


Predicting Code Churn from XML Metrics

Siim Karus
University of Tartu
Estonia




OSS Languages

Total

Language	Total	2010
XML	21%	14%
HTML	16%	11%
Java	10%	7%
Shell scripts, C	8%	5%
C++	6%	

Data from <http://www.ohloh.net>




OSS Code Churn

Churn in 2010


Language	Churn in 2010
XML	20%
Java	15%
C++	13%
C	11%
HTML	8%
Shell scripts	5%

Data from <http://www.ohloh.net>



Can we ...

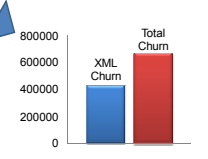
- Predict yearly XML Churn based on code metrics?
- Predict yearly LOC Churn based on code metrics?
- Predict yearly XML Churn based on organisational metrics?
- Predict yearly LOC Churn based on organisational metrics?




Theme

13 OSS projects
12 years of data

Version Control System
Revision Repository



Metric	Value
XML Churn	~400,000
Total Churn	~700,000



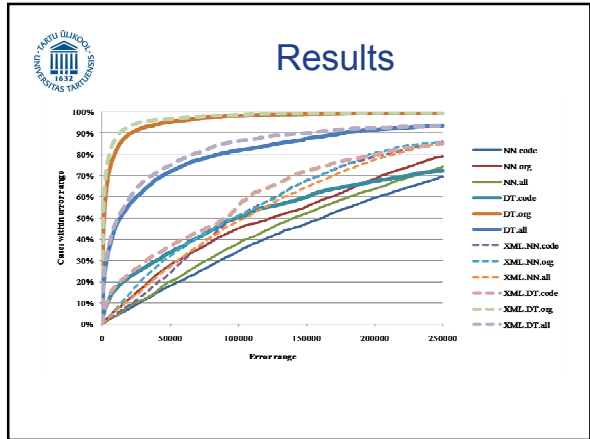
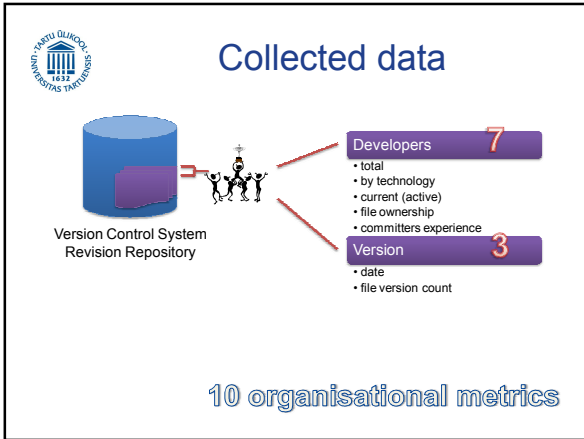
Collected Data

Version Control System
Revision Repository

project.xmlit

- Text file: 1
 - lines of code
- XML file: 7
 - nodes
 - nodes by type (elements, attributes, etc.)
 - max depth
 - avg depth
- XSL file: 55
 - expressions
 - expressions by type (inline, simple, complex)
 - elements by name (if, template, etc.)
 - etc.

63 code metrics



Results

Confidence	Error	XML Error
0.90	22 kLOC	13 kLOC
0.95	49 kLOC	27 kLOC
0.99	169 kLOC	107 kLOC
NMAE	0.0320	0.0314

-
- Conclusions**
- Models based on organisational data are highly suitable for code churn estimation (mean error less than 4%)
 - XML files offer better estimations of the whole project's churn than the churn of XML files
 - Organisational metrics extracted from version control systems are better predictors of code churn than code metrics

Thank you!

- Any questions?