OXPath

Scalable, Memory-efficient Data Extraction from Web Applications

Tim Furche

July 3rd, 2012 @ Department of Computer Science, Oxford University

joint work with Georg Gottlob, Giovanni Grasso, Christian Schallhart, Andrew Sellers
Find Homes

Address or Neighborhood or City or ZIP

For Sale, Zestimate® Home Values, Rentals

Oxford Homes For Sale

4 Peabody Dr
3 beds, 2,287 sqft, 2.5 baths
For Sale: $198,000

3606 Oxford Millville Rd
2 beds, 1,672 sqft, 3.0 baths
For Sale: $124,900

4747 Booth Rd
2 beds, 1,008 sqft, 1.0 baths
For Sale: $49,900

Find Mortgage Rates

Ohio

Low Rates, Trusted Lenders, Anonymous

Zillow Mortgage Marketplace

Special Offer

For a limited time
GET UP TO $1,000
in Lowe’s gift card credit

Get Started

On top of already low rates and fees
Seattle

4829 results. 23 unmapped.

- For Sale: 4837
- For Rent: 935
- Make Me Move: 1727
- Recently Sold: 7156

Price
Min - Max

Beds
0+ 0+

Baths
0+ 0+

Sort: Featured

2004 34th Ave S, Seattle, WA (Mt. Baker)
- For Sale: $1,095,000
- Zestimate®: $679,700
- Mortgage: $5,259/mo
- Beds: 3
- Baths: 2.25
- Sqft: 2,380
- Days on Zillow: 73
- Built: 1924
- Single Family
- Price/sqft: $460
<table>
<thead>
<tr>
<th>Address</th>
<th>City, State, Zip</th>
<th>Status</th>
<th>Price</th>
<th>Beds</th>
<th>Baths</th>
<th>Days on Zillow</th>
<th>Built</th>
<th>Type</th>
<th>Price/sqft</th>
</tr>
</thead>
<tbody>
<tr>
<td>4128 SW Austin St, Seattle, WA</td>
<td>Seattle, WA</td>
<td>For Sale</td>
<td>$459,900</td>
<td>4</td>
<td>2.5</td>
<td>104</td>
<td>1920</td>
<td>Single Family</td>
<td>$134</td>
</tr>
<tr>
<td>3632 Courtland Pl S, Seattle, WA</td>
<td>Seattle, WA</td>
<td>Pending</td>
<td>$219,900</td>
<td>3</td>
<td>2.5</td>
<td>113</td>
<td>2003</td>
<td>Miscellaneous</td>
<td>$109</td>
</tr>
<tr>
<td>2716 Elliott Ave APT 1005, Seattle, WA</td>
<td>Seattle, WA</td>
<td>For Sale</td>
<td>$250,000</td>
<td>1</td>
<td>1.0</td>
<td>2</td>
<td>2003</td>
<td>Condo</td>
<td>$393</td>
</tr>
<tr>
<td>1709-A Dexter Ave N, Seattle, WA</td>
<td>Seattle, WA</td>
<td>For Rent</td>
<td>$3,300/mo</td>
<td>3</td>
<td>3.5</td>
<td>2</td>
<td>--</td>
<td>Condo</td>
<td>$411</td>
</tr>
<tr>
<td>831 NW 63rd St APT A, Seattle, WA</td>
<td>Seattle, WA</td>
<td>For Sale</td>
<td>$389,000</td>
<td>3</td>
<td>2.5</td>
<td>2</td>
<td>2001</td>
<td>Single Family</td>
<td>$266</td>
</tr>
<tr>
<td>5627 45th Ave SW, Seattle, WA</td>
<td>Seattle, WA</td>
<td>For Sale</td>
<td>$437,500</td>
<td>4</td>
<td>2.0</td>
<td>2</td>
<td>1911</td>
<td>Single Family</td>
<td>$227</td>
</tr>
<tr>
<td>Address</td>
<td>Neighborhood</td>
<td>Status</td>
<td>Price</td>
<td>Price Cut</td>
<td>Beds</td>
<td>Baths</td>
<td>Sqft</td>
<td>Built</td>
<td>Type</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------------</td>
<td>-----------------</td>
<td>------------------</td>
<td>-------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>----------</td>
</tr>
<tr>
<td>4128 SW Austin St, Seattle, WA</td>
<td>Gatewood</td>
<td>For Sale</td>
<td>$459,900</td>
<td>$15,000</td>
<td>4</td>
<td>2.5</td>
<td>3,420</td>
<td>1920</td>
<td>Single Family</td>
</tr>
<tr>
<td>3632 Courtland Pl S, Seattle, WA</td>
<td>Genesee</td>
<td>Pending</td>
<td>$219,900</td>
<td>$5,000</td>
<td>3</td>
<td>2.5</td>
<td>2,010</td>
<td>2003</td>
<td>Miscellaneous</td>
</tr>
<tr>
<td>2716 Elliott Ave APT 1005, Seattle, WA</td>
<td>Belltown</td>
<td>For Sale</td>
<td>$250,000</td>
<td></td>
<td>1</td>
<td>1.0</td>
<td>636</td>
<td>2003</td>
<td>Condo</td>
</tr>
<tr>
<td>1709-A Dexter Ave N, Seattle, WA</td>
<td>Westlake</td>
<td>For Rent</td>
<td>$3,300/mo</td>
<td></td>
<td>3</td>
<td>3.5</td>
<td>1,875</td>
<td>--</td>
<td>Condo</td>
</tr>
<tr>
<td>831 NW 63rd St APT A, Seattle, WA</td>
<td>East Ballard</td>
<td>For Sale</td>
<td>$389,000</td>
<td></td>
<td>3</td>
<td>2.5</td>
<td>1,460</td>
<td>2001</td>
<td>Single Family</td>
</tr>
<tr>
<td>5627 45th Ave SW, Seattle, WA</td>
<td>Seaview</td>
<td>For Sale</td>
<td>$437,500</td>
<td></td>
<td>4</td>
<td>2.0</td>
<td>1,920</td>
<td>1911</td>
<td>Single Family</td>
</tr>
<tr>
<td>Address</td>
<td>Location</td>
<td>Type</td>
<td>Bedrooms</td>
<td>Bathrooms</td>
<td>Price</td>
<td>Days on Zillow</td>
<td>Built</td>
<td>Price/sqft</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------</td>
<td>------------</td>
<td>----------</td>
<td>-----------</td>
<td>----------------</td>
<td>----------------</td>
<td>-------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>4128 SW Austin St, Seattle, WA</td>
<td>Gatewood</td>
<td>House For Sale</td>
<td>4</td>
<td>2.5</td>
<td>$459,900</td>
<td>104</td>
<td>1920</td>
<td>$134</td>
<td></td>
</tr>
<tr>
<td>3632 Courtland Pl S, Seattle, WA</td>
<td>Genesee</td>
<td>Pending</td>
<td>3</td>
<td>2.5</td>
<td>$219,900</td>
<td>113</td>
<td>2003</td>
<td>$109</td>
<td></td>
</tr>
<tr>
<td>2716 Elliott Ave APT 1005, Seattle, WA</td>
<td>Belltown</td>
<td>Home For Sale</td>
<td>1</td>
<td>1.0</td>
<td>$250,000</td>
<td>2</td>
<td>2003</td>
<td>$393</td>
<td></td>
</tr>
<tr>
<td>1709-A Dexter Ave N, Seattle, WA</td>
<td>Westlake</td>
<td>Home For Rent</td>
<td>3</td>
<td>3.5</td>
<td>$3,300/mo</td>
<td>2</td>
<td>--</td>
<td>$411</td>
<td></td>
</tr>
<tr>
<td>831 NW 63rd St APT A, Seattle, WA</td>
<td>East Ballard</td>
<td>House For Sale</td>
<td>3</td>
<td>2.5</td>
<td>$389,000</td>
<td>2</td>
<td>2001</td>
<td>$266</td>
<td></td>
</tr>
<tr>
<td>5627 45th Ave SW, Seattle, WA</td>
<td>Seaview</td>
<td>House For Sale</td>
<td>4</td>
<td>2.0</td>
<td>$437,500</td>
<td>2</td>
<td>1911</td>
<td>$227</td>
<td></td>
</tr>
</tbody>
</table>
3907 14th Ave S
Seattle, WA 98108

