. This allows for software agents to be able to understand the ontology and automatically gather information from an Internet site. The approach works quite well, and has been tested on more than 17 newspaper sites. We were very pleased with the overall performance of this ontological approach. Our retrieval agents were able to process the ontologies with ease, and the simplicity of the ontologies provided great flexibility during the project. Additionally, several RDF ontologies have been created by non-programmers in a matter of hours.

This approach holds enormous promise for the conversion of existing HTML pages to XML. In a matter of minutes, this system can convert the key components of an HTML site into XML. Furthermore, the ontology that drives this process can be quickly developed by virtually anyone.

One aspect of XML that we were not able to exploit, but we believe is very powerful, is the ability to search articles over time. Since the articles are encoded in XML, along with considerable metadata, it would be fairly simple to use the XML tags to perform complex queries over the set of articles. For example, to determine what author wrote the most articles on Osama Bin Laden in the first quarter of the year.

Another aspect that we have not addressed would be the use of style sheets to tailor the way that an article is presented to the analysts. This again would be a trivial enhancement, but would allow the analysts to customize the way they see information. Some would like to see additional information about an article, such as the copyright holder for the article. Style sheets would also allow a consistent formatting of articles across the varied newspaper sites.

Several issues remain to be solved, first of which is the automatic creation of an RDF file. There are thousands of Internet newspapers and manually creating an RDF file for each paper is a relatively time intensive process. Although much of the information required for an RDF ontology can be automatically created, there are a few pieces that cannot be currently automated. For example, determining the start and end of the article text is quite challenging, particularly for a complex paper like the BBC, where sidebars, picture captions, and headline links can easily be mistaken for the start of an article.

Another issue is how to adapt to changes in a newspaper site. During this project, a handful of RDF ontologies had to be modified due to changes in an Internet site. These changes occur unpredictably. Automatically detecting changes at a site is quite a challenge but is related to the automatic creation of RDFs. Progress in one of these areas may leverage progress in the other.

6. Conclusion

The VIPAR project uses an RDF ontology to convert Internet newspapers from an HTML format to an XML format. We believe that this is the first such application of RDF to implement an HTML to XML conversion. This system is currently operational, and is being used by the US Pacific Command to convert Internet newspaper articles to XML on a daily basis. This conversion to XML provides the ability to use the full value of XML, i.e., to tailor of the view of an article, or to query a collection of articles based on XML tags.

The RDF ontology that we have defined is based on five key elements of an Internet newspaper site. These elements are: 1) article metadata, 2) traversal directives, 3) traversal maps, 4) article delimiters, and 5) article structuring rules. RDF provides a rich and compact means of representing this information. Software retrieval agents are able to unambiguously parse the RDF files enabling them to autonomously traverse a site, retrieve articles of interest, and convert the articles from HTML to XML.

We have applied this conversion process to 17 Internet newspapers provided to us by the US Pacific Command, ranging from very simple regional sites, to very complex international sites. Our software retrieval agents have used the site-specific RDF ontologies to successfully convert articles from HTML to XML in all cases. This is a very significant result, considering the vast diversity of the Internet sites.

This automatic conversion process holds great promise for performing larger scale conversion of HTML documents to XML. There are still several hurdles to clear however, such as automatically generating RDF ontologies, or automatically adapting ontologies to site changes. We believe that this approach brings us a step closer to transforming the HTML-based Web of today into the Semantic Web concept of tomorrow.

7. References

[1] T. E. Potok, M. T. Elmore, and N. Ivesic, “Collaborative Management Environment” *Proceedings of the InForum’99 Conference*, http://www.osti.gov/inforum99/papers/collmgmt.html, 1999.

[2]T. E. Potok, N. Ivezic, and B. A. Singletary, “XML For Web Based Collaboration.” *Proceedings of XML98 Conference*, Chicago, IL, 1998.

[3] N. Ashish and C. Knoblock, "Wrapper Generation for Semi-structured Internet Sources", *Workshop on Management of Semistructured Data*, SIGMOD ’97, 1997.

[4] N. Kushmerick, D. Weld and R. Doorenbos, "*Wrapper induction for information extraction.",* IJCAI-97, 1997.

[5] L. Liu, C. Pu, and W. Han, "XWRAP: An XML-enabled wrapper construction system for web information sources", *International Conference on Data Engineering (ICDE)*, 2000, pp.611--621.

[6] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web", *Scientific American*, http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html, May 2001.

[7] K. Barber, "About the DAML Language" http://www.daml.org/about.html.

[8] E. Miller, R. Swik, D. Brickley, and B. McBride, "Semantic Web Activity" http://www.w3.org/2001/sw/, 2001.

[9] S. Decker, S. Melnik, F. Van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks, “The Semantic Web: The Roles of XML and RDF”, *IEEE Expert*, 15(3), October 2000.

[10] O. Lassila, and R. Swick, "Resource Description Framework (RDF) Model and Syntax Specification", W3C Recommendation, http://www.w3.org/TR/REC-rdf-syntax/, February 1999.

[11] “The DARPA Agent Modeling Language” http://www.daml.org.

[12] M. Dean, "Language Feature Comparison", http://www.daml.org/language/features.html.

[13] T. Gruber, “A translation Approach to portable ontology specifications”, *Knowledge Acquisition,* 5, 1993, pp199-220.