Exploratory research on language usage

Ralf Lämmel
Software Languages Team, University of Koblenz-Landau
Empirical language analysis in software linguistics

Jean-Marie Favre¹, Dragan Gasevic², Ralf Lämmel³, and Ekaterina Pek³

¹ OneTree Technologies, Luxembourg
² Athabasca University, Canada
³ Universität Koblenz-Landau, Germany

Abstract. Software linguistics is the science of software languages. In this short paper, we sketch the general discipline of software linguistics, but our focus is on one part of it: empirical analysis of software languages. Such analysis is concerned with understanding language usage on the grounds of a corpus. In this short paper, we sketch a survey on empirical language analysis, and we argue that the research method of content analysis is needed for a thorough survey.
SLE vs. SL

• Software Language Engineering (SLE), or "Software Languages are Software too"

• Software Linguistics* (SL), or "Software Languages are Languages too"

* -- [Misek-Falkoff82]
Software Linguistics

• Comparative linguistics
• Historical linguistics
• Geo-linguistics
• Socio-linguistics
• Corpus linguistics
Development of strong typing in programming languages happened cotemporally with software engineering research on reliability and later, software reuse. It seems clear that the concepts of strong typing and software reliability reinforced each other, especially since the early pioneers in imperative language design were also active in software engineering research. Further, as type systems for programming languages developed notions of genericity and polymorphism, these were directly related to issues of software reuse.

The main imperative languages designed in the 1960s and 1970s—Algol 60, Pascal, and PL/I—were designed at a time when the software engineering community was worried about the reliability of code. In the mid-1970s, Barry Boehm defined software engineering to be "the practical application of scientific knowledge in the design and construction of computer programs and the associated documentation required to develop, operate and maintain them" [Boehm 1976].

Peter Wegner [1984] talked of the 1950s as a time of stand-alone programs and the 1960s as a time of development of operating systems and databases. He stated that the 1970s saw the birth of software engineering, referring to similarities in the construction of large software systems and large physical structures such as bridges. The 1980s, in Wegner’s view, saw the development of interface technologies and the personal computer revolution. The 1990s was a time of knowledge engineering, the use of intelligent components to build systems (e.g., adaptation). In this temporal framework, the 1970s—the era of Algol 60 and Pascal—was a key time for software engineering, so that the design of strongly typed languages, which were type safe, coincided with the beginning of the discipline of engineering software.

C.A.R. Hoare, a member of the IFIP Working Group in Algol (WG 2.1) active in the original design of Algol 68 [Bergin and Gibson 1996], was the keynote ACM Transactions on Software Engineering and Methodology, Vol. 14, No. 4, October 2005.
The main imperative languages designed in the 1960s and 1970s—Algol 60, Pascal, and PL/I—were designed at a time when the software engineering community was worried about the reliability of code. In the mid-1970s, Barry Boehm defined software engineering to be "the practical application of scientific knowledge in the design and construction of computer programs and the associated documentation required to develop, operate and maintain them" [Boehm 1976].

Peter Wegner [1984] talked of the 1950s as a time of stand-alone programs and the 1960s as a time of development of operating systems and databases. He stated that the 1970s saw the birth of software engineering, referring to similarities in the construction of large software systems and large physical structures such as bridges. The 1980s, in Wegner’s view, saw the development of interface technologies and the personal computer revolution. The 1990s was a time of knowledge engineering, the use of intelligent components to build systems (e.g., adaptation). In this temporal framework, the 1970s—the era of Algol 60 and Pascal—was a key time for software engineering, so that the design of strongly typed languages, which were type safe, coincided with the beginning of the discipline of engineering software.

C.A.R. Hoare, a member of the IFIP Working Group in Algol (WG 2.1) active in the original design of Algol 68 [Bergin and Gibson 1996], was the keynote ACM Transactions on Software Engineering and Methodology, Vol. 14, No. 4, October 2005.
Fig. 1. Time line for data abstraction.

Corpus linguistics

Designing, producing, annotating, analyzing, and sharing corpora as supported, for example, by the European Language Resources Association (ELRA).

We will discuss “corpus linguistics” (engineering) elsewhere in this course.
What is empirical language analysis, anyway?
Survey

Full collection of 52 papers on empirical analysis
Selective collection -- 17 papers

http://toknow.sourceforge.net/
Research questions

- What are the characteristics of the corpora?
- What are the objectives of the empirical analysis?
- What are the analyses leveraged by the empirical analysis?
## Corpus characteristics

<table>
<thead>
<tr>
<th>Language</th>
<th>Sources</th>
<th>Items</th>
<th>Unit</th>
<th>Accessibility</th>
<th>Reproducibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlvesV05</td>
<td>SDF</td>
<td>8</td>
<td>27</td>
<td>grammars</td>
<td>partial</td>
</tr>
<tr>
<td>BaxterFNRSVM06</td>
<td>Java</td>
<td>17</td>
<td>56</td>
<td>projects</td>
<td>full</td>
</tr>
<tr>
<td>ChevanceH78</td>
<td>COBOL</td>
<td>1</td>
<td>50</td>
<td>programs</td>
<td>none</td>
</tr>
<tr>
<td>CollbergMS04</td>
<td>Java</td>
<td>1</td>
<td>1132</td>
<td>JAR files</td>
<td>full</td>
</tr>
<tr>
<td>CookL82</td>
<td>Pascal</td>
<td>1</td>
<td>264</td>
<td>programs</td>
<td>none</td>
</tr>
<tr>
<td>CranorESMC08</td>
<td>P3P</td>
<td>3</td>
<td>?</td>
<td>policies</td>
<td>full</td>
</tr>
<tr>
<td>GilM05</td>
<td>Java</td>
<td>4</td>
<td>14</td>
<td>projects</td>
<td>full</td>
</tr>
<tr>
<td>HageK08</td>
<td>Haskell</td>
<td>1</td>
<td>68000</td>
<td>compilations</td>
<td>none</td>
</tr>
<tr>
<td>Hahsler04</td>
<td>Java</td>
<td>1</td>
<td>988</td>
<td>projects</td>
<td>full</td>
</tr>
<tr>
<td>Hautus02</td>
<td>Java</td>
<td>2</td>
<td>9</td>
<td>packages</td>
<td>full</td>
</tr>
<tr>
<td>KimSNM05</td>
<td>Java</td>
<td>2</td>
<td>2</td>
<td>programs</td>
<td>full</td>
</tr>
<tr>
<td>Knuth71</td>
<td>Fortran</td>
<td>2</td>
<td>440</td>
<td>programs</td>
<td>none</td>
</tr>
<tr>
<td>LaemmelKR05</td>
<td>XSD</td>
<td>2</td>
<td>63</td>
<td>schemas</td>
<td>partial</td>
</tr>
<tr>
<td>LaemmelP10</td>
<td>P3P</td>
<td>1</td>
<td>3227</td>
<td>policies</td>
<td>full</td>
</tr>
<tr>
<td>ReayDM09</td>
<td>P3P</td>
<td>1</td>
<td>2287</td>
<td>policies</td>
<td>full</td>
</tr>
<tr>
<td>SaalW77</td>
<td>APL</td>
<td>6</td>
<td>32</td>
<td>workspaces</td>
<td>none</td>
</tr>
<tr>
<td>Visser06</td>
<td>XSD</td>
<td>9</td>
<td>9</td>
<td>schemas</td>
<td>full</td>
</tr>
</tbody>
</table>
Objectives

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>language adoption</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>language habits</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>language taming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>user feedback</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>language evolution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>user behavior</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>implementor feedback</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Objectives of the selected publications

Without going into detail, the available data caters for various observations. For instance, we realize that research on language adoption is generally not exercised for programming languages. It appears that online communications but not scientific publications are concerned with such adoption.

3.6 Analyses of the papers

Based on our preliminary analysis of the paper collection, we have come up with a simple hierarchical classification of typically automated analyses that are leveraged in the empirical research projects, see Fig. 6 for the classification, see Fig. 7 for corresponding metadata for the selective paper collection.

The presented classification focuses on prominent forms of static and dynamic analysis. In our paper collection, static analysis is considerably more common, and there is a substantial variety of different analyses. We also indicate two additional dimensions for analyses. An analysis is concerned with evolution when different versions of items, sources, or languages are considered. The dimension of clustering generalizes the linguistic of Sec. 4. By no means, our classification scheme is complete. For instance, we currently miss characteristics regarding data analysis in terms of the involved statistical methods, and the presentation of research results in terms of the leveraged tables, charts, etc.

