# Przetwarzanie dokumentów XML w oparciu o zdarzenia

Bartłomiej Świercz

Katedra Mikroelektroniki i Technik Informatycznych

Łódź, 21 października 2005 roku

SAX – Simple API for XML

SAX jest szeregowym interfejsem opartym o zdarzenia. Dostarcza mechanizm odczytywania danych z dokumentów XML. SAX został zaprojektowany przez członków listy pocztowej xml-dev bez formalnego wsparcia instytucji takich jak W3C. Początkowym liderem projektu był David Megginson, który pracował nad implementacją SAX dla języka Java. Obecnie stroną domową projektu SAX jest:

`http://www.saxproject.org`

# SAX - odczytywanie dokumentu XML

SAX traktuje dokument XML jako strumień danych, który jest odczytywany sukcesywnie w czasie parsowania. Oznacza to, że nie da się przetworzyć ponownie wcześniej doczytanego elementu, bez wznowienia odczytywania całego dokumentu. API SAX oparte jest o mechanizm zdarzeń, które są generowane podczas parsowania elementu. Zdarzenie wywołuje zarejestrowaną funkcję zwrotną (ang. callback).

# Funkcje zwrotne - przypomnienie

## Język C

```
int add (int x, int y)
{
        return x + y;
}


void sum (int a, int b, int (*my_add) (int, int))
{
        . . . .
        my_add (a, b);
        . . . .
}
```

# Funkcje zwrotne - przypomnienie

## Język Python

```
def add (x, y):
        return x + y


def sum (a, b, fun):
        . . . .
        fun (a, b)
        . . . .
```

# SAX kontra DOM

API oparte o drzewa (DOM) Algorytm mapowania dokumentu XML na drzewo elementów przechowywane w całości w pamięci.

API oparte o zdarzenia (SAX) Algorytm parsowania dokumentów XML generujący zdarzenia po napotkaniu elementów dokumentu. Aplikacja musi dostarczyć funkcje obsługi zdarzeń podobnie jak aplikacja GUI dostarcza funkcje obsługi zdarzeń elementów graficznego interfejsu użytkownika.

Wady i zalety obydwu rozwiązań ...

# Przetwarzanie dokumentów oparte o zdarzenia

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Rozpoczynamy przetwarzanie ...

# Przetwarzanie dokumentów oparte o zdarzenia

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 1

Nazwa zdarzenia: Początek dokumentu

# Przetwarzanie dokumentów oparte o zdarzenia

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 2

Nazwa zdarzenia: Początek elementu: “dvd”

# Przetwarzanie dokumentów oparte o zdarzenia

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 3

Nazwa zdarzenia: Początek elementu: “tytul”

# Przetwarzanie dokumentów oparte o zdarzenia

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 4

Nazwa zdarzenia: Tekst: “Rambo I”

# Przetwarzanie dokumentów oparte o zdarzenia

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 5

Nazwa zdarzenia: Koniec elementu: “tytul”

# Przetwarzanie dokumentów oparte o zdarzenia

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 6

Nazwa zdarzenia: Początek elementu: “aktor”

# Przetwarzanie dokumentów oparte o zdarzenia

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 7

Nazwa zdarzenia: Tekst: “Sylvester Stallone”

# Przetwarzanie dokumentów oparte o zdarzenia

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 8

Nazwa zdarzenia: Koniec elementu: “aktor”

# Przetwarzanie dokumentów oparte o zdarzenia

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 9

Nazwa zdarzenia: Koniec elementu: “dvd”

# Przetwarzanie dokumentów oparte o zdarzenia

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 10

Nazwa zdarzenia: Koniec dokumentu

# Interfejs SAX

Interfejs SAX składa się z czterech głównych obiektów:

**ContentHandler** Metody tego obiektu wywoływane są podczas generowania zdarzeń. Jest to główny obiekt (interfejs) modułu SAX.

**DTDHandler** Wywoływany jest do przetwarzania zdarzeń związanych z obsługą DTD.

**EntityResolver** Obiekt odpowiada za obsługę zewnętrznych jednostek dokumentu.

**ErrorHandler** Obiekt służy do generowania informacji o błędach w parsowaniu dokumentu XML.

# Obiekt ContentHandler

## Klasa xml.sax.handler.ContentHandler:

```
def ContentHandler:
        def startDocument(self):
        ...
        def endDocument(self):
        ...
        def startElement(self, name, attrs):
        ...
        def endElement(self, name):
        ...
        def characters(self, content):
        ...
```

# Obiekt ContentHandler

## Dziedziczenie po klasie handler.ContentHandler:

```
from xml.sax import handler

def myHandler (handler.ContentHandler):
        def startDocument(self):
                print ‘‘Poczatek dokumentu’’
        ...
        def endDocument(self):
                print ‘‘Koniec dokumentu’’
        ...
```

# Zdarzenia obiektu handler.ContentHandler

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Rozpoczynamy przetwarzanie ...

# Zdarzenia obiektu handler.ContentHandler

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 1

Nazwa zdarzenia: startDocument

# Zdarzenia obiektu handler.ContentHandler

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 2

Nazwa zdarzenia: startElement

# Zdarzenia obiektu handler.ContentHandler

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 3

Nazwa zdarzenia: startElement

# Zdarzenia obiektu handler.ContentHandler

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 4

Nazwa zdarzenia: characters

# Zdarzenia obiektu handler.ContentHandler

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 5

Nazwa zdarzenia: endElement

# Zdarzenia obiektu handler.ContentHandler

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 6

Nazwa zdarzenia: startElement

# Zdarzenia obiektu handler.ContentHandler

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 7

Nazwa zdarzenia: characters

# Zdarzenia obiektu handler.ContentHandler

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 8

Nazwa zdarzenia: endElement

# Zdarzenia obiektu handler.ContentHandler

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 9

Nazwa zdarzenia: endElement

# Zdarzenia obiektu handler.ContentHandler

## Plik XML:

```
<?xml version=‘‘1.0’’?>
<dvd>
        <tytul>Rambo I</tytul>
        <aktor>Sylvester Stallone</aktor>
</dvd>
```

## Zdarzenie generowane podczas przetwarzania:

Zdarzenie numer: 10

Nazwa zdarzenia: endDocument