For Sale: $359,950
Zestimate: $376,800
Mortgage payment: $1,996/mo

Bedrooms: 4
Bathrooms: 3.5
Sq ft: 2,340
Lot size: 4,810 sq ft / 0.11 acres
Property type: Single Family
Year built: 1999
Parking type: Garage - Attached
Cooling system: Forced air
Heating system: Yes
Fireplace: Yes
Days on Zillow: 27
Last sold: April 25 2002
MLS number: 257063

Description:
Prepare to TURN that KEY & joyfully MOVE IN! Fabulous, light & bright home in quiet North Beacon Hill location! Spacious living spaces! GLEAMING hardwood floors. Unbeatable circular floor plan! Lovely open Scandinavian kitchen- maple, gas cooktop, stainless! Breakfast bar & wonderful dining room...

Your e-mail
Message (optional)
I am interested in 3907 14th Ave S, Seattle, WA.

Contact Agent
3907 14th Ave S
Seattle, WA 98108

For Sale: $359,950
Zestimate®: $376,800
Mortgage payment: $1,996/mo

Bedrooms: 4
Bathrooms: 3.5
Sqft: 2,340
Lot size: 4,810 sq ft / 0.11 acres
Property type: Single Family
Year built: 1999
Parking type: Garage - Attached
Cooling system: --
Heating system: Forced air
Fireplace: Yes
Days on Zillow: 27
Last sold: April 25 2002
MLS number: 257063
Open house: Sun08/28, 1:00pm - 4:00pm

Description
Prepare to TURN that KEY & joyfully MOVE IN! Fabulous, light & bright home in quiet North Beacon Hill location! Spacious living spaces! GLEAMING hardwood floors. Unbeatable circular floor plan! Lovely open Scandinavian kitchen- maple, gas cooktop, stainless! Breakfast bar & wonderful dining room...

More facts

Contact agent
Janine Duncan
Windermere Real Estate Call: (888) 855-0416
Contributions: 66

Your e-mail

Message (optional)
I am interested in 3907 14th Ave S, Seattle, WA.

Contact Agent

Charts and Data

<table>
<thead>
<tr>
<th>Value</th>
<th>Range</th>
<th>30-day change</th>
<th>$/sqft</th>
<th>Last updated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Combined Facts                     | Public Facts | Visit county website?
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bedrooms: 4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Bathrooms: 3.5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Sqft: 2,340</td>
<td>2,010</td>
<td></td>
</tr>
<tr>
<td>Lot size: 4,810 sq ft / 0.11 acres</td>
<td>4,810 sq ft / 0.11 acres</td>
<td></td>
</tr>
<tr>
<td>Property type: Single Family</td>
<td>Single Family</td>
<td></td>
</tr>
<tr>
<td>Year built: 1999</td>
<td>1999</td>
<td></td>
</tr>
<tr>
<td>Parking type: Garage - Attached</td>
<td>Garage - Attached</td>
<td></td>
</tr>
<tr>
<td>Cooling system:</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Heating system: Forced air</td>
<td>Forced air</td>
<td></td>
</tr>
<tr>
<td>Fireplace: Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Days on Zillow: 27</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Last sold: April 25 2002</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>MLS number: 257063</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Basement: Finished</td>
<td>Finished</td>
<td></td>
</tr>
<tr>
<td>Basement sqft: 450</td>
<td>--</td>
<td></td>
</tr>
</tbody>
</table>
A Call for **Action** in Web Extraction!

**Past:** Form Filling + HTML Patterns

**Now:** Interaction + DOM Patterns

- getting to the data requires interaction not just form filling
- identifying relevant data from rendered DOMs
  - including computed style and geometric information
- access to all CSS properties, but less rich relations than in SXPath
Kleene Star. Finally, we add the Kleene star, as in \[12\]. For example, the following expression queries Google for "Oxford", traverses all accessible result pages and extracts all links.

\[
\text{doc("google.com")/descendant::field()[1]/"Oxford"/following::field()[1]/click/(descendant::a:<Link=(@href)>[.#="Next"])/click\}^*
\]

To limit the range of the Kleene star, one can specify upper and lower bounds on the multiplicity, e.g., \((\ldots)^{3,8}\).

\[
\text{doc("zillow.com")/descendant::field()[1]/"Seattle"/following::*#GOButton/click/descendant::input[@type='checkbox'][2]/uncheck/following::field[]{uncheck}///div[.="Beds"]//select/{"3+"}///a[./text()="More filters"]/click//input[following-sibling[contains(.,"Multi Family")]]/uncheck/(//span.arrowNext/a/click)^*//ul#search-results/li:<property>
\]

\section*{References}

\begin{itemize}
\end{itemize}
The nesting in the result mirrors the structure of the OXPath expression: extraction markers in a predicate (title, source) represent attributes to the last marker outside the predicate (story).

Kleene Star. Finally, we add the Kleene star, as in [12]. For example, the following expression queries Google for "Oxford", traverses all accessible result pages and extracts all links.

```xml
doc("google.com")/descendant::field()[1]/"Oxford"/following::field()[1]/{"click /}
```

To limit the range of the Kleene star, one can specify upper and lower bounds on the multiplicity, e.g., (...)*{3,8}.

```xml
doc("zillow.com")/descendant::field()[1]/"Seattle"/following::*#GOButton/{click/}
```

2. REFERENCES


<table>
<thead>
<tr>
<th>Address</th>
<th>Type</th>
<th>Beds</th>
<th>Baths</th>
<th>Sqft</th>
<th>Days on Zillow</th>
<th>Built</th>
<th>Price/sqft</th>
</tr>
</thead>
<tbody>
<tr>
<td>4128 SW Austin St, Seattle, WA</td>
<td>House For Sale</td>
<td>4</td>
<td>2.5</td>
<td>3,420</td>
<td>104</td>
<td>1920</td>
<td>$134</td>
</tr>
<tr>
<td>3632 Courtland Pl S, Seattle, WA</td>
<td>Pending</td>
<td>3</td>
<td>2.5</td>
<td>2,010</td>
<td>113</td>
<td>2003</td>
<td>$109</td>
</tr>
<tr>
<td>2716 Elliott Ave APT 1005, Seattle, WA</td>
<td>Home For Sale</td>
<td>1</td>
<td>1.0</td>
<td>636</td>
<td>2</td>
<td>2003</td>
<td>$393</td>
</tr>
<tr>
<td>1709-A Dexter Ave N, Seattle, WA</td>
<td>Home For Rent</td>
<td>3</td>
<td>3.5</td>
<td>1,875</td>
<td>2</td>
<td>--</td>
<td>$411</td>
</tr>
<tr>
<td>831 NW 63rd St APT A, Seattle, WA</td>
<td>House For Sale</td>
<td>3</td>
<td>2.5</td>
<td>1,460</td>
<td>2</td>
<td>2001</td>
<td>$266</td>
</tr>
<tr>
<td>5627 45th Ave SW, Seattle, WA</td>
<td>House For Sale</td>
<td>4</td>
<td>2.0</td>
<td>1,920</td>
<td>2</td>
<td>1911</td>
<td>$266</td>
</tr>
</tbody>
</table>
Kleene Star. Finally, we add the Kleene star, as in [12]. For example, the following expression queries Google for "Oxford", traverses all accessible result pages and extracts all links.

doc("google.com")/descendant::field()[1]/"{click /}(*

doc("zillow.com")/descendant::field()[1]/"{click /}*/( //a:"{click /}*/)*/

2. REFERENCES
The nesting in the result mirrors the structure of the OXPath expression: extraction markers in a predicate (title, source) represent attributes to the last marker outside the predicate (story).