4. Outlook on research methodology

The survey is work in progress. We are in the process of assessing the feasibility of such a survey, discovering research questions, and studying applicable research methodology.
### Analyses

#### Static analysis
- **Source code or other static entities are analyzed.**
- **Validity**
  - The validity of items in terms of syntax or type system is analyzed.
- **Metrics**
  - Some sort of metrics are analyzed.
  - **Size**
    - The size of items is analyzed, e.g., in terms of lines of code.
  - **Complexity**
    - The complexity of items is analyzed, e.g., the McCabe complexity.
- **Structural properties**
  - Example: the depth of inheritance hierarchy in OO programs.
- **Coverage**
  - The coverage of language constructs is analyzed.
- **Styles**
  - The usage of coding styles is analyzed.
- **Patterns**
  - The usage of patterns, e.g., design patterns, is analyzed.
- **Cloning**
  - Cloning across items of the corpus is analyzed.
- **Bugs**
  - The items are analyzed w.r.t. bugs that go beyond syntax and type errors.

#### Dynamic analysis
- **Actual program runs are analyzed.**
- **Profiles**
  - Execution frequencies of methods, for example, are analyzed.
- **Traces**
  - Execution traces of method calls, for example, are analyzed.

#### Dimensions of analysis
- **Orthogonal dimensions applicable to analyses.**
- **Evolution**
  - An analysis is carried out comparatively for multiple versions.
- **Clustering**
  - The corpus is clustered by metadata such as country, team size, or others.
## Analyses

### Static Dimensions
- **Validity**
- **Metrics**
- **Coverage**
- **Styles**
- **Patterns**
- **Cloning**
- **Bugs**

### Dynamic Dimensions
- **Profiles**
- **Traces**
- **Evolution**
- **Clustering**

<table>
<thead>
<tr>
<th>Publication</th>
<th>Validity</th>
<th>Metrics</th>
<th>Coverage</th>
<th>Styles</th>
<th>Patterns</th>
<th>Cloning</th>
<th>Bugs</th>
<th>Profiles</th>
<th>Traces</th>
<th>Evolution</th>
<th>Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlvesV05</td>
<td></td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BaxterFNSM06</td>
<td></td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChevanceH78</td>
<td></td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CollbergMS04</td>
<td></td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CookL82</td>
<td></td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CranorESMC08</td>
<td></td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GIM05</td>
<td></td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HageK08</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hahsler04</td>
<td></td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hauss02</td>
<td></td>
<td>•</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KimSNN05</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knuth71</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LaemmelKR05</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LaemmelP10</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReayDM09</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SaalW88</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visser06</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Evolution
- An analysis is carried out comparatively for multiple versions.

### Clustering
- The corpus is clustered by metadata such as country, team size, or others.

**Fig. 4.** Classification of analyses

**Fig. 5.** Analyses of the selected publications

---

| © 2005-2013 Ralf Lämmel, Software Languages Team, and collaborators |
Applied research method: Content analysis

- Formulate research questions
- Formulate inclusion criteria for papers
- Collect data
- Evaluate data
- Analyze data
- Publish analysis

In another lecture, we will look at systematic literature surveys in more detail.
Analysis of XML schema usage

Ralf Lämmel and Stan Kitsis and Dave Remy

Microsoft Corporation

Abstract

XML schema analysis aims to extract quantitative and qualitative information from actual XML schemas. To this end, XML schemas are measured through systematic algorithms, on the basis of the intrinsic feature model of the XSD language. XML schema analysis is a derivative of software analysis (program analysis) and of software code metrics, in particular. The present article introduces essential concepts of XML schema analysis and applies them to the important problem of understanding XML schema usage in practice. Analyses for feature counts, idiosyncrasy counts, size metrics, complexity metrics, and XML schema styles are executed on a large corpus of real-world XML schemas.

Published in XML’05
Questions about W3C XML Schema

Size
How “big” and “complex” are real-world XML schemas?

Features
What XSD features are used in practice and with what frequency?

Styles
Which styles of schema organization are used in practice?
Agenda

- XSD harness
- XSD metrics
- XSD features
- XSD styles
- XSD-isms
- XSD plans
XSD harness under study

- 63 schemas from various IT sectors
- HL7, IFX, MS Office, OAGIS, …
- Approx. 6000 individual schema files
- Approx. 100,000 combined global element declarations and global complex-type definitions
XSD metrics

- XML-agnostic size
- XSD-agnostic size
- XSD-aware counts
- McCabe for XSD
- Code-oriented breadth
- Instance-oriented breadth
- Code-oriented depth
- Instance-oriented depth
XML agnostic – The KB metrics

<table>
<thead>
<tr>
<th>Size of schema projects in KB (6 KB to 18 MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100,000</td>
</tr>
<tr>
<td>10,000</td>
</tr>
<tr>
<td>1,000</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Schema projects (sorted by KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 10 20 30 40 50 60</td>
</tr>
</tbody>
</table>
XML-agnostic – The LOC metrics

[Graph showing the relationship between the size of schema projects in LOC and the number of schema projects (again, sorted by KB).]
## LOC-based size categorization

<table>
<thead>
<tr>
<th>LOC-based category</th>
<th>Definition</th>
<th>Schema count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mini</td>
<td>0 – 100</td>
<td>0</td>
</tr>
<tr>
<td>Small</td>
<td>100 – 1,000</td>
<td>12</td>
</tr>
<tr>
<td>Medium</td>
<td>1,000 – 10,000</td>
<td>24</td>
</tr>
<tr>
<td>Large</td>
<td>10,000 – 100,000</td>
<td>23</td>
</tr>
<tr>
<td>Huge</td>
<td>100,000 – …</td>
<td>4</td>
</tr>
</tbody>
</table>
Obvious problems with LOC

<!-- counts as 1 LOC -->
<xs:complexType name="foo"/>

<!-- counts as 2 LOC -->
<xs:complexType name="foo">
  <xs:complexType>
  </xs:complexType>
</xs:complexType>

<!-- counts as 3 LOC -->
<xs:complexType name="foo">
  <xs:complexType>
    <xs:complexType>
    </xs:complexType>
  </xs:complexType>
</xs:complexType>
XSD-agnostic – The #NODE metrics

Size of schema projects in nodes (excluding some extra-large projects)

- Nodes
- Elements
- Attributes
- Annotations
- Documentation
- AppInfo
**XSD–aware counts**

- **#EL\(_g\)**: Number of global element declarations
- **#CT\(_g\)**: Number of global complex–type definitions
- **#ST\(_g\)**: Number of global simple–type definitions
- **#MG\(_g\)**: Number of global model–group definitions
- **#AG\(_g\)**: Number of global attribute–group declarations
- **#AT\(_g\)**: Number of global attribute declarations

<table>
<thead>
<tr>
<th>Measure</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>#EL(_g)</td>
<td>2</td>
<td>75</td>
<td>3549</td>
<td>31416</td>
</tr>
<tr>
<td>#CT(_g)</td>
<td>0</td>
<td>52</td>
<td>16775</td>
<td>50568</td>
</tr>
<tr>
<td>#ST(_g)</td>
<td>0</td>
<td>24</td>
<td>3044</td>
<td>10753</td>
</tr>
<tr>
<td>#MG(_g)</td>
<td>0</td>
<td>0</td>
<td>643</td>
<td>1615</td>
</tr>
<tr>
<td>#AG(_g)</td>
<td>0</td>
<td>0</td>
<td>2526</td>
<td>3294</td>
</tr>
<tr>
<td>#AT(_g)</td>
<td>0</td>
<td>0</td>
<td>65</td>
<td>292</td>
</tr>
</tbody>
</table>
What’s a good count-based measure?

Candidates:

- \(#\text{EL}_g + \#\text{CT}_g\)
- \(#\text{EL}_g + \#\text{EL}_l\) (#EL)
- \(#\text{CT}_g + \#\text{CT}_l\) (#CT)

The former is often used in informal communication. #CT captures the number of hierarchical types (“concepts”).
## CT-based size categorization

<table>
<thead>
<tr>
<th>CT-based category</th>
<th>Definition</th>
<th>Schema count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mini</td>
<td>0 – 32</td>
<td>13</td>
</tr>
<tr>
<td>Small</td>
<td>32 – 100</td>
<td>12</td>
</tr>
<tr>
<td>Medium</td>
<td>100 – 256</td>
<td>14</td>
</tr>
<tr>
<td>Large</td>
<td>256 – 1000</td>
<td>12</td>
</tr>
<tr>
<td>Huge</td>
<td>1000 – ...</td>
<td>12</td>
</tr>
</tbody>
</table>
McCabe complexity

Count binary decisions.

