

# Data Extraction and Annotation for Dynamic Web Pages

Hui Song / Suraj Giri / Fanyuan Ma

PS WIE – Paper #20

by Seferovic Edvin ( e0325189 )



# Data Extraction and Annotation for Dynamic Web Pages

---

- Paper presented on 2004 IEEE International Conference on e-Technology, e-Commerce and e-Services ( Tapei, Taiwan )
- 100 dynamic web pages : 1 static web page
- Lot of irrelevant data ( ads, images, etc. )
- Need for data extraction and annotation



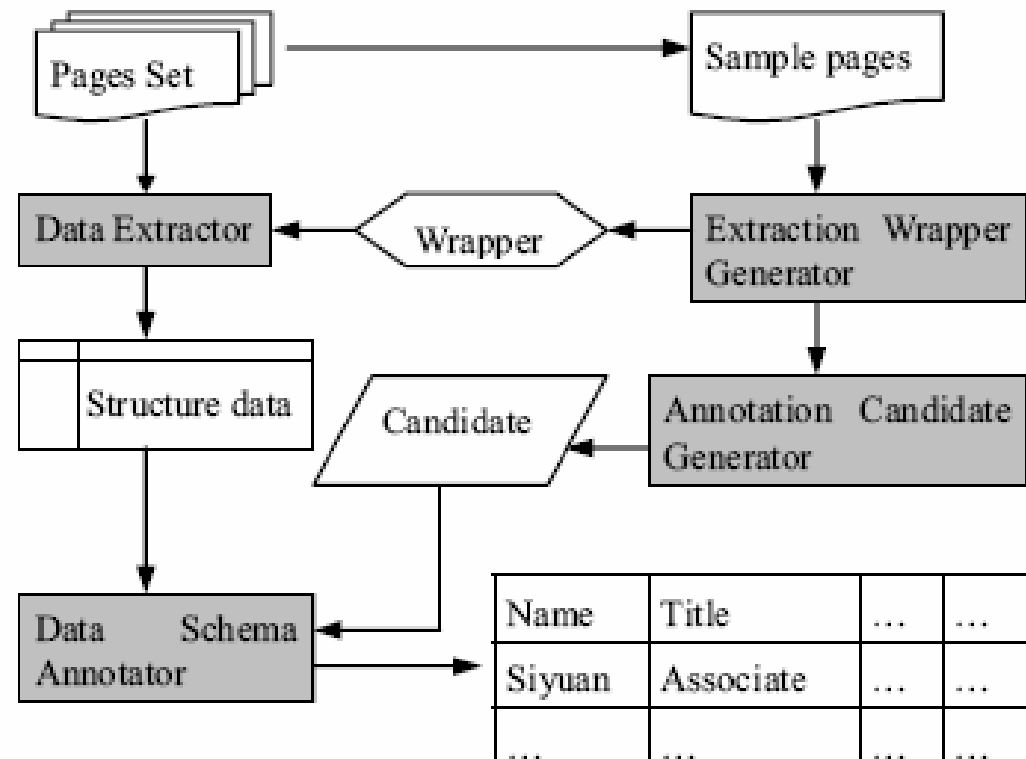


# ADeaD – an Introduction

---

- Non supervised method of data extraction from a page set
  - by addressing an algorithm to automatically generate a wrapper
  - extracted data without a semantic meaning
- Data annotation of extracted data
  - heuristics to associate data items and description items on a web page

# ADeaD – system overview





# ADeaD – main system parts

---

- Wrapper generation
- Data extractor
- Annotation candidate generator
- Data annotator



