

# Report on the paper

## „Data Extraction and Annotation for Dynamic Web Pages“

by Edvin Seferovic  
e0325189@student.tuwien.ac.at  
TU Vienna

April 2005

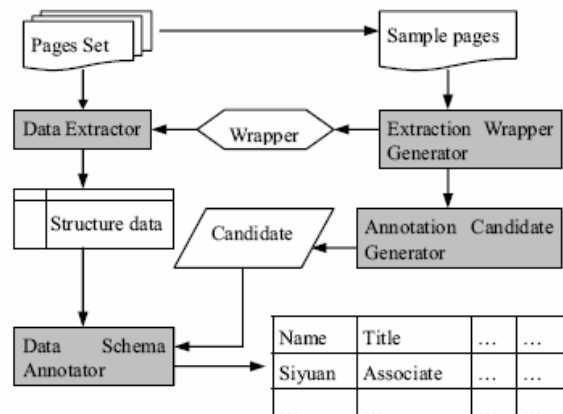
## Introduction

My paper was presented on the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Services that took place in Taipei, Taiwan. As we all know most of the web site ( commercial as well as non-commercial ) have a database as a backend for storing data. Those data are being presented to us on a websites with complex structures and with lot of another ( irrelevant ) elements. In the paper "Data Extraction and Annotation for Dynamic Web Pages" Hui Song presented a rather simple non-supervisor method to extract and annotate data from web pages.

## System architecture

Because of his architecture and main components, the ADeaD system is able to recognize data structure of a web page and extract the data-rich portions. Here are the main elements of the system :

The main elements are :

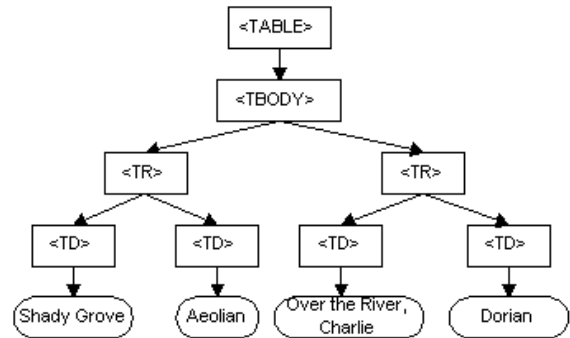


- **Wrapper Generator** – generates a wrapper based on tag tree comparison. This element obtains only the data structure.
- **Data Extractor** – which extracts data with extraction wrapper
- **Annotation Candidate Generator** – generates candidates for annotation of extracted data from the web page
- **Data Annotator** – associates the data elements with annotation candidates

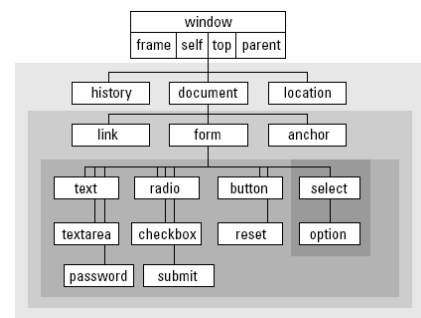
On the following pages I will explain the three steps of IE found in ADeaD system.

## Wrapper generation

In the process of wrapper generation, ADeaD system divides web pages into so called multi template units which are being used to deduce the actual data schema and page template. The generation of the minimum extract tree or the data-rich section of the web page is based on DOM tree. The Document Object Model, or DOM, is an interface that allows scripts or programs to access and manipulate the contents of a web page (or document). It provides a structured, object-oriented representation of the individual elements and content in a page.



The DOM specification "defines the Document Object Model, a platform- and language-neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of documents. The Document Object Model provides a standard set of objects for representing HTML and XML documents, a standard model of how these objects can be combined, and a standard interface for accessing and manipulating them. Vendors can support the DOM as an interface to their proprietary data structures and APIs, and content authors can write to the standard DOM interfaces rather than product-specific APIs, thus increasing interoperability on the Web." [from WD-DOM-19980318]



By comparing two or more web pages the system is able to find the minimum extract tree and discard all other elements that does not contain valueable data. After generation of minimum extract tree, this tree is divided into subtrees to identify template units. The system traverses two trees in depth-first order and compares tag pairs. The comparison is done node-by-node. When two nodes match, it means that the two pages have the same content - template. This leads to a conclusion that repeated elements are not recognized. All other results like tag mismatch or partial matches are signes that those subtrees contain data. This process of comparison starts at lowest level od template unit set and moves to the upper level. In my opinion – this procedure is very useful when it comes to data presented in tables ( see picture ). Aparently other sets without table tags are processed with different method that is not explained nor mentioned in this paper. When the process of the wrapper generation and data schema generation is over, a procedure similar to wrapper generation is used for data extraction.

## Data extraction

Data extraction is the process of getting the valuable data which is defined by data schema from a given pages set and entering this data into a database or XML-based database. As already mentioned – this procedure is similar to wrapper generation. A DOM parser presents each web page in the same tree structure. After the removal of irrelevant tree parts, minimum extract tree is located. By comparing the page template and data schema which is to be extracted, text in table data slots are extracted. Because of the complex structure, XML is used to record the extracted data. To present the collected data, the extracted data have to be annotated to some page elements that present the semantic meaning of the data values.

## Data annotation

**Annotation** is extra information associated with a particular point in a document. In the case of a web page there are data tags that represent extra information ( usually a semantic meaning ) of single/multiple data values. In the ADeaD system the fact that minimum extract tree covers all data items ( annotation strings as well as pure data ) is being used to find those annotation information. Another fact is that the annotation text is visually close to the data element. The authors define two kinds of ranking patterns of annotation text with data elements:

- annotation text is adjacent with the data value in the depth-first traverse order
- annotation text followed by multi data value ( table )

Under those circumstances several heuristics were developed to find annotation text and to compute association possibility of an annotation candidate :

- annotation candidate is found above or below the data element
- the text found to the left or above the data element has a high priority to become annotation candidate

Annotation of data found on websites is used so that machines can also understand and associate data with a semantic meaning. Semantic annotation is used today for semantic web. One of good working systems for IE and semantic annotation is a system called AMILCARE made by Fabio Ciravegna. This system uses a supervised (LP)<sup>2</sup> algorithm for NLP and it has to be trained so it can generate rules and generalize them. Other example for annotation for semantic web would be RDF, a resource description framework that is used for representation of metadata. A good example for this framework is IM application called Trillian that implements RDF metadatabase of Wikipedia.