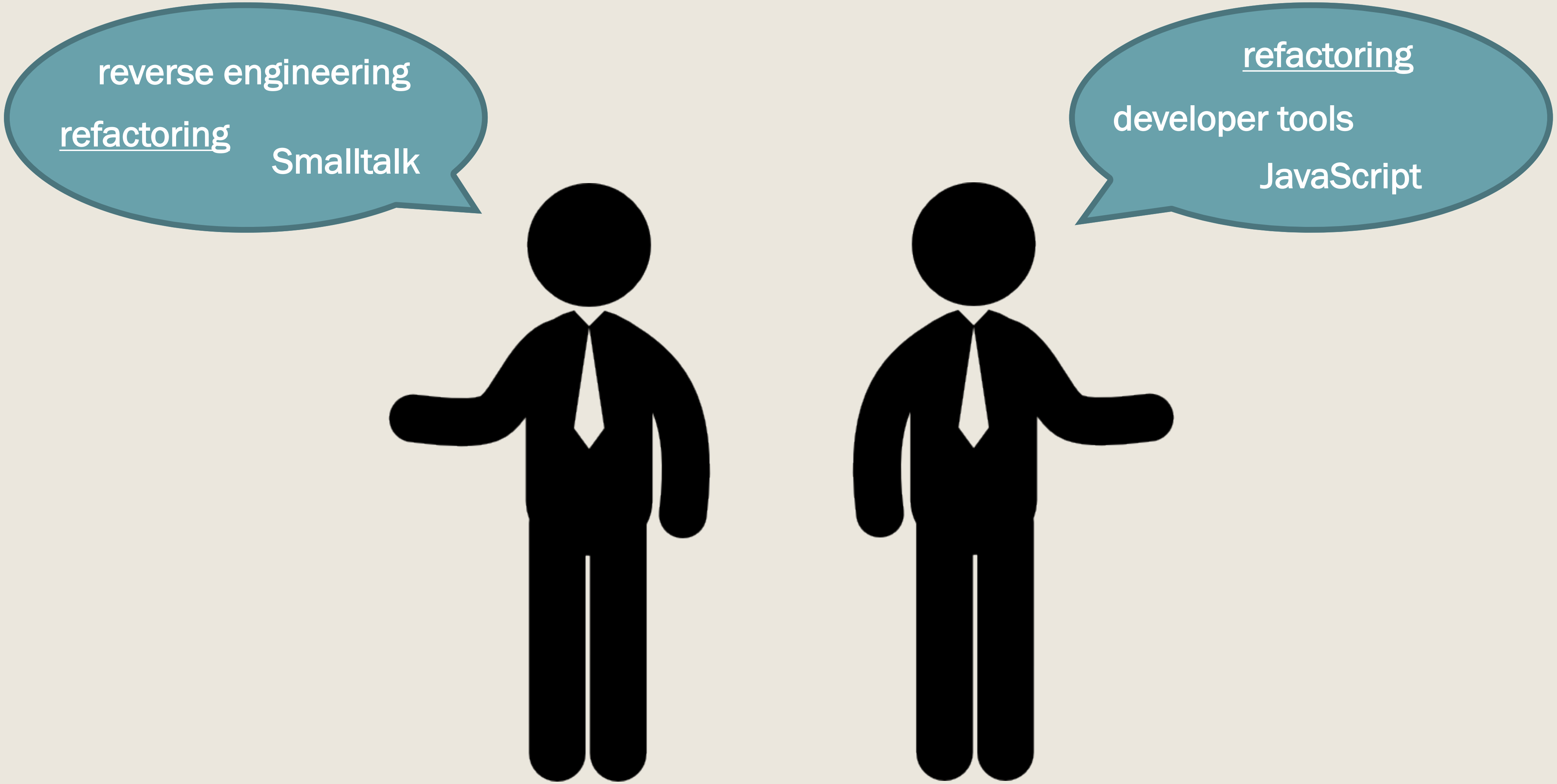
VISUALLY MINING SCIENTIFIC COMMUNITIES

Bachelor Thesis by Silas Berger (silas.berger@students.unibe.ch),
supervised by Leonel Merino

Who could I join forces with?



reverse engineering

refactoring

Smalltalk

refactoring

developer tools

JavaScript

Image source: https://www.shareicon.net

Who could I join forces with?

