Who could I join forces with?

reverse engineering
refactoring
Smalltalk

refactoring
developer tools
JavaScript
Who could I join forces with?

Combine refactoring knowledge for Smalltalk and JavaScript?
Not an easy question...

reverse engineering, refactoring, Smalltalk

refactoring, developer tools, JavaScript

Image source: https://www.shareicon.net
A possible solution
A possible solution
The visualization

• Explorable graph
• Visualizes a scientific community in terms of papers, authors and authorship
• Additional visualization: word cloud for paper bodies

-> Live demo (corpus: 1’100 papers, published at OOPSLA, 1986-2015)
This was just one possible query...

- How active was a certain field within the last couple of years?
- Is there a field that lately hasn’t been covered anymore?
- Has a specific field recently gained more interest?
- What else can we find out about this community?
- ...
Information is hidden in corpuses of papers

Why Smalltalk Wins the Host Languages Shootout

Lukas Renggli
renggli@iam.unibe.ch
Software Composition Group, University of Bern, Switzerland
http://scg.unibe.ch/

Tudor Girba
girba@iam.unibe.ch

Runtime bytecode transformation for Smalltalk

Marcus Denker\textsuperscript{a,}\textsuperscript{*}, Stéphane Ducasse\textsuperscript{a,}\textsuperscript{b}, Éric Tanter\textsuperscript{c}

\textsuperscript{a}Software Composition Group, IAM, Universitét Bern, Switzerland
\textsuperscript{b}Language and Software Evolution Group, LISTIC, Université de Savoie, France
\textsuperscript{c}Center for Web Research, DCC, University of Chile, Santiago, Chile

Received 17 September 2005

On the Integration of Smalltalk and Java

Practical Experience with STX:LIBJAVA

Marcel Hlopk\textsuperscript{a}
Czech Technical University in Prague
marcel.hlopk@ft.cvut.cz

Jan Kurš
Software Composition Group, University of Bern
kurs@iam.unibe.ch

Jan Vraný
Czech Technical University in Prague,
eXept Software AG
jan.vran@fit.cvut.cz

Claus Gittinger
eXept Software AG
cg@exapt.de
What we need
What we need
What we need

Image source: https://img.clipartfest.com
What we need

Image source: https://img.clipartfest.com
What we built

• A pipeline to convert PDFs into metadata-XML
• A query-able model of the underlying scientific community
• A graph and word cloud to visualize the model
• Builds on “EggShell”, by Dominik Seliner
The metadata extraction

- PDFBox: PDF to text
- ParsCit: text to metadata-XML

- Use third-party command line tools, used through controllers in ExtendedEggShell
- First PDF to text, then metadata extraction from text

**PDFBox**

**Command-Line Tools**

PDFBox comes with a series of command-line utilities. They are available as standard Java applications. See the Dependencies page for instructions on how to set your classpath in order to run PDFBox tools as Java applications.

**Table of Contents**
- Encrypt
- EncryptImages
- Extract
- ExtractCaption
- ExtractTable
- OverlayPDF
- PrintPDF
- PDFDebugger
- PDFReader
- PDFMerge
- PDFJoin
- PDFStitch
- PDFImage
- TextToPDF
- WriteDecodedDoc

**ParsCit**

The Scientific Community model

- **ScientificCommunity**
  - **PapersCollection**
    - **PaperModel**
      - title: String
      - sectionHeaders: Collection<String>
      - citedPapers: PapersCollection
      - bodyText: String
      - fileReference: FileReference
  - **AuthorsCollection**
    - **AuthorModel**
      - authorName: String
      - affiliation: String
      - email: String
  - **AbstractCommunityItem**
    - communityItemId: int
    - isCollected: boolean
    - additionalInfo: Dictionary
Initial model isn’t perfect

• Non-alphanumeric characters in author names
• Multiple spellings of the same name
• Bad name extraction
Model clean-up: heuristics framework
Model clean-up: heuristics framework
Model clean-up: heuristics runner

CleanupHeuristicsRunner

- heuristics: Collection<AbstractCleanupHeuristic class>
- scientificCommunity

+ on: (aScientificCommunity)
+ addHeuristic(aHeuristicClass)
+ addDefaultHeuristics() 
+ run()
Model clean-up: heuristics framework
Model clean-up: example heuristics

AbstractAuthorCleanupHeuristic

- ACHBlacklist
- ACHCommonSubstring
- ACHSameCharClass
Model clean-up: before and after
Conclusion

- Started with a complex, large dataset and built an explorable model and visualization, that allows for extracting insights from and about the dataset
- Users can create custom queries and have them answered visually
- Query-able model allows for easy creation of custom visualizations
- Scalability is limited
Future work

- Further improve the model
- Web crawler for fetching cited papers
- Suggest related papers for some paper or author
Some handy model query methods

**PaperModel**
- `numberOfBodyOccurrences(query)`
- `bodyContains(query)`
- `bodyContainsAll(query)`
- `bodyContainsAny(query)`

**PapersCollection**
- `atTitle(query)`
- `atTitleLike(query)`
- `atAuthorNameSubstring(query)`
Some handy model query methods

AuthorsCollection

+atAuthorNameLike(query)
+atAuthorNameLike(query)
+atAuthorNameSubstring(query)
sc := DemoModelLoader oopslaCollection.

graph := SCGraph new
  on: sc;
  title: 'OOPSLA collection SC graph';
  load.

hlFactory := GraphHighlightingFactory new graph: graph.

wordCloudView := SCWordCloudView new.
relatedPapers := sc papers whereBodyContainsAny: #( 'programming' 'object-oriented' 'software' ).
relatedPapersHL := hlfactory
    highlight: relatedPapers
    callIt: 'Some related papers'.

relatedPapersHL remove.

relatedPapers := sc papers atTitleSubString: 'refactoring'.
relatedPapersHL := hlFactory
  highlight: relatedPapers
  callIt: 'Papers about refactoring'.
relatedPapersHL remove.
relatedPapersHL := hlFactory
    highlight: relatedPapers
callIt: 'Papers about refactoring'
normalizerBlock: [:each | each year].
interestingAuthors := {
    sc authors atAuthorName: 'Anders Møller'.
    sc authors atAuthorName: 'Asger Feldthaus'. }

interestingAuthorsHL := hlFactory
    highlight: interestingAuthors
    callIt: 'Some interesting authors'.

papersMollerFeldthaus := sc papersByAny: #('Anders Møller', 'Asger Feldthaus').
wordCloudView buildCloudFor: papersMollerFeldthaus.