



OXPath

(Oxford XPath)

Andrew Sellers

Georg Gottlob

Giovanni Grasso

Tim Furche

Christian Schallhart

Outline

- Introduction
- Actions
- Additional Axes
- Web Form Field Association
- Result Extraction
- Architecture

Goals and Motivations

- Provide a declarative formalism that extends XPath
 - Simulate user actions in order to query and manipulate a Dynamic DOM
 - Provide a means for automated systems to specify and maintain Deep Web queries
- Develop a lightweight, scalable tool for building domain-specific knowledge bases from existing, disparate web data

Goals and Motivations

- DIADEM
 - will generate (by learning) suitable “access paths” for retrieving and aggregating data from websites
 - On the fly evaluation on thousand of websites and interlinked pages containing forms and menus
- Main issues
 - Web forms population
 - Result pages navigation and data extraction
 - Scalability

XPath

- Extension of XPath
- Facilitates querying web form and retrieving returned data
- Simulates user actions (mouse events/keyboard strokes) for filling out web forms and collecting data from multiple pages
- Highly parallelizable
- Navigation across multiple pages

Actions

- Standard XPath expressions are used for node identification and navigation
- Actions to be taken on the context node are expressed by {..} brackets
- Three type of “actions”
 - Explicit input reference
 - e.g., entering “London” in a text field
 - entering values read from external sources (DB, XML, separate XPath expression, file)
 - Events
 - submit, click, unclick, mouse events
 - Options selection by index
 - e.g., suggested options in a drop-down list, multiselection

User actions simulation



```
www.google.com//input[@name='q']/{“Oxford”}/  
following::input[@name='btnG']/{click}
```

A single XPath expression can also represent several queries
(non-ground expression)

```
www.google.com//input[@name='q']/{“Oxford”, “Cambridge”,  
“London”}/ following::input[@name='btnG']/{click}
```

Its instantiation produces three different ground expressions,
asking “Oxford”, “Cambridge”, and “London”, respectively.

In case of multiple selection allowed, it is possible to specify values using inner { } brackets

By index { {1,4,6,7} }

By values { {"val1",..., "valn"} }

All values { {*} } and { {regex} }



Input can be also provided

from external sources by special predicates such as

{ *fromDB(db url, sql query)* }

{ *fromFile(url)* }

{ *fromRDF(url, sparql query)* }

{ *fromXML(document url, XPath query)* }

Additional Axes for XPath

Facilitate the navigation between form fields

- Next-Field
 - selects the next field, in document order
- Previous-Field (as inverse of Next-Field)
 - selects the previous field, in document order
- Following-Field
 - returns all the following fields, in document order.
- Preceding-Field (as inverse of Following-Field)
 - returns all the preceding fields, in document order

Additional Axes

- Support conditional predicates and node tests
 - /next-field::input[@type=radio]
 - finds the first radio button in document order
- Expressions become far shorter and more intuitively

www.google.com//input[@name='q']/{“Oxford”}/
following::input[@name='btnG']/{click}



www.google.com//next-field::*/{“Oxford”}/next-field::*/{click}

Form Field Associations

Rent from: to £'s per month

...//next-field::<*/{ *fromFile(myFile)* }/

next-field::<*/ { *fromFile(myFile)* }/



...//next-field::<*/{ *X = fromFile(myFile)* }/

next-field::<*/ { *fromFile(myFile) > X* }/

Allows for much more efficient evaluation than Cartesian products of field values

Result Extraction

```
..../next-field::*/{“Renting”}/.../{...}/.../{“Submit”} /foreach(...)/...  
/{scrape(loc)}/...
```

UK International

I'm interested in Buying
 Renting
 New developments


Maximum price

Minimum bedrooms

Area
e.g Surrey, SW4, Chelsea

GO

George Street, OX1
Monthly Rental of GBP 1,200




A top floor apartment offering spacious accommodation. Situated in the heart of the City, with convenient access to amenities including the train station. Ideal for professional sharers or a couple.

Flat | 3 Bedrooms | 1 Reception | 1 Bathroom
Furnished | Long let - more than six months

Oxford Lettings
+44 (0)1865 723557

[View Details](#)
[Select Property](#)

Rewley Road, OX1
Monthly Rental of GBP 950



A spacious, well presented modern two bedroom apartment situated on the second floor of this purpose built building. Within close proximity to Oxford City Centre and station.

Flat | 2 Bedrooms | 1 Reception | 2 Bathrooms
Own Parking | Furnished | Long let - more than six months

Oxford Lettings
+44 (0)1865 723557

[View Details](#)
[Select Property](#)

Atomic results
regardless of
presentation (list, table,
etc.)

Thank you