For XML schemas:
- Choices
- Occurrence constraints
- Element references to substitution groups
- Type references to subtyped types
- The multiplicity of root elements
- Nullable elements
Absolute McCabe complexity
Relative McCabe complexity

![Bar chart showing relative McCabe complexity](chart.png)
MCC-based size categorization based on (#NODE–#ANN)/MCC

<table>
<thead>
<tr>
<th>MCC Complexity Level</th>
<th>Definition</th>
<th>Schema count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trivial</td>
<td>20 – …</td>
<td>7</td>
</tr>
<tr>
<td>Simple</td>
<td>10 – 20</td>
<td>22</td>
</tr>
<tr>
<td>Difficult</td>
<td>4 – 10</td>
<td>34</td>
</tr>
<tr>
<td>Intractable</td>
<td>0 – 4</td>
<td>7</td>
</tr>
</tbody>
</table>
## Breadth and depth measures

<table>
<thead>
<tr>
<th></th>
<th>Code-oriented</th>
<th>Instance-oriented</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Breadth</strong></td>
<td>“Parties” in content models</td>
<td>Sibling children and attributes in XML instance tree</td>
</tr>
<tr>
<td><strong>Depth</strong></td>
<td>Descriptive depth of XML types</td>
<td>Normal notion of depth for XML instance tree</td>
</tr>
</tbody>
</table>

We are primarily interested in bounds per schema.
Code-oriented breadth

![Graph showing the relationship between schema projects and the number of parties not showing some outliers. The graph plots two sets: one for all parties and another for those without attributes.](image-url)
Instance-oriented breadth

![Graph showing the relationship between the number of children (not showing some outliers) and schema projects (sorted by maximum children including attributes). The graph includes two lines: one for "all children" and another for "w/o attributes." ](image-url)
Code-oriented depth

![Bar chart showing the number of schemas with the selected depth for elements only and full descriptional depth.](chart.png)
Instance-oriented depth

“Early ceasing” – Optional particles have depth 0.

“Late ceasing” – Attempt to count optional particles.
XSD features

- Model-group operators
- Simple-type features
- Occurrence features
- Subtyping and friends
- Mixed content
- Wildcards
- Identity constraints
- Modularization
## Model–group operators

<table>
<thead>
<tr>
<th>Operator</th>
<th>Schema count</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td>sequence</td>
<td>63</td>
<td>53903</td>
</tr>
<tr>
<td>choice</td>
<td>53</td>
<td>19950</td>
</tr>
<tr>
<td>all</td>
<td>6</td>
<td>40</td>
</tr>
</tbody>
</table>

© 2005-2013 Ralf Lämmel, Software Languages Team, and collaborators
## Simple-type features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Schema count</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td>restriction</td>
<td>59</td>
<td>10299</td>
</tr>
<tr>
<td>list</td>
<td>15</td>
<td>73</td>
</tr>
<tr>
<td>union</td>
<td>21</td>
<td>2038</td>
</tr>
<tr>
<td>pattern</td>
<td>32</td>
<td>925</td>
</tr>
<tr>
<td>enumeration (groups)</td>
<td>57</td>
<td>4663</td>
</tr>
<tr>
<td>enumeration (constants)</td>
<td>57</td>
<td>117923</td>
</tr>
</tbody>
</table>
### Occurrence features

<table>
<thead>
<tr>
<th>Operator</th>
<th>Schema count</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td>all &quot;occurs&quot;</td>
<td>63</td>
<td>275661</td>
</tr>
<tr>
<td>“non-DTD occurs”</td>
<td>32</td>
<td>2235</td>
</tr>
<tr>
<td>nillable=&quot;true&quot;</td>
<td>7</td>
<td>121832</td>
</tr>
<tr>
<td>fixed=&quot;...&quot;</td>
<td>22</td>
<td>33209</td>
</tr>
<tr>
<td>default=&quot;...&quot;</td>
<td>31</td>
<td>18240</td>
</tr>
<tr>
<td>use=&quot;required&quot;</td>
<td>49</td>
<td>7866</td>
</tr>
<tr>
<td>use=&quot;prohibited&quot;</td>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td>Specified default</td>
<td>47</td>
<td>66095</td>
</tr>
</tbody>
</table>
## Subtyping and friends

<table>
<thead>
<tr>
<th>Feature</th>
<th>Schema count</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract elements and types</td>
<td>21</td>
<td>298</td>
</tr>
<tr>
<td>Substitution groups</td>
<td>12</td>
<td>165</td>
</tr>
<tr>
<td>Simple-type restriction</td>
<td>59</td>
<td>10299</td>
</tr>
<tr>
<td>Complex-type extension; simple content</td>
<td>40</td>
<td>6026</td>
</tr>
<tr>
<td>Complex-type extension; complex content</td>
<td>36</td>
<td>5948</td>
</tr>
<tr>
<td>Complex-type restriction; simple content</td>
<td>5</td>
<td>41</td>
</tr>
<tr>
<td>Complex-type restriction; complex content</td>
<td>17</td>
<td>1020</td>
</tr>
<tr>
<td>Redefine</td>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>blockDefault/block</td>
<td>4</td>
<td>39</td>
</tr>
<tr>
<td>finalDefault/final</td>
<td>7</td>
<td>399</td>
</tr>
</tbody>
</table>
# Mixed content

<table>
<thead>
<tr>
<th>Feature</th>
<th>Schema count</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed</td>
<td>30</td>
<td>761</td>
</tr>
</tbody>
</table>
# Wildcards

<table>
<thead>
<tr>
<th>Feature</th>
<th>Schema count</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td>any</td>
<td>33</td>
<td>825</td>
</tr>
<tr>
<td>anyAttribute</td>
<td>13</td>
<td>371</td>
</tr>
</tbody>
</table>
# Identity constraints

<table>
<thead>
<tr>
<th>Feature</th>
<th>Schema count</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td>unique</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>key</td>
<td>5</td>
<td>41</td>
</tr>
<tr>
<td>keyref</td>
<td>2</td>
<td>333</td>
</tr>
</tbody>
</table>
### Modularization

<table>
<thead>
<tr>
<th>Feature</th>
<th>Schema count</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple files</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td><strong>No targetNamespace</strong></td>
<td>20</td>
<td><strong>101</strong></td>
</tr>
<tr>
<td>Chameleon include</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Imports</td>
<td>26</td>
<td><strong>1348</strong></td>
</tr>
<tr>
<td>Includes</td>
<td>30</td>
<td><strong>19106</strong></td>
</tr>
</tbody>
</table>
XSD styles

- Recall folklore styles
  - Venetian Blind
  - Salami Slice
  - Russian Doll
  - Garden of Eden

- Formalization of styles
- Styles measurements
- Open questions
<xsd:schema ...

  <xsd:element name="orders" type="ordersType"/>
  <xsd:complexType name="ordersType">
    <xsd:sequence>
      <xsd:element name="order" type="orderType"
        minOccurs="0" maxOccurs="unbounded"/>
    </xsd:sequence>
  </xsd:complexType>

  <xsd:complexType name="orderType">
    <xsd:sequence>
      <xsd:element name="orderid" type="xsd:string"/>
      <xsd:element name="item" type="itemType" maxOccurs="unbounded"/>
      <xsd:element name="address" type="addressType"/>
    </xsd:sequence>
  </xsd:complexType>

  <!-- Some types ELIDED ... -->

</xsd:schema>
<xsd:schema ...>
  <xsd:element name="orders">
    <xsd:complexType>
      <xsd:sequence>
        <xsd:element ref="order" minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="order">
    <xsd:complexType>
      <xsd:sequence>
        <xsd:element name="orderid" type="xsd:string"/>
        <xsd:element ref="item" maxOccurs="unbounded"/>
        <xsd:element ref="address"/>
      </xsd:sequence>
    </xsd:complexType>
  </xsd:element>
</xsd:schema>
<xsd:schema ...
  