Kleene Star. Finally, we add the Kleene star, as in [12]. For example, the following expression queries Google for "Oxford", traverses all accessible result pages and extracts all links.

doc("google.com")/descendant::field()[1]/{"Oxford"}/following::field()[1]/{click /}/*{3,8}.

To limit the range of the Kleene star, one can specify upper and lower bounds on the multiplicity, e.g., {(#="Next")/}{click /}.

REFERENCES
The nesting in the result mirrors the structure of the OXPath expression: extraction markers in a predicate (title, source) represent attributes to the last marker outside the predicate (story).

Kleene Star. Finally, we add the Kleene star, as in [12]. For example, the following expression queries Google for "Oxford", traverses all accessible result pages and extracts all links.

doc("google.com")/descendant::field()[1]/"Oxford"/following::field()[1]/click/{click/}(descendant::a:[@href]/*[#="Next"]/{click/})*/

to limit the range of the Kleene star, one can specify upper and lower bounds on the multiplicity, e.g., 

doc("zillow.com")/descendant::field()[1]/"Seattle"/following::*#GOButton/{click/}(descendant::input[@type='checkbox'][2]/uncheck/following::field()[1]/uncheck//div[.='Beds']/following-sibling::select/"3+"//a[./text()='More filters']/click/)//input[following-sibling[contains(.,"Multi Family")]/uncheck/)//div[.='Beds']/following-sibling::select/"3+"//a[./text()='More filters']/click/)//ul#search-results/li:<property>

References


Wrapper Babel

- Wrapper induction & data extraction systems
  - each invent their **own** wrapper language
  - often **separate** navigation and matching
- Main classes:
  1. pattern matching + imperative navigation
     - XPath
     - Finite & Tree Automata
     - Token Prefix/Suffix
  2. Datalog
     - E-Log (Lixto)
Why OXPath?

scalability

an XPath for data extraction

web applications

familiarity

simplicity
OXPath
OXPath = XPath + 4

action   extraction
style   iteration
Start at kayak.co.uk:

doc("kayak.co.uk")
Start at kayak.co.uk:

```javascript
doc("kayak.co.uk")
```

To select an airport, type a few letters and select from completion list
```html
//field().destination/{"Sea" /}
//div#smartbox//li[1]/{click /
```
Start at kayak.co.uk:

doc("kayak.co.uk")

To select an airport, type a few letters and select from completion list

//field().destination/{"Sea" />

//div#smartbox//li[1]/{click /}

Submit the form
<table>
<thead>
<tr>
<th>Price (£)</th>
<th>Airline</th>
<th>Take-off</th>
<th>Landing</th>
<th>Stops</th>
</tr>
</thead>
<tbody>
<tr>
<td>503</td>
<td>US Airways</td>
<td>LGW 09:40</td>
<td>SEA 20:55</td>
<td>1 19h 15m</td>
</tr>
<tr>
<td>503</td>
<td>US Airways</td>
<td>SEA 11:15</td>
<td>LHR 10:00</td>
<td>1 14h 45m</td>
</tr>
<tr>
<td>509</td>
<td>Expedia.co.uk</td>
<td>Majortavel.co.uk</td>
<td>Lastminute.com</td>
<td>Expedia.co.uk</td>
</tr>
</tbody>
</table>

*Found on 4 Websites*
Refine the results by unchecking the “2+ stops”:

```javascript
///*#stops2/ {uncheck}
```
Refine the results by unchecking the "2+ stops": `//*#stops2/{uncheck }`

On all result pages
`/(//a[.='Next']/{click /})/*`
Refine the results by unchecking the “2+ stops”:
/**#stops2/ {uncheck }**

On all result pages
/**(//a[.='Next']/{click /})/**

and for each flight
/ //body.resultrow:**<flight>**
Extract the attributes
- Extract the attributes
- Mouseover the ! to extract flight quality warnings

```
//span.qualityWarningIcon/{mouseover }/}
```
Extract the attributes
Mouseover the ! to extract flight quality warnings
Click on the details to extract layovers
Actions correspond to **DOM events**, e.g.,

<table>
<thead>
<tr>
<th>Document</th>
<th><code>doc(&quot;rightmove.co.uk&quot;)</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>Click</td>
<td><code>{click}</code></td>
</tr>
<tr>
<td>Fill</td>
<td><code>{&quot;Sea&quot;}</code></td>
</tr>
<tr>
<td>Mouseover</td>
<td><code>{mouseover}</code></td>
</tr>
</tbody>
</table>

- Executed once on **each context node**
- Return **context nodes** (contextual actions) or **root nodes for new DOM** (absolute actions)
Extraction: Compact Tree Construction

- Extraction marker **select nodes for extraction**
  - record markers: `<flight>`
  - attribute markers: `<price=string(.)>`

- Extracted data **has tree shape**
  - nesting of extraction markers in OXPath expression defines
    nesting of records and attribute-record associations in the output
2 Extraction: Compact Tree Construction

- Extraction marker **select nodes for extraction**
  - record markers:  `<flight>`
  - attribute markers:  `<price=string(.)>`

- Extracted data **has tree shape**
  - nesting of extraction markers in OXPath expression defines
  - nesting of records and attribute-record associations in the output

```xml
<story>
  <title>Tax cuts ...</title>
  <source>Washington Post</source>
  <source>Wall Street Journal</source>
</story>
```
Most web sites use **pagination for results**

- traversing paginated results require **iteration**
- extraction from any **unbounded component** of a link graph

**Kleene Star from Regular XPath [Marx TODS ‘05]**

- extended to OXPath, i.e., with action in the iterated expression
  
  ```
  //a[.='Next']/{click /})*
  ```

**OXPath’s Page-at-a-time algorithm**

- buffers in practice only a constant number of pages
- even for very large components
4 Style: Querying Visual Attributes

Access to all CSS properties via **style axis**

<table>
<thead>
<tr>
<th></th>
<th><strong>style::display</strong> or <strong>style::visibility</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Visibility</td>
<td></td>
</tr>
<tr>
<td>Font size</td>
<td><strong>style::font-size</strong></td>
</tr>
<tr>
<td>Geometry</td>
<td><strong>style::top, style::left, ...</strong></td>
</tr>
<tr>
<td>Color</td>
<td><strong>style::color</strong> or <strong>style::background-color</strong></td>
</tr>
</tbody>
</table>

**Joins** on style properties possible

- but: no rich spatial relations as in SXPath
OXPath = XPath + 4

- action
- extraction
- style
- iteration
Analysis
Semantics – Overview

- Semantics defined over **relational structure**
  - as in XPath Leashed [Benedikt 08]
  - but: extended to **multiple** documents and with **action relations**
    - action relations form tree → no two actions lead to same DOM
  - but: instead of single node set, set of relations for **extracted tree**