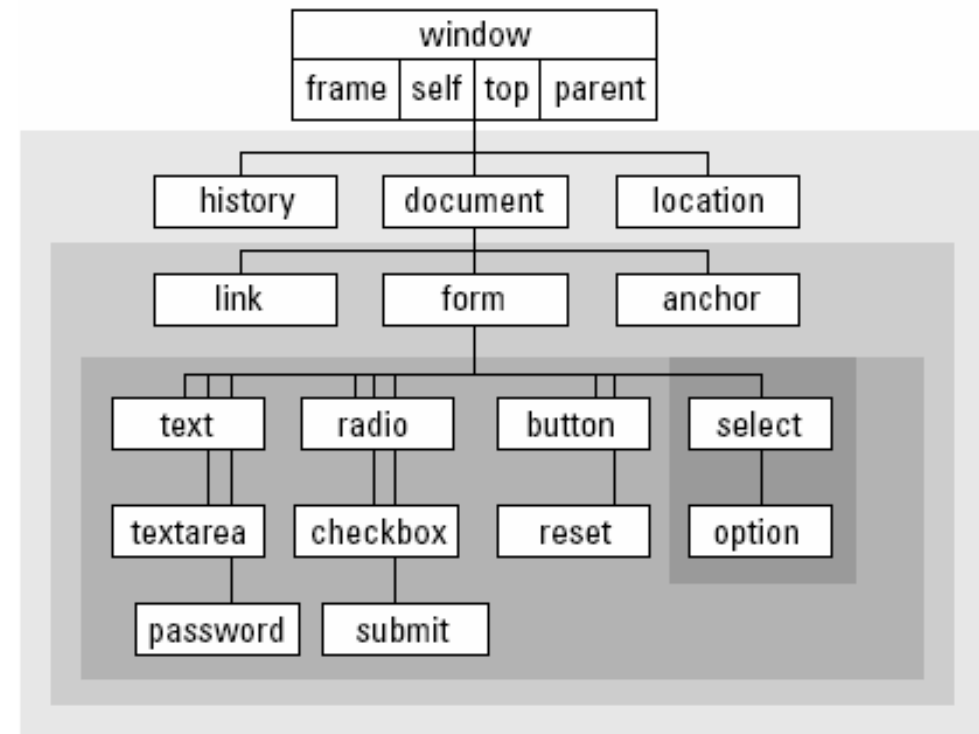
# Wrapper generation

---

- Multi template units
  - Data schema & page template
- DOM tree
  - Minimum extract tree
- Identification of template units

# Document Object Model Tree

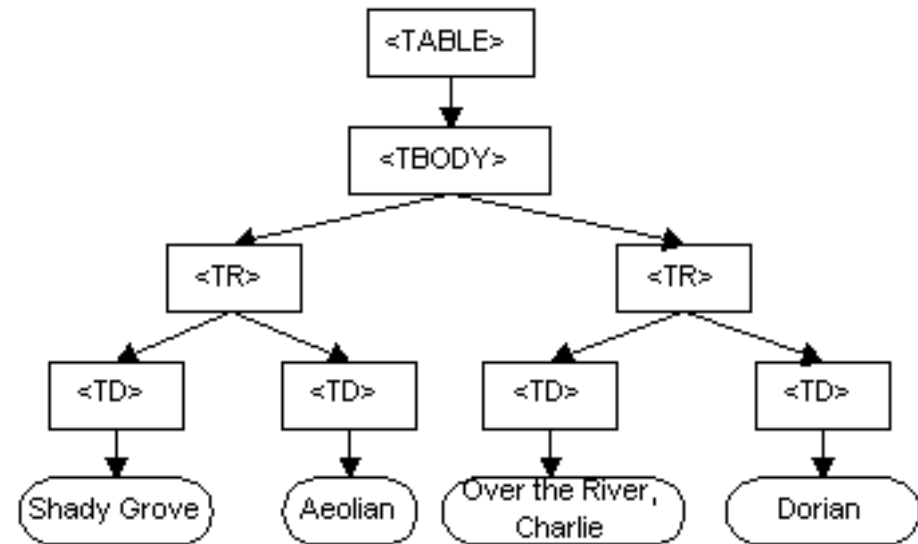
- ...a platform- and language-neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of documents !



# Identification of template units

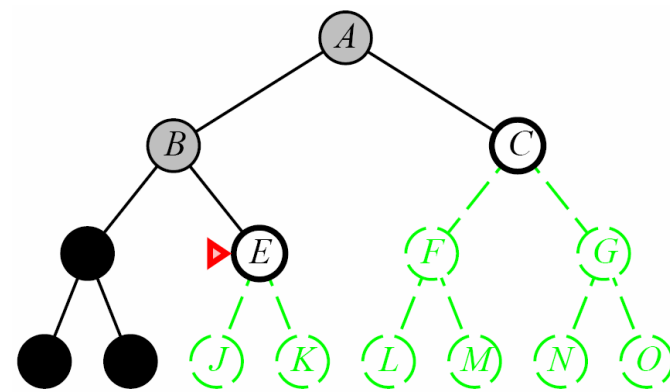
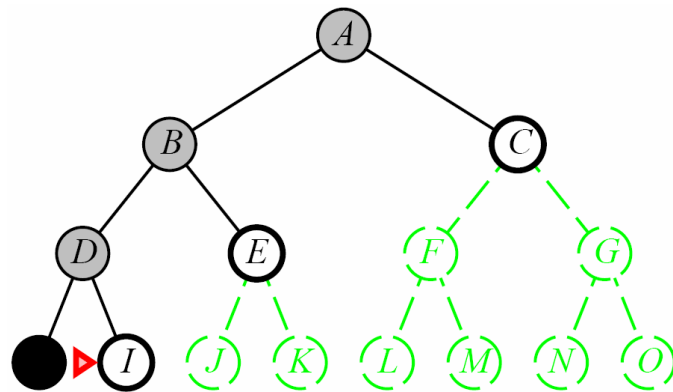
---

- Location of nodes containing textual data
- Depth-first traversing & comparison of tag pairs