  <xsd:element name="orders">
    <xsd:complexType>
      <xsd:sequence>
        <xsd:element name="order" minOccurs="0" maxOccurs="unbounded">
          <xsd:complexType>
            <xsd:sequence>
              <xsd:element name="orderid" type="xsd:string"/>
              <xsd:element name="item" maxOccurs="unbounded">
                <!-- Type ELIDED ... -->
              </xsd:element>
            </xsd:sequence>
          </xsd:complexType>
        </xsd:element>
        <xsd:element name="address">
          <!-- Type ELIDED ... -->
        </xsd:element>
        </xsd:sequence>
      </xsd:complexType>
    </xsd:element>
  </xsd:sequence>
</xsd:complexType>
</xsd:element>
</xsd:complexType>
</xsd:element>
</xsd:schema>
<xsd:schema ...>
  <xsd:element name="orders" type="ordersType"/>
  <xsd:complexType name="ordersType">
    <xsd:sequence>
      <xsd:element ref="order" minOccurs="0" maxOccurs="unbounded"/>
    </xsd:sequence>
  </xsd:complexType>
  <xsd:element name="order" type="orderType"/>
  <xsd:complexType name="orderType">
    <xsd:sequence>
      <xsd:element name="orderid" type="xsd:string"/>
      <xsd:element ref="item" maxOccurs="unbounded"/>
      <xsd:element ref="address"/>
    </xsd:sequence>
  </xsd:complexType>
</xsd:schema>

<!– Some types ELIDED ... -->
</xsd:schema>
# Formalization of styles

<table>
<thead>
<tr>
<th>Strict style</th>
<th>#EL&lt;sub&gt;g&lt;/sub&gt;</th>
<th>#CT&lt;sub&gt;g&lt;/sub&gt;</th>
<th>#EL&lt;sub&gt;l&lt;/sub&gt;</th>
<th>#CT&lt;sub&gt;l&lt;/sub&gt;</th>
<th>#EL&lt;sub&gt;r&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Russian Doll”</td>
<td>&gt; 0</td>
<td>= 0</td>
<td>&gt; 0</td>
<td>&gt; 0</td>
<td>= 0</td>
</tr>
<tr>
<td>“Salami Slice”</td>
<td>&gt; 0</td>
<td>= 0</td>
<td>= 0</td>
<td>&gt; 0</td>
<td>&gt; 0</td>
</tr>
<tr>
<td>“Venetian Blind”</td>
<td>&gt; 0</td>
<td>&gt; 0</td>
<td>&gt; 0</td>
<td>= 0</td>
<td>= 0</td>
</tr>
<tr>
<td>“Garden of Eden”</td>
<td>&gt; 0</td>
<td>&gt; 0</td>
<td>= 0</td>
<td>= 0</td>
<td>&gt; 0</td>
</tr>
</tbody>
</table>
Styles measurements

- **Schema files in total:** 5970
- **Files per strict style:**
  - “Russian Doll”: 85
  - “Salami Slice”: 263
  - “Venetian Blind”: 2728
  - “Garden of Eden”: 344
- **Strict styling:**
  - Files: 3420 (57.29%)
  - Affected schemas: 45 (71.43%)
  - Fully styled schemas: 4 (6.35%)
  - “No Trees”: 845
Problems with XSD styles

- There are more styles.
- XSD abstractions not mentioned by styles
- Modular schemas with different styles
- No expression of style intentions
- ...

© 2005-2013 Ralf Lämmel, Software Languages Team, and collaborators
How should we call this style?

```xml
<xsd:schema ...>
  <xsd:element name="orders">
    <xsd:complexType>
      <xsd:sequence>
        <xsd:element name="order" type="orderType"
                      minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
    </xsd:complexType>
  </xsd:element>
  <xsd:complexType name="orderType">
    <xsd:sequence>
      <xsd:element name="orderid" type="xsd:string"/>
      <xsd:element name="item" type="itemType" maxOccurs="unbounded"/>
      <xsd:element name="address" type="addressType"/>
    </xsd:sequence>
  </xsd:complexType>
</xsd:schema>
```

<!– Some types ELIDED ... -->

© 2005-2013 Ralf Lämmel, Software Languages Team, and collaborators
XSD-isms

- Anonymous compositors
- Ambiguous selectors
- Global symbol namespaces
- Local symbol namespaces
- Case sensitivity
<xs:element name="foo">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="bar" type="xs:int"/>
      <xs:choice>
        <xs:element name="xyz" type="xs:int"/>  
        <xs:element name="zyx" type="xs:int"/>
      </xs:choice>
    </xs:sequence>
  </xs:complexType>
</xs:element>
Strongly anonymous form

```xml
<xs:element name="foo">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="bar"/>
      <xs:sequence maxOccurs="unbounded">
        <xs:element name="xyz"/>
        <xs:element name="zyx"/>
      </xs:sequence>
    </xs:sequence>
  </xs:complexType>
</xs:element>
```
Ambiguous selectors

<!-- Three bars are all at the same level. -->
<xs:element name="foo">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="bar" type="xs:int"/>
      <xs:element name="bar" type="xs:int"/>
      <xs:element name="xyz" type="xs:int"/>
      <xs:element name="bar" type="xs:int"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
Global symbol namespaces

```xml
<?xml version="1.0" encoding="utf-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <!-- An element foo and a simple type foo -->
  <xs:element name="foo" type="foo"/>
  <xs:simpleType name="foo">
    <xs:restriction base="xs:int"/>
  </xs:simpleType>
</xs:schema>
```
Local symbol namespaces

<!-- An element bar and an attribute bar -->
<xs:element name="foo1">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="bar" type="xs:int" />
    </xs:sequence>
    <xs:attribute name="bar" type="xs:string" />
  </xs:complexType>
</xs:element>
# XSD-ism in numbers

<table>
<thead>
<tr>
<th>Friction issue (&quot;clashes&quot;)</th>
<th>Affected schemas</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anonymous compositors</td>
<td>41</td>
<td>6737</td>
</tr>
<tr>
<td>Strong anonymity</td>
<td>28</td>
<td>288</td>
</tr>
<tr>
<td>Ambiguous selectors <em>(case sensitive)</em></td>
<td>27</td>
<td>3591</td>
</tr>
<tr>
<td>Ambiguous selectors <em>(case insensitive)</em></td>
<td>27</td>
<td>3591</td>
</tr>
<tr>
<td>Clashing globals <em>(case sensitive)</em></td>
<td>17</td>
<td>634</td>
</tr>
<tr>
<td>Clashing globals <em>(case insensitive)</em></td>
<td>27</td>
<td>2179</td>
</tr>
<tr>
<td>Element vs. attributes <em>(case sensitive)</em></td>
<td>30</td>
<td>33</td>
</tr>
<tr>
<td>Element vs. attributes <em>(case insensitive)</em></td>
<td>33</td>
<td>158</td>
</tr>
</tbody>
</table>
Future work

- Applications to XML data binding
- Applications to schema design
- Research on instance-based analyses
- More research on styles
Understanding privacy policies
A study in empirical analysis of language usage

Ralf Lämmel · Ekaterina Pek

Abstract Their is growing recognition that users of web-based systems want to understand, if not control, what customer’s data is stored by whom, for what purpose, for what duration, and with whom it is shared. We inform current language-based privacy efforts with an empirical study of P3P—the W3C domain-specific language for privacy policies. We use methods of software language engineering to study usage profiles, correctness of policies, metrics, cloning, and language extensions. The study supports the conclusion that P3P’s approach to policy validation is too weak to ensure correct use of the language. The study also discovers common, dominating policies, which may suggest a simpler approach to web privacy. Further, the study investigates a range of metrics for policies in an attempt to discover particularly interesting or complex policies. Finally, the study also attempts to discover symptoms of the need for extending the P3P language, but the found results are not conclusive here.

Published in ESE Journal, 2013
TOC

• What’s web privacy?
• The language-based approach of P3P
• Research challenges
• An empirical study on P3P usage
What’s web privacy?
[COMPANY NAME] is the sole owner [or describe other entities, e.g., parent, affiliates, who have an ownership interest] of the information collected on [NAME OF SITE]. [COMPANY NAME] collects personally identifiable information from our users at several different points on our Web site.

**Orders**
If you purchase a product or service from us, we request certain personally identifiable information from you on our order form. You must provide contact information (such as name, email, and shipping address) and financial information (such as credit card number, expiration date).

We use this information for billing purposes and to fill your orders. If we have trouble processing an order, we will use this information to contact you.

Source: [www.truste.org/docs/Model_Privacy_Policy_Disclosures.doc](http://www.truste.org/docs/Model_Privacy_Policy_Disclosures.doc)
The TRUSTe seal for web privacy
Social Security Numbers

We require your social security number to provide [describe the service you provide here.] When you enter your social security number on our registration [order form, application form], we encrypt it using secure socket layer technology (SSL).

TRUSTe (an organization that grants seals for web privacy) recommends that web sites collect a social security number ONLY when it is required for performing a service the site provides.

Source: www.truste.org/docs/Model_Privacy_Policy_Disclosures.doc
Some “textual” policies

http://www.facebook.com/privacy
http://www.facebook.com/about/privacy/
http://www.google.com/policies/privacy/
http://www.microsoft.com/privacy
https://twitter.com/privacy
...

URL rewritten to my settings
Effective per 1 March
General info, not a policy
The language-based approach of P3P
P3P 1.0: A New Standard in Online Privacy

"The World Wide Web Consortium, the group that designs standards for the Web, is creating a new way [P3P] for Web sites to transmit the site's privacy policy automatically, and allow users to signal only the information they are willing to share."


http://www.w3.org/P3P/brochure.html

last revised $Date: 2006/05/12 15:20:18$ by $Author: rigo$
A quick intro to P3P
The “full privacy” policy

<POLICY>
  <STATEMENT>
    <NON-IDENTIFIABLE/>
  </STATEMENT>
</POLICY>
Full privacy with Tor?