- Context extended by last match for **parent extraction marker**
  - necessary to construct the extraction tree
  - last match for **sibling extraction** used for
    - **compact** specification of records with mixed attributes and nested records
\[
\begin{align*}
\text{N1} & \quad \left[\text{estep/path}\right]_N(c) = \{c'' \mid c' \in \left[\text{estep}\right]_N(c) \land c'' \in \left[\text{path}\right]_N(c')\} \\
\text{N2} & \quad \left[\text{axis::nodes}\right]_N(c) = \{\langle n', c.p, c.l \rangle \mid R_{\text{axis}}(c.n, n') \land n' \in R_{\text{nodes}}\} \\
\text{N3} & \quad \left[\text{step}[q]\right]_N(c) = \{c' \in \left[\text{step}\right]_N(c) \mid \left[\text{q}\right]_B(\langle c'.n, c'.l, c'.l\rangle)\} \\
\text{N4} & \quad \left[\text{step}^\pm[qp]\right]_N(c) = \{c' \in \left[\text{step}^\pm\right]_N(c) \mid C = \left[\text{step}^\pm\right]_N(c) \land \left[\text{REWRITE}^\pm(qp, C, c')\right]_B(\langle c'.n, c'.l, c'.l\rangle)\} \\
\text{N5} & \quad \left[\{\text{action} /\}\right]_N(c.n) = \{\langle n', c.p, c.l \rangle \mid n' \text{ such that } R_{\text{action}}(c.n, n')\} \\
\text{N6} & \quad \left[\{\text{action}\}\right]_N(c) = \left[\text{AFP}(\text{action}, c.n)\right]_N(\left[\{\text{action} /\}\right]_N(c)) \\
\text{N7} & \quad \left[\text{step} : \langle M[= v]\rangle\right]_N(c) = \{\langle c'.n, c'.p, \text{OUT}(c'.n, M)\rangle \mid \exists c' \in \left[\text{step}\right]_N(c)\} \\
\text{N8} & \quad \left[\text{(path)}^\ast\right]_N(c) = \{c_r \mid \exists r \geq 0 \forall 0 \leq s \leq r : c_{s+1} \in \left[\text{path}\right]_N(c_s) \land c_0 = c\} \\
\text{N9} & \quad \left[\text{(path)}^\ast\{v, w\}\right]_N(c) = \{c_r \mid \exists v \leq r \leq w \forall 0 \leq s \leq r : c_{s+1} \in \left[\text{path}\right]_N(c_s) \land c_0 = c\} \\
\text{N10} & \quad \left[\text{expr}[qp]\right]_N(c) = \{c' \mid c' \in C \land C = \left[\text{expr}\right]_N(c) \land \left[\text{REWRITE}^+(qp, C, c')\right]_B(\langle c'.n, c'.l, c'.l\rangle)\} \\
\text{N11} & \quad \left[\text{doc(uri)}\right]_N(c) = F_{\text{doc}}(uri)
\end{align*}
\]
### Table of XPath Expressions

<table>
<thead>
<tr>
<th>Expression</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ estep/path ] (N) ((c))</td>
<td>({c''</td>
</tr>
<tr>
<td>[ axis::nodes ] (N) ((c))</td>
<td>({\langle n', c.p, c.l \rangle</td>
</tr>
<tr>
<td>[ step[q] ] (N) ((c))</td>
<td>({c' \in [ step ] (N) ((c))</td>
</tr>
<tr>
<td>[ step\pm[qp] ] (N) ((c))</td>
<td>({c' \in [ step\pm ] (N) ((c))</td>
</tr>
<tr>
<td>[ {action /} ] (N) ((c.n))</td>
<td>({\langle n', c.p, c.l \rangle} ) with (n') such that (R_{action}(c.n, n'))</td>
</tr>
<tr>
<td>[ {action} ] (N) ((c))</td>
<td>([ AFP(action, c.n) ] (N) ([ {action /} ] (N) ((c))) )</td>
</tr>
<tr>
<td>[ step : \langle M[= v] \rangle ] (N) ((c))</td>
<td>({\langle c'.n, c'.p, OUT(c'.n, M)}</td>
</tr>
<tr>
<td>( (path)* ) (N) ((c))</td>
<td>({c_r \mid \exists r \geq 0 \forall 0 \leq s \leq r : c_{s+1} \in [ path ] (N) ((c_s) \land c_0 = c})</td>
</tr>
<tr>
<td>( (path)* {v, w} ) (N) ((c))</td>
<td>({c_r \mid \exists v \leq r \leq w \forall 0 \leq s \leq r : c_{s+1} \in [ path ] (N) ((c_s) \land c_0 = c})</td>
</tr>
<tr>
<td>( (expr)[qp] ) (N) ((c))</td>
<td>({c'</td>
</tr>
<tr>
<td>[ doc(uri) ] (N) ((c))</td>
<td>(F_{doc}(uri))</td>
</tr>
<tr>
<td>Rule</td>
<td>Definition</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
</tr>
<tr>
<td>N1</td>
<td>([ estep/path ]_N (c) = { c''</td>
</tr>
<tr>
<td>N2</td>
<td>([ axis::nodes ]_N (c) = { \langle n', c.p, c.l \rangle</td>
</tr>
<tr>
<td>N3</td>
<td>([ step[q] ]_N (c) = { c' ∈ [ step ]_N (c)</td>
</tr>
<tr>
<td>N4</td>
<td>([ step_±[qp] ]<em>N (c) = { c' ∈ [ step</em>± ]_N (c)</td>
</tr>
<tr>
<td>N5</td>
<td>([ { action / } ]<em>N (c.n) = { \langle n', c.p, c.l \rangle } ) with ( n' ) such that ( R</em>{action} (c.n, n') )</td>
</tr>
<tr>
<td>N6</td>
<td>([ { action } ]_N (c) = [ \text{AFP} (action, c.n) ]_N ( [ { action / } ]_N (c) ) )</td>
</tr>
<tr>
<td>N7</td>
<td>([ step : \langle M[= v] \rangle ]_N (c) = { \langle c'.n, c'.p, \text{OUT}(c'.n, M) \rangle</td>
</tr>
<tr>
<td>N8</td>
<td>([ (path)* ]_N (c) = { c_r</td>
</tr>
<tr>
<td>N9</td>
<td>([ (path)* { v, w } ]_N (c) = { c_r</td>
</tr>
<tr>
<td>N10</td>
<td>([ (expr)[qp] ]_N (c) = { c'</td>
</tr>
<tr>
<td>N11</td>
<td>([ doc(uri) ]<em>N (c) = F</em>{doc} (uri) )</td>
</tr>
</tbody>
</table>
Recently, the notion of self-similarity has been shown to apply to World Wide Web (WWW) transfers. In this paper, we show evidence that the subset of network traffic that is due to World Wide Web (WWW) transfers can show characteristics that are consistent with self-similarity. Despite its increasing role in communication, the World Wide Web (WWW) remains the least controlled medium: any individual or institution can create websites with unrestricted access. This paper seeks to describe the aims, data model, and protocols needed to implement the "information universe" into existence using available technology.
Abstract—Recently, the notion of self-similarity has been shown to apply to local-area network traffic. In this paper, we show evidence that the subset of network traffic that is due to World Wide Web (WWW) transfers can show characteristics that are consistent with a fractal distribution. Despite its increasing role in communication, the World Wide Web (WWW) remains the least controlled medium: any individual or institution can create websites with unrestricted access. Weaving the information universe into existence using available technology. This paper seeks to describe the traps that result in poorly designed sites. It's for experienced designers who have designed websites and for novice inexperienced designers who are interested in understanding the people who have already designed sites. The diameter of the World Wide Web (WWW) initiative is a practical project designed to bring a global information universe into existence using available technology. This paper seeks to describe how the concept of self-similarity is due to World Wide Web (WWW) transfers can show characteristics that are consistent with a fractal distribution.
The diameter of the *world wide web*

R Albert, H Jeong, AL Barabási - Arxiv preprint cond-mat/9907038, 1999 - arxiv.org

Despite its increasing role in communication, the world wide web is a controlled medium: any individual or institution can create websites with unrestricted access to the information universe into existence using available technology. This paper seeks to describe the aims, data model, and protocols needed to implement the world wide web.

