

Hypermedia and the Web – Lecture Notes

SAX Parsing

SAX is an event-based XML parser.

How to use it:

- Create an instance of XMLReader
 - Do this by using the XMLReaderFactory class in org.xml.sax.helpers package
 - createXMLReader() method
- Create an implementation of the ContentHandler interface
 - ContentHandler has methods that are called as an XML document is parsed
 - At the start & end the document
 - At the start & end of an element
 - You write your code to implement these methods, and perform your processing on the document
 - Example: Arriving at the start/end of an element constitutes an “event”
 - Code typically takes the form of state machines, where some events cause progression through the state machine

ContentHandler interface:

```
package org.xml.sax;

public interface ContentHandler {
    public void setDocumentLocator(Locator locator);
    public void startDocument( ) throws SAXException;
    public void endDocument( ) throws SAXException;
    public void startPrefixMapping(String prefix, String uri)
        throws SAXException;
    public void endPrefixMapping(String prefix) throws SAXException;
    public void startElement(String namespaceURI, String localName,
        String qualifiedName, Attributes atts) throws SAXException;
    public void endElement(String namespaceURI, String localName,
        String qualifiedName) throws SAXException;
    public void characters(char[] text, int start, int length)
        throws SAXException;
    public void ignorableWhitespace(char[] text, int start, int length)
        throws SAXException;
    public void processingInstruction(String target, String data)
        throws SAXException;
    public void skippedEntity(String name) throws SAXException;
```

Most important/useful of these are startDocument/endDocument, startElement/endElement, and characters.

With startDocument, you typically initialize state machines used in processing the XML document. endDocument prints or writes parsing results.

startElement/endElement are used to look for specific elements, and advance state machines based on seeing those elements. startElement also gives you access to a list of attributes defined on the element.

characters gives you the contents of an element.

Study example in XML in a Nutshell, p. 319.