**Anonymity Online**

Protect your privacy. Defend yourself against network surveillance and traffic analysis.

- Tor prevents anyone from learning your location or browsing habits.
- Tor is for web browsers, instant messaging clients, remote logins, and more.
- Tor is free and open source for Windows, Mac, Linux/Unix, and Android

**What is Tor?**

Tor is free software and an open network that helps you defend against a form of network surveillance that threatens personal freedom and privacy, confidential business activities and relationships, and state security known as traffic analysis.

Learn more about Tor »

**Why Anonymity Matters**

Tor protects you by bouncing your communications around a distributed network of relays run by volunteers all around the world: it prevents somebody watching your Internet connection from learning what sites you visit, and it prevents the sites you visit from learning your physical location. Tor works with many of your existing applications, including web browsers, instant messaging clients, remote login, and other applications based on the TCP protocol.

© 2005-2013 Ralf Lämmel, Software Languages Team, and collaborators
The “logging only” policy

```xml
<POLICY>
  <STATEMENT>
    <PURPOSE><admin/><current/><develop/></PURPOSE>
    <RECIPIENT><ours/></RECIPIENT>
    <RETENTION><indefinitely/></RETENTION>
    <DATA-GROUP>
      <DATA ref="#dynamic.clickstream"/>
      <DATA ref="#dynamic.http"/>
    </DATA-GROUP>
  </STATEMENT>
</POLICY>
```
A variation on “logging only” with less exposure

```xml
<POLICY>
 <STATEMENT>
   <PURPOSE><admin/><current/><develop/></PURPOSE>
   <RECIPIENT><ours/></RECIPIENT>
   <RETENTION><legal-requirement/></RETENTION>
   <DATA-GROUP>
     <DATA_ref="#dynamic.clickstream"/>
   </DATA-GROUP>
 </STATEMENT>
</POLICY>
```
Use of “wildcards” and categories

```xml
<Data ref="#dynamic.miscdata">
  <Categories>
    <political/>
  </Categories>
</Data>
```
The study's abstract syntax of P3P

The P3P specification does not commit to a specific language for user preferences, but W3C's APPEL is explicitly mentioned as an option; other options have been proposed. Fundamentally, languages for policies versus preferences differ as follows: A policy language is assumed to make statements about data being optionally or mandatorily required by the website for certain purposes, to be stored for a certain duration, and to be shared with certain recipients. In contrast, preferences can be thought of as constraints or rules to be applied to policies which are hence to be accepted or rejected.

Our paper focuses entirely on P3P as a policy language.

3- Syntax of P3P

In our study, we are only concerned with the formal statements about data collection, use, storage, and sharing. We are usually not concerned with the entity that issues the policy or a policy's consequences which are given in natural language. The remaining P3P syntax is described in Fig. 5, which can be seen as an abstract syntax derived from comprehensive concrete syntax definitions. It is important to note that this abstract syntax has no counterpart in the P3P specification. Instead, we have designed this abstract syntax to precisely include those language elements that are of interest for our purposes.

According to the figure, a P3P policy consists of any number of statements, each of which signifies data references for the collected data, purposes of collecting data, recipients that receive the data, a retention level which defines how long data is stored, and a few other directives. The figure clarifies that the major syntactical domains are finite enumeration types and there is no recursion involved. We should add that P3P is somewhat prepared to deal with unforeseen concepts. There is, for example, a purpose called 'other purpose', and there is a general extension mechanism that we discuss in §;

Data references either refer to the Base Data Schema of P3P or a Custom Data Schema that is provided with the policy. The BDS, which is part of the P3P specification, (organizes user data, business data, third-party data, and so-called dynamic data with diverse roles in a hierarchical manner; see Fig. 5 for an illustration. Data references can be annotated with categories such as 'purchase' or 'health'. In fact, most data references in the BDS are implicitly associated with one or more categories. Comprehensive concrete syntax definitions of P3P are available in different forms:

- XSD: http://www.w3.org/2002/01/P3Pv1.xsd
- Relax NG: http://yupotan.sppd.ne.jp/relax-ng/p3pv1.rng
- RDF: http://www.w3.org/TR/p3p-rdfschema/

BDS: http://www.w3.org/TR/P3P11/-base_data_structure
http://www.w3.org/TR/P3P11/-schema_detail

© 2005-2013 Ralf Lämmel, Software Languages Team, and collaborators
In need of a normal form

• Styles:
  ‣ Use **one** statement with **multiple** data refs
  ‣ Use **multiple** statements with **one** data ref
  ‣ ...

• Normal form

  \[
  \begin{align*}
  & \text{r-purpose}(\text{data}, \text{purpose}, \text{required}) \\
  & \text{r-recipient}(\text{data}, \text{recipient}, \text{required}) \\
  & \text{r-retention}(\text{data}, \text{retention}) \\
  & \text{r-data}(\text{data}, \text{optional, identifiable}) \\
  & \text{r-category}(\text{data}, \text{category})
  \end{align*}
  \]
The P3P process

Participants, Supporters, Developers

The following companies and organizations have been active participants in developing P3P.

- America Online
- AT&T
- Center for Democracy & Technology
- Citigroup
- Crystaliz
- Direct Marketing Association
- Electronic Network Consortium
- GeoTrust
- Gesellschaft für Mathematik und Datenverarbeitung (GMD)
- Hewlett Packard
- IBM
- IDcide
- International Security, Trust, and Privacy Alliance
- Internet Alliance
- Jotter Technologies Inc.
- Microsoft
- NCR
- NEC
- Netscape
- Nokia
- Ontario Office of the Information and Privacy Commissioner
- Phone.com, Inc.
- Privacy Commission of Schleswig-Holstein, Germany
- TRUSTe

http://www.w3.org/P3P/brochure.html
last revised $Date: 2006/05/12 15:20:18$ by $Author: rigo$
Status of P3P
Status: P3P Work suspended

After a successful Last Call, the P3P Working Group decided to publish the P3P 1.1 Specification as a Working Group Note to give P3P 1.1 a provisionally final state.

The P3P Specification Working Group took this step as there was insufficient support from current Browser implementers for the implementation of P3P 1.1. The P3P 1.1 Working Group Note contains all changes from the P3P 1.1 Last Call. The Group thinks that P3P 1.1 is now ready for implementation. It is not excluded that W3C will push P3P 1.1 until Recommendation if there is sufficient support for implementation.

On the other hand, P3P keeps being the basis of a number of research directions in the area of privacy world wide. One might cite the PRIME Project as well as the Policy aware Web. Many other approaches also follow the descriptive metadata approach started by P3P. Such projects are invited to send email to <rigo@w3.org> to be listed here.

http://www.w3.org/P3P/
Arguably, the largest barrier to P3P adoption has not been problems with the P3P vocabulary or difficulties with the technical mechanisms, but rather lack of incentives to adopt. As Dyson observed in 1997, “Industry disclosure schemes often founder without strong government/public pressure. Otherwise, companies are simply too busy to adopt them, and customers don’t factor the information disclosed into their buying habits.” By the time the P3P specification was released in 2002, government pressure had subsided and industry had largely lost interest in P3P.

http://www.ftc.gov/os/comments/privacyreportframework/00453-58003.pdf

[Emphasis added by this speaker]
Looking Back at P3P: Lessons for the Future
by Ari Schwartz
November 11, 2009

• Building and Over-Building

• Caught Up in the Politics of Privacy: Criticized by some ...
  • ... privacy advocates as an industry subterfuge
  • ... industry advocates as a substitute for legislation
  • ... in industry for providing consumers too much transparency

• Web sites build to the implementation, not the specification.

Research challenges
Research challenges

• Integration of privacy and access control
• Understandability of privacy policies
• Effective control of privacy
• Privacy for users or for legislators
• Privacy as specified or as implemented
• Learn from the past -- scientifically
A “Nutrition Label” for Privacy

Patrick Gage Kelley,* Joanna Bresee,* Lorrie Faith Cranor,† Robert W. Reeder**

*Carnegie Mellon University  †Microsoft

Symposium On Usable Privacy and Security (SOUPS) 2009, July 15-17, 2009, Mountain View, CA, USA.
© 2005-2013 Ralf Lämmel, Software Languages Team, and collaborators
A “Nutrition Label” for Privacy

Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, Robert W. Reeder

* Carnegie Mellon University

© 2005-2013 Ralf Lämmel, Software Languages Team, and collaborators

** Microsoft

Symposium On Usable Privacy and Security (SOUPS) 2009, July 15-17, 2009, Mountain View, CA, USA.
Some related work on “enforcement”


Method and system for implementing privacy policy enforcement with a privacy ... Paul Anthony Ashley et al

A method is presented for enforcing a privacy policy concerning management of personally identifiable information in a centralized manner through a privacy proxy agent. A proxy intercepts a message from a first system to a second system, e.g., from a server to a client, and determines whether the message is associated with an operation on personally identifiable information; if not, then the proxy sends the message to the second system, but if so, then the proxy determines whether the operation on the personally identifiable information is compliant with a privacy policy and with user preference information with respect to the privacy policy for a user who is associated the personally identifiable information. If the message is compliant with the privacy policy and user preference data, then the proxy sends the first message to the second system; otherwise, an error indication is returned to the first system.