**Abstract**—Recently, the notion of self-similarity has been shown to apply to the traffic of the World Wide Web (WWW). In this paper, we show evidence that the subset of network traffic that is due to World Wide Web (WWW) transfers can show characteristics that are consistent with self-similarity. The diameter of the wide web remains the least explored region of the information universe, where web designers who, from the start, want to avoid local-area network traffic. In this paper, we show evidence that the traffic is due to World Wide Web (WWW) transfers can show characteristics consistent with self-similarity.

**Self-similarity in World Wide Web traffic: evidence and possible causes**

ME Crovella, A Bestavros - Networking, IEEE/ACM ..., 2002 - ieeexplore.ieee.org

Abstract—Recently, the notion of self-similarity has been shown to apply to the traffic of the World Wide Web (WWW). In this paper, we show evidence that the subset of network traffic that is due to World Wide Web (WWW) transfers can show characteristics consistent with self-similarity.
Summary of Complexity

Combined: $\text{PTIME-hard}$

Data: $\text{NLOGSPACE}$
Summary of Complexity

Combined: $\texttt{PTIME}$-hard $\texttt{PTIME}$-hard

Data: $\texttt{NLogSpace}$ $\texttt{LogSpace}$
Summary of Complexity

Combined: **PTIME-hard**

Data: **NLogSpace**  **LogSpace**

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>OXPath <strong>w/o Actions &amp; Kleene</strong></td>
<td>O(</td>
<td>O(</td>
</tr>
</tbody>
</table>
### Summary of Complexity

**Combined:** $\text{PTIME-hard}$  $\text{PTIME-hard}$

**Data:** $\text{NLogSpace}$  $\text{LogSpace}$

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>OXPath $w/o$ Actions &amp; Kleene</td>
<td>$O($ $n^4 \cdot q^2)$</td>
<td>$O($ $n^3 \cdot q^2)$</td>
</tr>
</tbody>
</table>
Summary of Complexity

Combined: **PTIME-hard**

Data: **NLOGSPACE**  **LOGSPACE**

<table>
<thead>
<tr>
<th>OXPath <strong>w/o Actions &amp; Kleene</strong></th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$O(n^3 \cdot q^2)$</td>
<td>$O(n^3 \cdot q^2)$</td>
</tr>
</tbody>
</table>
Summary of Complexity

Combined: \textbf{PTIME-hard} \textbf{PTIME-hard}

Data: \textbf{NLogSpace} \textbf{LogSpace}

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>OXPath \textbf{w/o Actions &amp; Kleene}</td>
<td>(O(n^4 \cdot q^2))</td>
<td>(O(n^3 \cdot q^2))</td>
</tr>
<tr>
<td>OXPath \textbf{w/o Kleene}</td>
<td>(O(n^3 \cdot q^2))</td>
<td>(O(n \cdot q^2))</td>
</tr>
</tbody>
</table>
Summary of Complexity

Combined: **PTIME-hard**

Data: **NLogSpace**  **LogSpace**

<table>
<thead>
<tr>
<th>OXPath w/o Actions &amp; Kleene</th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>O( ( n^4 \cdot q^2 ) )</td>
<td>O( ( n^3 \cdot q^2 ) )</td>
<td></td>
</tr>
</tbody>
</table>

Actions = multiple pages
Summary of Complexity

Combined: **PTIME-hard** **PTIME-hard**

Data: **NLogSpace** **LogSpace**

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>OXPath w/o Actions &amp; Kleene</td>
<td>$O(\cdot)$</td>
<td>$O(n^3 \cdot q^2)$</td>
</tr>
<tr>
<td>OXPath w/o Kleene</td>
<td>$O(n^4 \cdot q^2)$</td>
<td>$O(n^3 \cdot q^2)$</td>
</tr>
</tbody>
</table>

Contextual actions (action free prefix)
Summary of Complexity

**Combined:** \textbf{PTIME-hard PTIME-hard}

**Data:** \textbf{NLogSpace LogSpace}

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>OXPath \textit{w/o Actions &amp; Kleene}</td>
<td>$O(\quad)$</td>
<td>$O(n^4 \cdot q^2)$</td>
</tr>
<tr>
<td>OXPath \textit{w/o Kleene}</td>
<td>$O(\quad)$</td>
<td>_</td>
</tr>
<tr>
<td>OXPath \textit{w/o unbounded Kleene}</td>
<td>$O(\quad)$</td>
<td>_</td>
</tr>
</tbody>
</table>
## Summary of Complexity

### Combined: PTIME-hard

**Data:** NLogSpace   LogSpace

<table>
<thead>
<tr>
<th>OXPath</th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o Actions &amp; Kleene</td>
<td>$O(n)$</td>
<td>$O(n^4 \cdot q^2)$</td>
</tr>
<tr>
<td>w/o Kleene</td>
<td>$O(n)$</td>
<td>$O(n^4 \cdot q^2)$</td>
</tr>
<tr>
<td>w/o unbounded Kleene</td>
<td>$O(n)$</td>
<td>$O(n^3 \cdot q^2)$</td>
</tr>
<tr>
<td>(full)</td>
<td>$O(n)$</td>
<td>$O(n^3 \cdot q^2)$</td>
</tr>
</tbody>
</table>
### Summary of Complexity

**Combined:** PTIME-hard

**Data:** NLogSPACE LogSPACE

<table>
<thead>
<tr>
<th>OXPath</th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>w/o Actions &amp; Kleene</strong></td>
<td>O( )</td>
<td>O(n^4 \cdot q^2)</td>
</tr>
<tr>
<td><strong>w/o Kleene</strong></td>
<td>O( )</td>
<td>O( )</td>
</tr>
<tr>
<td><strong>w/o unbounded Kleene</strong></td>
<td>O( )</td>
<td>O( )</td>
</tr>
<tr>
<td><strong>(full)</strong></td>
<td>O( )</td>
<td>O( )</td>
</tr>
</tbody>
</table>

**Buffer bounded by page depth**
Page-At-A-Time (PAAT)

**node-set semantics**

**extraction semantics**

**complexity close to XPath**

**parallelizable**
Evaluation
Appendix B shows that memory is constant with respect to pages visited from different web shops. Figures 4rcs and firs show the results, and figure 3rcs and 3firs pages with many results. The memory use is due to the repeated ascends and descends of the citation hierarchy. Figure 4rbs shows the same test simulating Google's memory use.

OXPath extracts over 6 million pieces of data from 3rz pages on our web server to avoid overly taxing Google's memory hierarchy. Figure 4rbs shows the same test simulating Google's memory use is due to the repeated ascends and descends of the citation hierarchy. The jitter in time, but memory size remains constant even as both the number of pages and results extracted are linear. The browser rendering time, even for complex queries on small web pages, affects the scaling behaviour or the dominance of page rendering. None of the extensions of OXPath in Section fix1s significantly affects the scaling behaviour or the dominance of page rendering. Pages retrieved and results extracted are linear, but memory size remains constant even if extracting millions of records from hundreds of thousands of web pages.