# Depth-first traversing of DOM tree

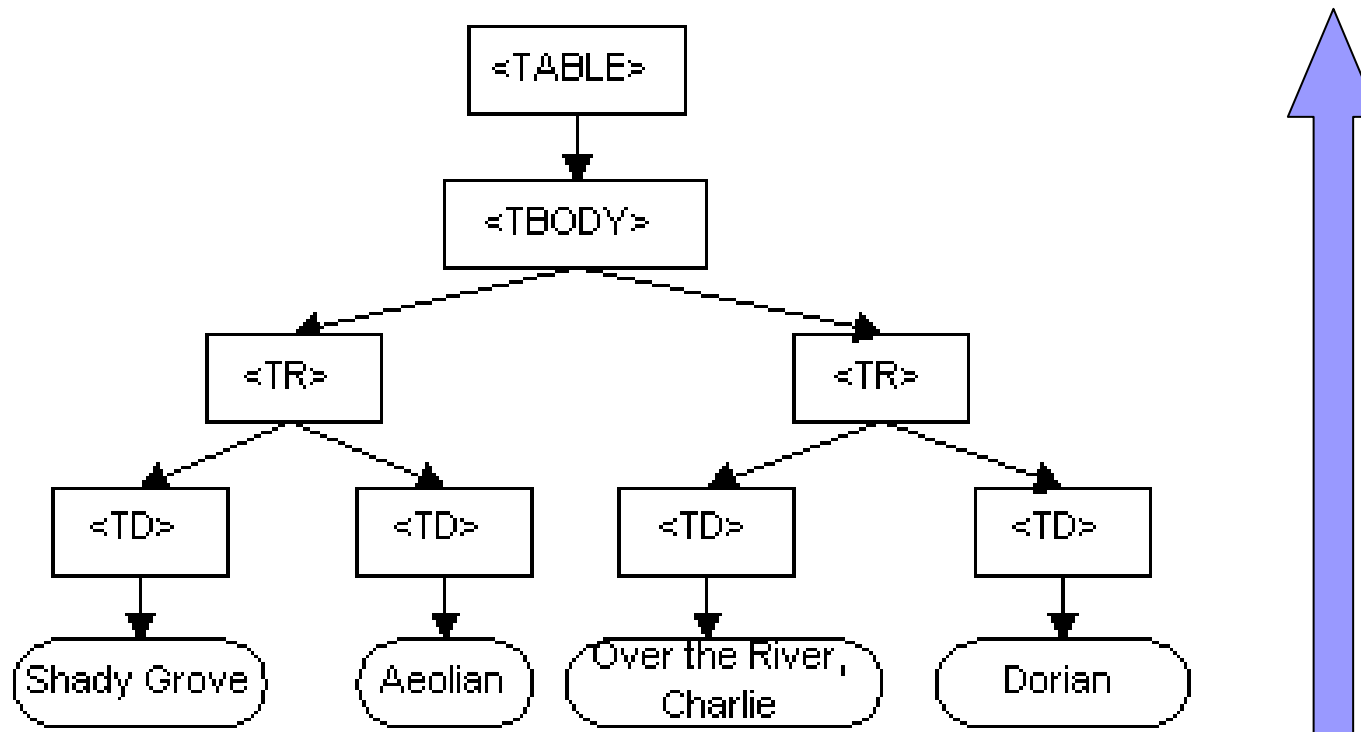
---



- Expand deepest unexpanded node

# Depth-first traversing and tag comparison

---



- Bottom – up search for template units



# Results of tag pair comparison

---

- Nodes match
  - Tag nodes **mismatch**
  - Text nodes **mismatch**
  - Text nodes **partial mismatch**
- 
- DATA SCHEMA

**!!! ONLY FOR TABLE DATA !!!**



# ADeaD – Data Extraction

---

- Similar to wrapper generation
- DOM tree parsing
  - Location of minimum extract tree
- Extraction of text in table data slots
- Data schema -> XML



# ADeaD – Data Annotation

---

- What is actually ment by „annotation“?
- Why the annotation?
- ..and the Oscar goes to... the web designer
  - annotation text is visually close to the data



# ADeaD – Data Annotation 2

---

- Ranking of data annotation candidates
- Heuristics for data annotation



# Ranking of data annotation

---

- Annotation text is adjacent with the data value found by depth-first algorithm
- Annotation text is followed by multi data value ( i.e. table )



# Heuristics for data annotation

---

- Visual placement of annotation text
  - Candidate found above or below
  - Text found on the left or above has high priority to represent annotation candidate
- Text nodes partial **mismatch**

# Semantic Web

---

- What ???
- A web of meaningful data





# Semantic Web by Tim Berners-Lee

---

- "The Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications. It promises to radically improve our ability to find, sort, and classify information, tasks that consume a majority of the time spent on and off-line"

# Semantic Web

---

- What ???
- A web of meaningful data
- Exchange languages
  - XML
  - RDF ( Resource Description Language )
  - DAML ( DARPA Agent Markup Language





# Semantic Web 2

---

- Semantic annotation
  - **AMILCARE** by Fabio Ciravegna
  - Uses supervised (LP)<sup>2</sup> algorithm
  - Mostly NLP algorithms used
  - Rule generation and generalisation
  - **Annotea** by W3C



# ADeaD – the conclusion

---

- No human involvement needed
- Use of minimum extract tree and template unit to increase processing time
- Cannot handle web sites that use CSS
- No explanation given for processing elements not included in the table structure
- Correct extraction & annotation of ~90% which is hard to believe