Inventors: Paul Anthony Ashley, Sridhar R. Muppidi, Mark Vandenauwer
Original Assignee: International Business Machines Corporation
Primary Examiner: Gilberto Barron, Jr.
Secondary Examiner: Venkat Perungavoor
Attorneys: Jeffrey S. LaBaw, David H. Judson
Current U.S. Classification: 726/1; 726/30

Source: http://www.google.com/patents/US7797726
An empirical study on P3P usage
Research questions

• What part of the vocabulary is used?
• Is the language correctly used?
• What is a significant policy?
• What are common policies?
• What language extensions circulate?
Our empirical work in comparison

http://softlang.uni-koblenz.de/p3p/

<table>
<thead>
<tr>
<th>Property</th>
<th>[18]</th>
<th>[19]</th>
<th>[20]</th>
<th>[21]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of corpus</td>
<td>3846?</td>
<td>1482?</td>
<td>3846?</td>
<td>3282</td>
</tr>
<tr>
<td>Number of sources</td>
<td>11</td>
<td>6</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Syntax error rates</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Boundary breakdown</td>
<td>–</td>
<td>●</td>
<td>–</td>
<td>●</td>
</tr>
<tr>
<td>Website evolution</td>
<td>–</td>
<td>●</td>
<td>●</td>
<td>–</td>
</tr>
<tr>
<td>P3P language coverage</td>
<td>–</td>
<td>–</td>
<td>○</td>
<td>–</td>
</tr>
<tr>
<td>Privacy classification</td>
<td>●</td>
<td>–</td>
<td>○</td>
<td>–</td>
</tr>
<tr>
<td>Legislation</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>●</td>
</tr>
<tr>
<td>HRPP interpretation</td>
<td>–</td>
<td>–</td>
<td>●</td>
<td>–</td>
</tr>
</tbody>
</table>
A P3P corpus
http://www.dmoz.org/
Results of URL scraping

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># Website URLs</td>
<td>4,009,337</td>
</tr>
<tr>
<td># Distinct website URLs</td>
<td>3,860,709</td>
</tr>
<tr>
<td># Distinct domains</td>
<td>2,957,657</td>
</tr>
</tbody>
</table>
Results of policy location and downloading

<table>
<thead>
<tr>
<th></th>
<th># Files</th>
<th># Distinct domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Located policy reference files</td>
<td>50,776</td>
<td>43,195</td>
</tr>
<tr>
<td>Referenced policy files</td>
<td>13,536</td>
<td>8,736</td>
</tr>
<tr>
<td>Downloaded policies</td>
<td>8,768</td>
<td>7,746</td>
</tr>
</tbody>
</table>

We refer to Appendix A where we provide some data on the disappearance of policies over time in terms of online availability for the policies in the corpus of the paper. All policies are permanently available from an open-source repository that goes with the paper.

2.2.3 Schema-based validation

The extracted policies are still subject to validation. Each policy file is supposed to contain one or more policies. However, not all policy files contain XML; files may happen to be empty or contain non-XML text. Further, not all policy files with XML content are well-formed XML, and not all well-formed XML content from policy files are valid with regard to P3P's XML schema. Counts of files on varying levels of validity are shown in Table 3. Again, numbers of files are annotated with the number of distinct domains [this can be viewed as a simple indicator of diversity].

We define our P3P corpus to consist only of those policies that are valid with regard to P3P's XML schema. For completeness’ sake, we mention that some P3P tools tolerate ill-formed or schema-invalid XML to some extent. Such tolerance is tool-specific though, and we do not try to use obviously incorrect policies in this study.

For convenience’s sake, Appendix A provides additional information about the diversity of the corpus in terms of top-level domains and ODP’s website categories. These considerations are inspired by studies of P3P adoption both geographically and in terms of website categories. Adoption and geographic or categorical diversity are not directly of interest in our study, but we do provide such extra diversity data so that others can assess the generality of our results.
Schema-based validation

Table 2
Results of policy location and downloading

<table>
<thead>
<tr>
<th>Criterion</th>
<th># Files</th>
<th># Policies</th>
<th># Distinct domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>All downloads</td>
<td>8,768</td>
<td>—</td>
<td>7,746</td>
</tr>
<tr>
<td>XML</td>
<td>7,899</td>
<td>—</td>
<td>7,554</td>
</tr>
<tr>
<td>Well-formed XML</td>
<td>7,675</td>
<td>—</td>
<td>7,371</td>
</tr>
<tr>
<td>Schema-validated XML</td>
<td>5,905</td>
<td>6,182</td>
<td>5,673</td>
</tr>
</tbody>
</table>

For convenience’s sake, Appendix A provides additional information about the diversity of the corpus in terms of top-level domains and ODP’s website categories. These considerations are inspired by studies of P3P adoption both geographically and in terms of website categories [8–38(397398]. Adoption and geographic or categorial diversity are not directly of interest in our study, but we do provide such extra diversity data so that others can assess the generality of our results.
Leveraged analyses

- Analysis of vocabulary
- Analysis of constraints
- Analysis of metrics
- Analysis of cloning
- Analysis of extensions
Leveraged analyses

• Analysis of vocabulary
• Analysis of constraints
• Analysis of metrics
• Analysis of cloning
• Analysis of extensions
Frequency of use for P3P’s syntactical domains

<table>
<thead>
<tr>
<th>Domain</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>DataRef</td>
<td>40979</td>
</tr>
<tr>
<td>Purpose</td>
<td>31254</td>
</tr>
<tr>
<td>Category</td>
<td>20896</td>
</tr>
<tr>
<td>Recipient</td>
<td>12496</td>
</tr>
<tr>
<td>Optional</td>
<td>10823</td>
</tr>
<tr>
<td>Retention</td>
<td>10623</td>
</tr>
<tr>
<td>Statement</td>
<td>10623</td>
</tr>
<tr>
<td>Required (purpose)</td>
<td>4171</td>
</tr>
<tr>
<td>Identifiable</td>
<td>1718</td>
</tr>
<tr>
<td>Required (recipient)</td>
<td>1155</td>
</tr>
</tbody>
</table>

Fig. 13

This definition omits tedious issues of folding and subsumption. We also omit the formal detail that Fig makes additional values equal.

4 Analyses of the study

In this section, we address the research questions of the study by analyses of vocabulary, language constraints, metrics, clones, and language extensions. Each subsection is dedicated to one of these analyses and its underlying research question. Analyses may break down into subanalyses. All subanalyses are subject to the following scheme: we motivate the analysis and describe technicalities, we list output data, and we discuss the results while referring to the research question again.

Analysis of vocabulary

The underlying research question is “What part of the vocabulary is used?” The main goal is to help with profiling P3P such that the core language is identified and simplifications or designated support for the profile may be possibly inferred.
Recipients

ours

delivery

© 2005-2013 Ralf Lämmel, Software Languages Team, and collaborators
Purposes

individual-decision
individual-analysis
pseudo-analysis
tailoring
telemarketing
contact

historical
other-purpose
pseudo-decision

admin
current
develop
Retention

stated-purpose

indefinitely

business-practices

legal-requirement

no-retention
Explicit categories

- computer
- demographic
- purchase
- content
- physical
- state
- navigation
- location
- financial
- interactive
- other-category
- uniqueid
- preference
Implicit categories

physical navigation

interactive computer online demographic
## Data refs

<table>
<thead>
<tr>
<th>Data Refs</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>dynamic.cookies</td>
<td>3351</td>
</tr>
<tr>
<td>dynamic.clickstream</td>
<td>3204</td>
</tr>
<tr>
<td>dynamic.http</td>
<td>3058</td>
</tr>
<tr>
<td>dynamic.miscdatalast</td>
<td>2753</td>
</tr>
<tr>
<td>dynamic.searchtext</td>
<td>1816</td>
</tr>
<tr>
<td>user.name</td>
<td>995</td>
</tr>
<tr>
<td>dynamic.clientevents</td>
<td>726</td>
</tr>
<tr>
<td>dynamic.interactionrecord</td>
<td>579</td>
</tr>
<tr>
<td>user.home-info</td>
<td>510</td>
</tr>
<tr>
<td>user.home-info.online.email</td>
<td>476</td>
</tr>
<tr>
<td>user.name.given</td>
<td>418</td>
</tr>
<tr>
<td>user.name.family</td>
<td>418</td>
</tr>
<tr>
<td>user.business-info.online.email</td>
<td>387</td>
</tr>
<tr>
<td>user.business-info</td>
<td>364</td>
</tr>
<tr>
<td>thirdparty.name</td>
<td>361</td>
</tr>
<tr>
<td>dynamic.http.useragent</td>
<td>333</td>
</tr>
<tr>
<td>user.home-info.telecom.telephone</td>
<td>280</td>
</tr>
</tbody>
</table>
Data refs

dynamic.clickstream
dynamic.searchtext
dynamic.miscedata
dynamic.interactionrecord
dynamic.clientevents
user.name
dynamic.cookies
dynamic.http
Second, the latest and provisional specification stipulates sanity checking rules for data references. We also obtained constraints from Yu et al.'s article on the relational semantics of P'P. This article covers key constraints as well as coherence constraints.