We illustrate its scaling behaviour by evaluating three classes of queries on Seattle r[4sx On each v we click on all links and extract the visualization of each result page. The first class v which searches for papers on “Seattle” in Google Scholar and repeatedly clicks on the “cited by” links of all results. The second experiment analyses the effect of OXPath's actions on query evaluation and the absolute actions. Our test performs actions on pages that do not result in new page retrievals. Figure 5rcs shows the results with many pages reached by the Google product search form. contextual actions suffer a small but insignificant penalty to evaluation time compared to their absolute equivalents. The second experiment analyses the effect of OXPath’s actions on query evaluation and the absolute actions. Our test performs actions on pages that do not result in new page retrievals. Figure 5rcs shows the results with many pages reached by the Google product search form. contextual actions suffer a small but insignificant penalty to evaluation time compared to their absolute equivalents.

In an extensive comparison with commercial or academic tools, we do not consider tools such as Lixto goes considerably beyondv Rickenfoot and Web Harvester. The open source extraction toolkit Web Harvest, as the first academic web automation and extraction system Rickenfoot and Lixto, Visual Web Ripper (VWR) and the
diadem-project.info tools are comparatively large and contain many links. Needed are scripting to traverse to new web pages. We also do not consider tools such as uScripter and iMacros needed. We do not consider tools such as uScripter and iMacros as they focus on automation only and offer no iterative constructs needed.

We profile each component of OXPath's evaluation, including browser initialization, PAAT (2%), extracted matches, visited pages, and time taken. Figure 6: Profiling OXPath's components. We conduct similar tests repeatedly clicking on all links for many pages. Many pages are involved, and lack the ability to traverse to new web pages as required for extraction tasks. We also do not consider tools such as uScripter and iMacros, which focus on automation only and offer no iterative constructs needed.
100,000+ pages, millions of results

Constant Memory

memory [MB]

extracted matches

visited pages

#pages [1000] / #results [100,000]

memory

time [h]

#matches [100] / #pages
it’s the browser

- page rendering
- browser initialization
- OXPath

- 2%
- 13%
- 85%
6. CONCLUSION AND FUTURE WORK

A crucial part of the toolset of developers interacting with the web is extracting data, and we believe that it can become an important part of the toolset of developers interacting with the web. In our experimental evaluation, we have shown that OXPath can achieve significant memory savings compared to other systems, with a clear linear trend. Even Web Harvest shows a clear linear trend, but the resulting program was too slow for the tests at any time. We also tried to simulate proper multi-way navigation, but requires only a single active OXPath instance to preserve page state. When a page is no longer cached, it does not need to be reloaded. This forces reloading when a page is no longer cached and does not require a ' ava', P-based system, but omits W' R and VWR as they show a clear linear trend in memory usage in the tests we were able to run. Among the systems that were considered, Web Harvest comes close to the memory usage of OXPath, but they were not able to run these tests. Again, both figures show a considerable advantage for OXPath, at least one order of magnitude. According to the image, the evaluation time for each system was recorded. Figure 3pcq shows the average evaluation time for each system, which is even more pronounced. With each system, we navigate the citation graph to a depth of five for papers on "Seattle" perpendix). We are committed to building a set of tools around OXPath to the best of our knowledge, OXPath is the first web extractor that provides a visual generator for OXPath expressions and allows for more expressive visual features and multi-property axes. The execution of actions on context sets with many nodes is an open issue.)3<

7. REFERENCES


6. CONCLUSION AND FUTURE WORK

Important part of the toolset of developers interacting with the webview.

Chickenfoot but the resulting program was too slow for the tests.

We also tried to simulate proper multi-way navigation.

This forces reloading when a page is no longer cached and does not...

We compensate in the experiments by using the browser's history.

Due to Chickenfoot's lack of support for multi-way navigation that...

Times more memory in absolute terms.

The constant memory is a constant memory just as OXPath through it uses about ten...

Even Web Harvest shows a clear linear trend.

Chickenfoot exhibits OXPath which is not surprising as it does not render pages.

Yet again XMPat which is not surprising as it does not render pages.

Yett in Figure 4t only Web Harvest comes close to the memory usage of...

Memory usage in the tests we were able to run.

Among the systems finally Figure 4 illustrates the memory use of these systems.

Again significant advantage for OXPath pat least one order of magnitude.

They were not able to run these tests.

Again both figures show a consu...

Differently browser engines or web cleaning approaches used in the sysu...

Graph to a depth of fi for papers on "Seattle" psee, appendix Bqv.

We are committed to building a strong set of tools around OXPath.

We record evaluation time and memory for each system.

&n Figure 8: Comparison: Memory...
6. CONCLUSION AND FUTURE WORK

Important part of the toolset of developers interacting with the web. In our experimental evaluation, we believe that it can become an important tool in the development process...

We are committed to building a strong set of tools around OXPath. To the best of our knowledge, OXPath is the first web extraction tool that...

We record evaluation time and memory for each system. Figure 4 illustrates the memory use of these systems. Again, both figures show a consistent advantage for OXPath, at least one order of magnitude, over the other systems.

We plan to further investigate language features such as the one we plan to address in future work are...
6. CONCLUSION AND FUTURE WORK

We are committed to building a strong set of tools around OXPath. To the best of our knowledge, OXPath is the first web extraction tool that integrates efficient and expressive queries into a visual web cleaning system that can be combined with arbitrary web cleaning approaches. Our approach is more expressive than traditional visual web cleaning tools, which either lack flexibility or are not as expressive as XPath. OXPath is amenable to significant optimization and a good target for automated generation.

Some of the issues raised by OXPath are:

1. A Java-based toolset for highly parallel execution of actions on context sets with many nodes is an open issue.
2. Different bindings for the same language are suited for highly parallel execution.
3. The execution of actions on context sets with many nodes is an open issue.
4. The performance advantage of OXPath is heavily scripted.

The evaluation of web extraction programs is an important part of the toolset of developers interacting with the web. The resulting program was too slow for the tests, but requires only a single active OXPath instance. We compensate in the experiments by using the browser's history, which is not surprising as it does not render pages. Yet, in Figure 4, only Web Harvest comes close to the memory usage of OXPath. OXPath, Web Harvest, and Lixto exclude W"R and VWR as systems affect the overall runtime considerably. Figure 3pq shows the time required to navigate the citation graph to a depth of five for papers on "Seattle." Figure 7 shows the number of pages required to perform the evaluation time discounting the page loading, cleaning, and rendering process. This allows for a more balanced comparison as the different browser engines or web cleaning approaches used in the systems affect the overall runtime considerably.

We record evaluation time and memory for each system. Figure 8 shows the number of pages required to perform the time required to navigate the citation graph to a depth of five for papers on "Seattle." We plan to further investigate language featurest such as more expressive visual features and multi-property axes.
6. CONCLUSION AND FUTURE WORK

important part of the toolset of developers interacting with the web. In our experimental evaluation, we believe that it can become an essential system with strict memory guarantees, which reflect strongly on the performance of the systems.

Due to Chickenfoot’s lack of support for multi-way navigation, the resulting program was too slow for the tests. We also tried to simulate proper multi-way navigation for a more balanced comparison, but it requires only a single active-OM instance. This forces reloading when a page is no longer cached and does not preserve page state, but the resulting program was too slow for the tests at any time.