4.2.1 Key constraints

The key constraints of the relational schema for P'P's normal form, as of §', immediately constrain the abstract syntax of §'. We check these constraints naturally as we normalize policies by deriving relations from the P'P statements. Whenever we insert tuples into the relations, we admit identical tuples, but we do not admit tuples that violate key constraints. This is part of the normalization algorithm.

We refer to Table 6 for violation counts for the different key constraints. We note again that the key constraints are potentially debatable; see the discussion in §'.
Coverage of BDS with a threshold on the number of syntactically distinct policies

<table>
<thead>
<tr>
<th>Threshold</th>
<th>BDS coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77.16 %</td>
</tr>
<tr>
<td>2</td>
<td>63.88 %</td>
</tr>
<tr>
<td>3</td>
<td>59.56 %</td>
</tr>
<tr>
<td>4</td>
<td>50.30 %</td>
</tr>
<tr>
<td>5</td>
<td>44.44 %</td>
</tr>
<tr>
<td>6</td>
<td>38.88 %</td>
</tr>
<tr>
<td>7</td>
<td>35.80 %</td>
</tr>
<tr>
<td>8</td>
<td>31.79 %</td>
</tr>
<tr>
<td>9</td>
<td>30.24 %</td>
</tr>
<tr>
<td>10</td>
<td>29.93 %</td>
</tr>
</tbody>
</table>

© 2005-2013 Ralf Lämmel, Software Languages Team, and collaborators
Leveraged analyses

• Analysis of vocabulary
• Analysis of constraints
• Analysis of metrics
• Analysis of cloning
• Analysis of extensions
Key-constraint violations

<table>
<thead>
<tr>
<th>Relation</th>
<th># Policies</th>
<th># Syn. distinct policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>r-data</td>
<td>448</td>
<td>248</td>
</tr>
<tr>
<td>r-retention</td>
<td>139</td>
<td>104</td>
</tr>
<tr>
<td>r-purpose</td>
<td>50</td>
<td>33</td>
</tr>
<tr>
<td>r-recipient</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Any of them</td>
<td>544 (of 6,182)</td>
<td>312 (of 2,304)</td>
</tr>
</tbody>
</table>

r-purpose(data,purpose,required)
r-recipient(data,recipient,required)
r-retention(data,retention)
r-data(data,optional,identifiable)
r-category(data,category)
Data-schema constraint violations

### Table: Data-schema constraint violations

<table>
<thead>
<tr>
<th>Issue</th>
<th># Policies</th>
<th># Syn. distinct policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDS legacy</td>
<td>580</td>
<td>358</td>
</tr>
<tr>
<td>Unresolvable data reference</td>
<td>115</td>
<td>85</td>
</tr>
<tr>
<td>Missing CDS</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>CDS legacy</td>
<td>203</td>
<td>157</td>
</tr>
<tr>
<td>Any of them</td>
<td><strong>848</strong> (of 6,182)</td>
<td><strong>579</strong> (of 2,304)</td>
</tr>
</tbody>
</table>

The 'BDS legacy' issue refers to the problem that policies use the obsolete BDS of P3P where they are supposed to use BDS of P3P instead. “User agents are only required to validate P3P policy data elements according to a P3P data schema.”

The ‘Unresolvable data reference’ issue refers to the problem that policies refer to data references that are simply not declared by the BDS. The ‘Missing CDS’ issue refers to the problem that a linked CDS is non-resolvable.

The ‘CDS legacy’ issue refers to the problem that policies use an obsolete format for CDS where they are supposed to XSD. “Web sites using custom data schemas MUST publish these schemas in P3P format only.”

Discussion

Again, the practice of P3P validation is clearly shown to be insufficient. In particular, the attempt to reduce P3P validation to straightforward XML schema-based validation is shown to be ineffective. P3P schemas define a hierarchical namespace for data items and XML schema-based validation cannot be directly used to fully enforce correct use of data references. The policies in the corpus do not use XML Schema (XSD) for their Custom Data Schemas. One reason for the use of this outdated approach may be that many policies were already on the web, when the revised specification started to require XSD adoption may have ceased.

### 4.2.4 Coherence constraints

It is relatively easy to see that there exist certain coherence constraints for combining purposes, recipients, data references and others in a given policy. For instance, a policy would give a vacuous sense of privacy if it combined recipient ‘public’ with ‘non-definite retention’. ‘Once something has been fully published, one cannot effectively cease retention.’

We extracted coherence constraints from the P3P specification.
Coherence constraints

For example:

‘public recipient ⇒ indefinite retention’

<table>
<thead>
<tr>
<th>Source Grade</th>
<th>Constraint</th>
<th># Policies</th>
<th># Syn. distinct policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>‘ours ...’</td>
<td>77</td>
<td>45</td>
</tr>
<tr>
<td>30</td>
<td>‘public ...’</td>
<td>40</td>
<td>24</td>
</tr>
<tr>
<td>30</td>
<td>‘historical ...’</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>‘contact’</td>
<td>71</td>
<td>47</td>
</tr>
<tr>
<td>1</td>
<td>‘telemarketing’</td>
<td>353</td>
<td>107</td>
</tr>
<tr>
<td>1</td>
<td>‘individual-analysis’</td>
<td>451</td>
<td>94</td>
</tr>
<tr>
<td>1</td>
<td>‘individual-decision’</td>
<td>432</td>
<td>82</td>
</tr>
</tbody>
</table>

Any of them: 898 (of 6,182) 264 (of 2,304)
Summary of violations

<table>
<thead>
<tr>
<th>Constraint category</th>
<th># Policies</th>
<th># Syn. distinct policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key</td>
<td>544</td>
<td>312</td>
</tr>
<tr>
<td>Hierarchy</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>Data-schema</td>
<td>848</td>
<td>579</td>
</tr>
<tr>
<td>Coherence</td>
<td>898</td>
<td>264</td>
</tr>
<tr>
<td>Any of them</td>
<td><strong>2,193</strong> (of 6,182)</td>
<td><strong>1,083</strong> (of 2,304)</td>
</tr>
</tbody>
</table>

Important disclaimer: some details such as the choice of key constraints are subjective.
Leveraged analyses

• Analysis of vocabulary
• Analysis of constraints
• **Analysis of metrics**
• Analysis of cloning
• Analysis of extensions
Metrics

- Syntactical size
- Semantical size
- Vocabulary size
- Data size
Varying semantical size for given syntactical size; max ratio $\approx 22$. 
Here, — and — are the numbers of all purposes, recipients, and categories in the vocabulary of POLICY. Hence, we measure the degree of coverage of the relevant part of the vocabulary and we give equal weight to purposes, recipients, and categories.

Policies with only the NON-IDENTIFIABLE element may have a value of , for VOCA(·). Otherwise, the minimum value is approximately — in the case of one purpose, one recipient, and one category.

For example, all three different representations of the sample policy 'logging only' [Fig. — on p. ] and Fig. [ on p. ] have the same value for VOCA(·), namely, it is —: all policies have three distinct purposes and one distinct recipient.