We also tried to simulate proper multi-way navigation but requires only a single active-OM instance. This forces reloading when a page is no longer cached and does not preserve page state, but the resulting program was too slow for the tests at any time.

We also tried to simulate proper multi-way navigation but requires only a single active-OM instance. This forces reloading when a page is no longer cached and does not preserve page state, but the resulting program was too slow for the tests at any time.

We also tried to simulate proper multi-way navigation but requires only a single active-OM instance. This forces reloading when a page is no longer cached and does not preserve page state, but the resulting program was too slow for the tests at any time. This allows for a more balanced comparison as the differences in evaluation time up to 41x pagest but omits W"R and VWR as different browser engines or web cleaning approaches used in the systems.

Finally, Figure 4 illustrates the memory use of these systems, which is not surprising as it does not render pages. Yet, W"R and VWR are excluded, but they show a clear linear trend in the evaluation time discounting the page loading, cleaning, and rendering times.

The linear trend in the evaluation time discounting the page loading, cleaning, and rendering times is significant advantage for OXPath. At least one order of magnitude of the evaluation time can be observed for OXPath even on this small number of pages. Figure 3 shows the normalized evaluation time for each system.

Figure 8: Comparison: Memory usage

[Graph showing memory usage across different systems (OXPath, Lixto, Web Harvest, Chickenfoot).]

We record evaluation time and memory for each system. In Figure 3paq, we show the page averaged evaluation time for each system.}

**References**


6. CONCLUSION AND FUTURE WORK

In our experimental evaluation, we believe that it can become an alternative to other web extraction tools. However, the resulting program was too slow for the tests we conducted. We also tried to simulate proper multi-way navigation, but due to Chickenfoot's lack of support for multi-way navigation that is common in web pages, we compensate in the experiments by using the browser's history.

Even Web Harvest shows a clear linear trend. Chickenfoot exhibits an increase in memory usage in the tests we were able to run. Among the systems we tested, W"R and VWR are excluded, but they show a clear linear trend in the data. Finally, Figure 4 illustrates the memory use of these systems. Again, both figures show a considerable advantage for OXPath, at least one order of magnitude.

They were not able to run these tests. Again, both figures show a considerable advantage for OXPath, at least one order of magnitude. The memory usage was not constant in our tests, and we found that cleaning and refreshing processes such web pages are impossible. All other tools and browser state or do not render pages at all. OXPath and Web Harvest do not manage page loading, and browser state or do not render pages at all. OXPath is amenable to significant optimization and a good target for automated generation.

To the best of our knowledge, OXPath is the first web extraction program that we plan to address in future work. The language featurest such as variable can be filled into forms in parallel. The effective parallelization of web extraction programs is an open issue.

With each system, we navigate the citation graph to a depth of fi for papers on "Seattle" (see Appendix 5). A similar evaluation time up to 41x pages but omits W"R and VWR as systems affect the overall runtime considerably.

Figure 3 shows the execution of actions on context sets with many nodes is an open issue. The data suggest that we plan to address in future work are.

We are committed to building a strong set of tools around OXPath. To the best of our knowledge, OXPath is the first web extraction program that we plan to address in future work. The language features such as the ability to fill variables into forms in parallel. The effective parallelization of web extraction programs is an open issue.

7. REFERENCES

For legacy web data sources using Wfl, for highly parallel execution, different bindings for the same XPath expression can be generated. The effective parallelization of web extraction programs is an open issue.
Minimal

constant memory

minimal page buffer

browser bound

very low overhead
DEQA
Deep Web Q&A
Question Answering on the Deep Web

1. Extraction from Unstructured and Structured Data
   - URI1
     - rdf:type
     - "x y z..."
   - some:Class
     - 123,45

2. Data Integration and Enrichment
   - URI2
     - rdf:type
     - "x y z..."
   - some:Class
     - 123,45
   - URI3
     - rdf:type
     - some:Class
   - URI4

3. Web-Application
   - Regular Answer
   - Structured Query/Answer
   - Fallback
   - Search
   - Question Answering

4. Question Answering
**Question:**
House near a Kindergarden under 2,000,000 £?

**Answer:**

```
White_Road
  -- dd:hasPrice
    ↓ 1,499,950 £
  -- dbp:near
    Kindergarden_A
```

```
OXPath
```

```
White_Road
```

```
1,499,950 £
```

```
Kindergarden_B
```

```
Kindergarden_A
```

```
Linking-Metric
```

```
OXPath
```

```
TBSL
```

```
Domain Specific Triple Store
```

```
LIMES
```

```
Linked GeoData
```

```
White_Walls
```

```
Property in Oxford search results
```

```
Question:
House near a Kindergarden under 2,000,000 £?

Answer:

```
White_Road
  -- dd:hasPrice
    ↓ 1,499,950 £
  -- dbp:near
    Kindergarden_A
```

```
Kindergarden_B
```

```
Kindergarden_A
```

```
Linking-Metric
```

```
OXPath
```

```
TBSL
```

```
Domain Specific Triple Store
```

```
Linked GeoData
```

```
White_Walls
```

```
Property in Oxford search results
```
The property values of all schools from LinkedGeoData that were found to be close to houses extracted by OXPath were subsequently retrieved by Limes and loaded into the deqa triple store.

2.3 TBSL Question Answering

Figure 4 gives an overview of our TBSL (template based SPARQL query generator) question answering approach [27]. The system takes a natural language question as input and returns a SPARQL query and the corresponding answer(s) as output. First, the natural language question is parsed on the basis of its part-of-speech tags and a few domain-independent lexical entries comprising wh-words, determiners, numerals, and so on. The result is a semantic representation of the natural language query, which is then converted into a SPARQL query template. This template fixes the overall structure of the target query, including aggregation functions such as filters and counts, but leaves open
All houses in Abingdon with more than 2 bedrooms

```
SELECT ?y WHERE {
  ?y a <http://diadem.cs.ox.ac.uk/ontologies/real-estate#House> .
  ?y <http://diadem.cs.ox.ac.uk/ontologies/real-estate#bedrooms> ?y0 .
  FILTER(?y0 > 2) .
  FILTER(regex(?y1,'Abingdon','i')) .
}
```
Question Answering on the Deep Web

Edwardian houses close to supermarket for less than 1,000,000 in Oxfordshire