Table 11 shows that the — scale of the VOCA(·) metric is fully exercised by the corpus. When compared to the SYN(·) and SEM(·) metrics, the distribution of VOCA(·) is considerably less shifted to the lower end; compare the ratios of maximum to 3rd quartile for the metrics. Hence, even syntactically or semantically smaller policies exercise the vocabulary substantially.
Determine significant policies

• Short-list top-10 for all 4 metrics.
• Consider policies in more than 1 short list.
• De-prioritize syntactical size.
• Also, include top position on vocabulary size.
The sites with the most significant policies

- http://www.astrodata.ch
- http://www.50cent.com
- http://kerntrophies.com/

astrology
rapping
trophies
The most significant policies

• http://www.astrodata.ch/w3c/policy-general.xml
• http://privacy-policy.umusic.com/w3c/p3p.xml
• http://kerntrophies.com/w3c/p3p.xml

“404” per 28 Feb 2012
Leveraged analyses

- Analysis of vocabulary
- Analysis of constraints
- Analysis of metrics
- Analysis of cloning
- Analysis of extensions
Clone detection

- No parametrized clones
- No fragment clones
- Classification of clones
  - **Textual** clones
  - **Syntactical** clones
  - **Semantical** clones
Table 14

<table>
<thead>
<tr>
<th>Type</th>
<th># Clone groups</th>
<th># Added clone groups</th>
<th># Added cloned policies</th>
<th># Enlarged clone groups</th>
<th># Merged clone groups</th>
<th>% Clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>txt</td>
<td>296</td>
<td>296</td>
<td>1,941</td>
<td>—</td>
<td>—</td>
<td>31.40</td>
</tr>
<tr>
<td>syn</td>
<td>496</td>
<td>280</td>
<td>1,857</td>
<td>46</td>
<td>114</td>
<td>30.04</td>
</tr>
<tr>
<td>sem</td>
<td>351</td>
<td>20</td>
<td>69</td>
<td>0</td>
<td>32</td>
<td>1.12</td>
</tr>
<tr>
<td>Total</td>
<td>—</td>
<td>—</td>
<td>3,867</td>
<td>—</td>
<td>—</td>
<td>62.55</td>
</tr>
</tbody>
</table>

We are left with 1,385 semantically distinct policies of 4,869 semantically valid policies of a total of 6,182 policies in the corpus.
Top-10 textual clone groups

<table>
<thead>
<tr>
<th>Entity</th>
<th>Card.</th>
<th>Avg. dist.</th>
<th>SYN(·)</th>
<th>SEM(·)</th>
<th>VOCA(·)</th>
<th>DATA(·)</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Hensley</td>
<td>175</td>
<td>1.000</td>
<td>42</td>
<td>156</td>
<td>0.650</td>
<td>(9; 3; 12)</td>
</tr>
<tr>
<td>CybrHost</td>
<td>139</td>
<td>1.000</td>
<td>2</td>
<td>0</td>
<td>0.000</td>
<td>(0; 0; 0)</td>
</tr>
<tr>
<td>PhotoBiz</td>
<td>137</td>
<td>1.000</td>
<td>2</td>
<td>0</td>
<td>0.000</td>
<td>(0; 0; 0)</td>
</tr>
<tr>
<td>Boatventures</td>
<td>96</td>
<td>1.000</td>
<td>15</td>
<td>16</td>
<td>0.163</td>
<td>(2; 1; 3)</td>
</tr>
<tr>
<td>Real Estate</td>
<td>69</td>
<td>1.000</td>
<td>42</td>
<td>–</td>
<td>0.363</td>
<td>(7; 1; 6)</td>
</tr>
<tr>
<td>Hilton Hotels</td>
<td>68</td>
<td>0.779</td>
<td>9</td>
<td>6</td>
<td>0.123</td>
<td>(2; 1; 1)</td>
</tr>
<tr>
<td>NASA</td>
<td>54</td>
<td>0.000</td>
<td>13</td>
<td>29</td>
<td>0.180</td>
<td>(4; 1; 1)</td>
</tr>
<tr>
<td>Rezidor SAS</td>
<td>51</td>
<td>0.000</td>
<td>12</td>
<td>16</td>
<td>0.163</td>
<td>(2; 1; 3)</td>
</tr>
<tr>
<td>Bravenet</td>
<td>46</td>
<td>0.000</td>
<td>17</td>
<td>16</td>
<td>0.123</td>
<td>(2; 1; 1)</td>
</tr>
<tr>
<td>Wetpaint</td>
<td>38</td>
<td>0.237</td>
<td>19</td>
<td>24</td>
<td>0.180</td>
<td>(4; 1; 1)</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>–</td>
<td>0.602</td>
<td>17.30</td>
<td>26.30</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Top-10 syntactical clone groups

<table>
<thead>
<tr>
<th>Sample entity</th>
<th>Card.</th>
<th>Avg. dist.</th>
<th>SYN(·)</th>
<th>SEM(·)</th>
<th>VOCA(·)</th>
<th>DATA(·)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accountancy</td>
<td>331</td>
<td>0.000</td>
<td>36</td>
<td>44</td>
<td>0.367 (8; 2; 2)</td>
<td>33</td>
</tr>
<tr>
<td>CybrHost</td>
<td>284</td>
<td>0.109</td>
<td>2</td>
<td>0</td>
<td>0.000 (0; 0; 0)</td>
<td>0</td>
</tr>
<tr>
<td>IBM tracking</td>
<td>218</td>
<td>0.124</td>
<td>19</td>
<td>24</td>
<td>0.180 (4; 1; 1)</td>
<td>30</td>
</tr>
<tr>
<td>Johnston Press</td>
<td>116</td>
<td>0.586</td>
<td>20</td>
<td>24</td>
<td>0.220 (4; 1; 3)</td>
<td>30</td>
</tr>
<tr>
<td>IBM logging</td>
<td>102</td>
<td>0.098</td>
<td>9</td>
<td>12</td>
<td>0.137 (3; 1; 0)</td>
<td>28</td>
</tr>
<tr>
<td>IBM purchase</td>
<td>67</td>
<td>0.104</td>
<td>51</td>
<td>–</td>
<td>0.363 (5; 2; 6)</td>
<td>–</td>
</tr>
<tr>
<td>Beach Suites</td>
<td>64</td>
<td>0.125</td>
<td>31</td>
<td>29</td>
<td>0.353 (6; 2; 4)</td>
<td>30</td>
</tr>
<tr>
<td>1066 Pools Ltd</td>
<td>30</td>
<td>0.000</td>
<td>24</td>
<td>10</td>
<td>0.377 (6; 1; 8)</td>
<td>1</td>
</tr>
<tr>
<td>Art of War</td>
<td>24</td>
<td>0.083</td>
<td>12</td>
<td>17</td>
<td>0.117 (1; 1; 2)</td>
<td>30</td>
</tr>
<tr>
<td>WebSolutions</td>
<td>20</td>
<td>0.000</td>
<td>31</td>
<td>–</td>
<td>0.283 (7; 1; 2)</td>
<td>–</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>–</td>
<td>0.123</td>
<td>23.50</td>
<td>16.00</td>
<td></td>
<td>–</td>
</tr>
</tbody>
</table>
Leveraged analyses

• Analysis of vocabulary
• Analysis of constraints
• Analysis of metrics
• Analysis of cloning

• Analysis of extensions

© 2005-2013 Ralf Lämmel, Software Languages Team, and collaborators
Example of an extension

<RETENTION>
  <legal-requirement/>
  <EXTENSION optional="no">
    <use-duration>
      <one-year/>
    </use-duration>
  </retention-basis>
  <internal-company-regulation> [Text] </internal-company-regulation>
</EXTENSION>
</RETENTION>
retention : (retension-duration | retention-basis | use-duration)*

credit-privacy-law = (one-year)*
e-trade-law = (five-year)*
internal-company-regulation = ε
other-basis = ε
other-duration = ε
retension-duration = (other-duration)*
retention-basis = (credit-privacy-law | e-trade-law | internal-company-regulation | ...)*
use-duration = (other-duration | five-year | instance | one-month | one-year | ...)*
### Understanding Privacy Policies

**Statement:**
- `collection-method` = (other-method | delivery | document | punish | qnaboard | subscription | ...)
- `destruction-method` = (other-method | format | shatter)
- `other-method` = ε

**Purpose:**
- `purpose` = (account | browsing | delivery | feedback | finmgmt | government | login | marketing | ...)

**Recipient:**
- `recipient` = (delivery | jurisdiction | recipient-description | recipient-duration | recipient-name | same | delivery)

**Retention:**
- `retention` = (retention-duration | retention-basis | use-duration)

**Data-Group:**
- `data-group` = (data-group)

---

Some productions had to be cut off for scalability of presentation. Underlined tags correspond to elements without content. For brevity attributes of extensions are not shown in the grammar. When `=` is used as a rule separator, then text content instead of structured content was encountered for the element in question. All tags that use consistently P3P's XML namespace are shown in boldface. All tags that are used with different XML namespaces are shown in capitals and boldface. All the other tags use a unique XML namespace different from P3P's.
Results of the study

• Chosen approach to validation is too weak.

• There are common dominating policies.

• Interesting policies may be located through metrics.

• The study is non-conclusive on the need for extensions.

• The study promotes explorative, multi-dimensional analysis.
In this course, we could carry out research designs to study language usage in an exploratory or more definitive manner (e.g., in experiments). See also the related topic of API usage analysis.