```
SELECT ?x0 WHERE {
  ?x0 a <http://diadem.cs.ox.ac.uk/ontologies/real-estate#House> .
  ?v <http://purl.org/goodrelations/v1#include> ?x0 .
  ?x0 <http://www.w3.org/2006/vcard/ns#street-address> ?y0 .
  ?v <http://diadem.cs.ox.ac.uk/ontologies/real-estate#hasPrice> ?y1 .
  ?x0 <http://purl.org/goodrelations/v1#description> ?y .
  FILTER(regex(?y0,'Oxfordshire','i')) .
  FILTER(regex(?y,'Edwardian ','i')) .
  FILTER(?y1 < 1000000) .
}
```
Preliminary Evaluation

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>number of questions</td>
<td>100</td>
</tr>
<tr>
<td>— SPARQL queries created</td>
<td>71</td>
</tr>
<tr>
<td>— SPARQL queries returning results</td>
<td>63</td>
</tr>
<tr>
<td>— SPARQL queries with correct results</td>
<td>49</td>
</tr>
<tr>
<td>— exactly intended SPARQL query</td>
<td>30</td>
</tr>
<tr>
<td>— SPARQL queries with incorrect results</td>
<td>14</td>
</tr>
</tbody>
</table>

(a) Evaluation results

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>failures</td>
<td></td>
</tr>
<tr>
<td>— data coverage</td>
<td>9</td>
</tr>
<tr>
<td>— linguistic coverage</td>
<td>18</td>
</tr>
<tr>
<td>— POS tagging</td>
<td>2</td>
</tr>
<tr>
<td>— other reasons</td>
<td>6</td>
</tr>
</tbody>
</table>

(b) Failure reasons

after roughly 3 days of integration with 4 people

>75% with 2-3 weeks
OXPath Suite
3.4. REFINEMENT

Figure 3.6: User Interface – Before Expression Refinement

Figure 3.7: User Interface – After Expression Refinement
Visual OXPath: Semi-supervised Generation

Figure 3.6: User Interface – Before Expression Refinement

Figure 3.7: User Interface – After Expression Refinement

Interaction recording

Did you mean: XPath

Finding an OXPath to Cherries Hidden in the Scripted Web
T Furche, G Gottlob, G Grasso, C Schallhart... - ... rep., diadem-project info/oxpath
Cited by 2 - Related articles

Taking the oxpath down the deep web
A Sellers, T Furche, G Gottlob... - Proceedings of the ... 2011 - portal.acm.org
ABSTRACT Although deep web analysis has been studied extensively, there is no succinct formalism to describe user interactions with AJAX-enabled web applications. Toward this end, we introduce OXPath as a superset of XPath 1.0. Beyond XPath, OXPath is able (1) to fill web forms ...
Cited by 1 - Related articles - Check Availability - All 11 versions

Oxpath: little language, little memory, great value
A Sellers, T Furche, G Gottlob... - Proceedings of the ... 2011 - portal.acm.org
ABSTRACT Data about everything is readily available on the web—but often only accessible through elaborate user interactions. For automated decision support, extracting that data is es-sential, but infeasible with existing heavy-weight data ex traction systems. In this ...
Cited by 1 - Related articles - Check Availability - All 11 versions

OXPath: A language for scalable, memory-efficient data extraction from web applications
T Furche, G Gottlob, G Grasso, C Schallhart... - Proceedings of the VLDB ... 2011 - fox7.eu
ABSTRACT The evolution of the web has outpaced itself. The growing wealth of information ...
Visual OXPath: Semi-supervised Generation

Generation of OXPath expressions ranked by robustness & specificity

Taking the xpath down the deep web
A Sellers, T Furche, G Gottlob... - Proceedings of the ... 2011 - portal.acm.org

OXPath: little language, little memory, great value
A Sellers, T Furche, G Gottlob... - Proceedings of the ... 2011 - portal.acm.org

OXPath: A language for scalable, memory-efficient data extraction from web applications
T Furche, G Gottlob, C Schallhart... - Proceedings of the VLDB ... 2011 - fox7.eu

Extract publication at //div[@class='gs_r']
Marker publication
OXPath //div[@class='gs_r']
OXPath selects 10 nodes

Score Generator OXPath Count
403 position //div[@class='gs_r'] 10
358 position //div[@class='gs_r'] 10
250 position //div 34
166 position //div 34
100 canonical //html[1]/body[1]/div 15

Citation Finding an OXPath to Cherries Hidden in the Scripted Web
T Furche, G Gottlob, G Grasso, C Schallhart... - diadem-project.info/oxpath
Cited by 2 - Related articles - Check Availability - All 11 versions

Citation A language for scalable, memory-efficient data extraction from web applications
T Furche, G Gottlob, C Schallhart... - Proceedings of the VLDB ... 2011 - fox7.eu
Cited by 1 - Related articles - Check Availability - All 11 versions
OXPath Tracer: Debugging
OXPath Tracer: Debugging

Track selected **DOM nodes**
OXPath Tracer: Debugging

Trace page management
OXPath Tracer: Debugging

Fine grained execution control (filter, step-by-step, breaks)
Scaling Out Wrapper Execution

- **Three forms** of parallelism for wrapper execution
  - **Wrapper** parallelism
    - “naive” distribution of *multiple* wrappers
    - or *multiple instances* of the same wrapper (parameters)
  - **Inter-site** parallelism
    - if a single wrapper touches many sites
  - **Intra-site** parallelism
    - if parallel wrappers touch many pages from the same site
Scaling Out Wrapper Execution

- **Three forms** of parallelism for wrapper execution
  - **Wrapper** parallelism
    - “naive” distribution of *multiple* wrappers
    - or *multiple instances* of the same wrapper (parameters)
  - **Inter-site** parallelism
    - if a single wrapper touches many sites
  - **Intra-site** parallelism
    - if parallel wrappers touch many pages from the same site
Scaling Out Wrapper Execution

- **Three forms** of parallelism for wrapper execution

- **Wrapper parallelism**
  - “naive” distribution of *multiple* wrappers
  - or *multiple instances* of the same wrapper (parameters)

- **Inter-site** parallelism
  - if a single wrapper touches many sites

- **Intra-site** parallelism
  - if parallel wrappers touch many pages from the same site
Scaling Out Wrapper Execution

- **Three forms** of parallelism for wrapper execution
  - **Wrapper** parallelism
    - “naive” distribution of **multiple** wrappers
    - or **multiple instances** of the same wrapper (parameters)
  - **Inter-site** parallelism
    - if a single wrapper touches many sites
  - **Intra-site** parallelism
    - if parallel wrappers touch many pages from the same site
Basic Example

wrapper := doc('rightmove.co.uk')/descendant::field()[1]/{$postcode}/following::field()[1]/{click /} //div.result:<property>[///tr.asking:<price=string(.)>]

location = LOAD 'postcodes.txt' AS (postcode);
results = FOREACH location GENERATE wrapper(postcode);

<table>
<thead>
<tr>
<th>ID</th>
<th>PARENT</th>
<th>TAG</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>⊤</td>
<td>property</td>
<td>⊥</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>price</td>
<td>£120</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>price</td>
<td>£230</td>
</tr>
</tbody>
</table>

Ox Latin » The Language
(De-) Composition
Why Decomposition?

- Analog to **multi-query planning**
  - "decompose" wrappers at **multi-way navigation** points
    - due to actions on large context-node set
  - "compose" shared prefixes of wrappers
    - e.g., due to form filling
  - also enables sharing inner parts or suffices, but far less common
  - both can be done with the same mechanic: **continuation**

- OXPath **continuation**
  - enables us to halt wrapper execution and resume later
  - here: approximated as location + parent for **absolute** actions only
    - fails in same (script) cases, but those are infeasible (browser)
Decomposition: Principle & Example

\[ h/\{\alpha\}/t \]

\[ h/\{\alpha\}:<\text{new_marker}=\text{location}()>/t \]

\[ \text{doc}($\text{new_variable}$)/t \]
Decomposition: Principle & Example

\[
\text{h/\{α\}/t} \quad \rightarrow \quad \text{h/\{α\}:<new_marker=location()> doc($new_variable)/t}
\]

\[
\text{head := doc("news.google.com")//a/{click/}//h3[contains(.,'News')]}<news=string(.)>
\]

\[
\text{tail := doc($prefix)//h3[contains(.,'News')]}<news=string(.)>
\]

\[
\text{head_results = GENERATE head();}
\]

\[
\text{tail_inputs = FOREACH head_results FILTER tag == "prefix" GENERATE parent, value as location;}
\]

\[
\text{tail_results = FOREACH tail_inputs GENERATE tail(parent,location);}
\]
DIADEM - domain-centric intelligent automated data extraction methodology