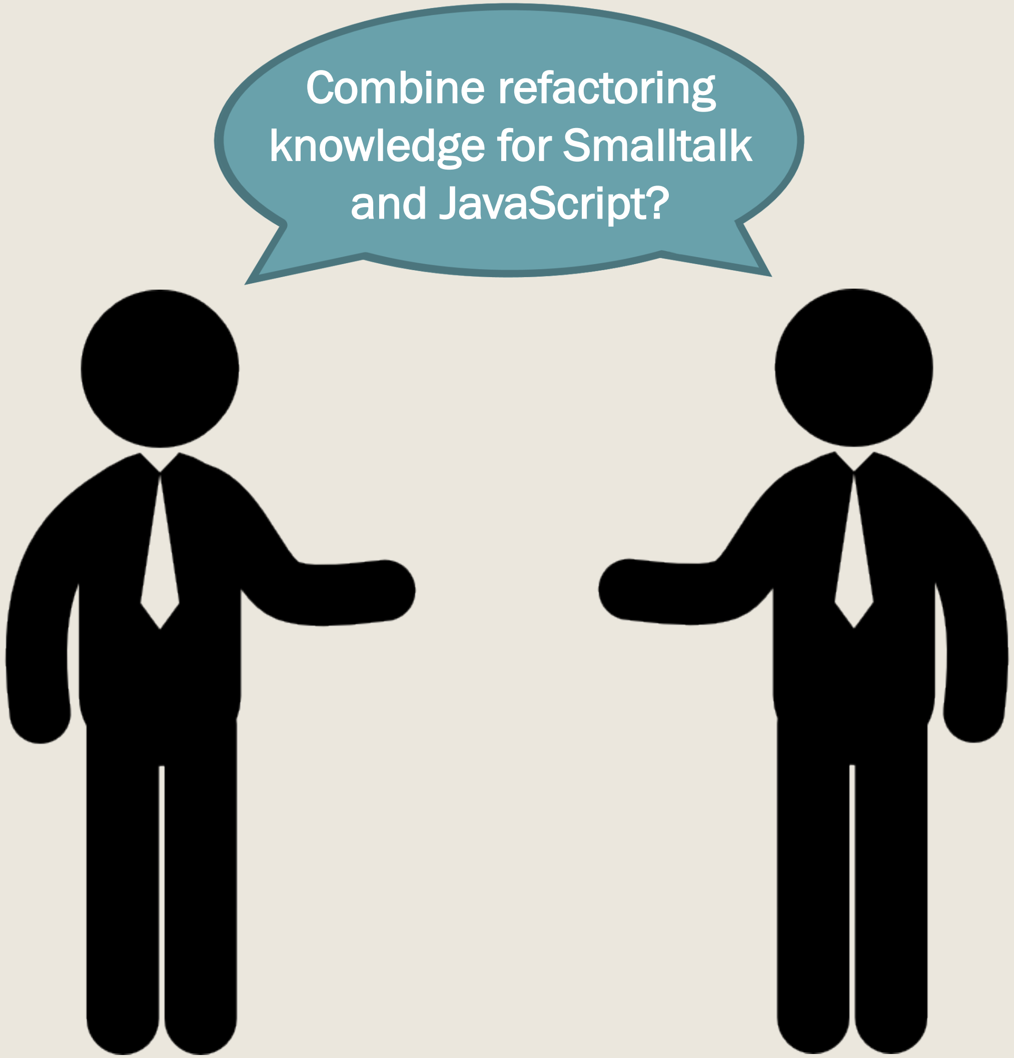


Image source: https://www.shareicon.net

Not an easy question…

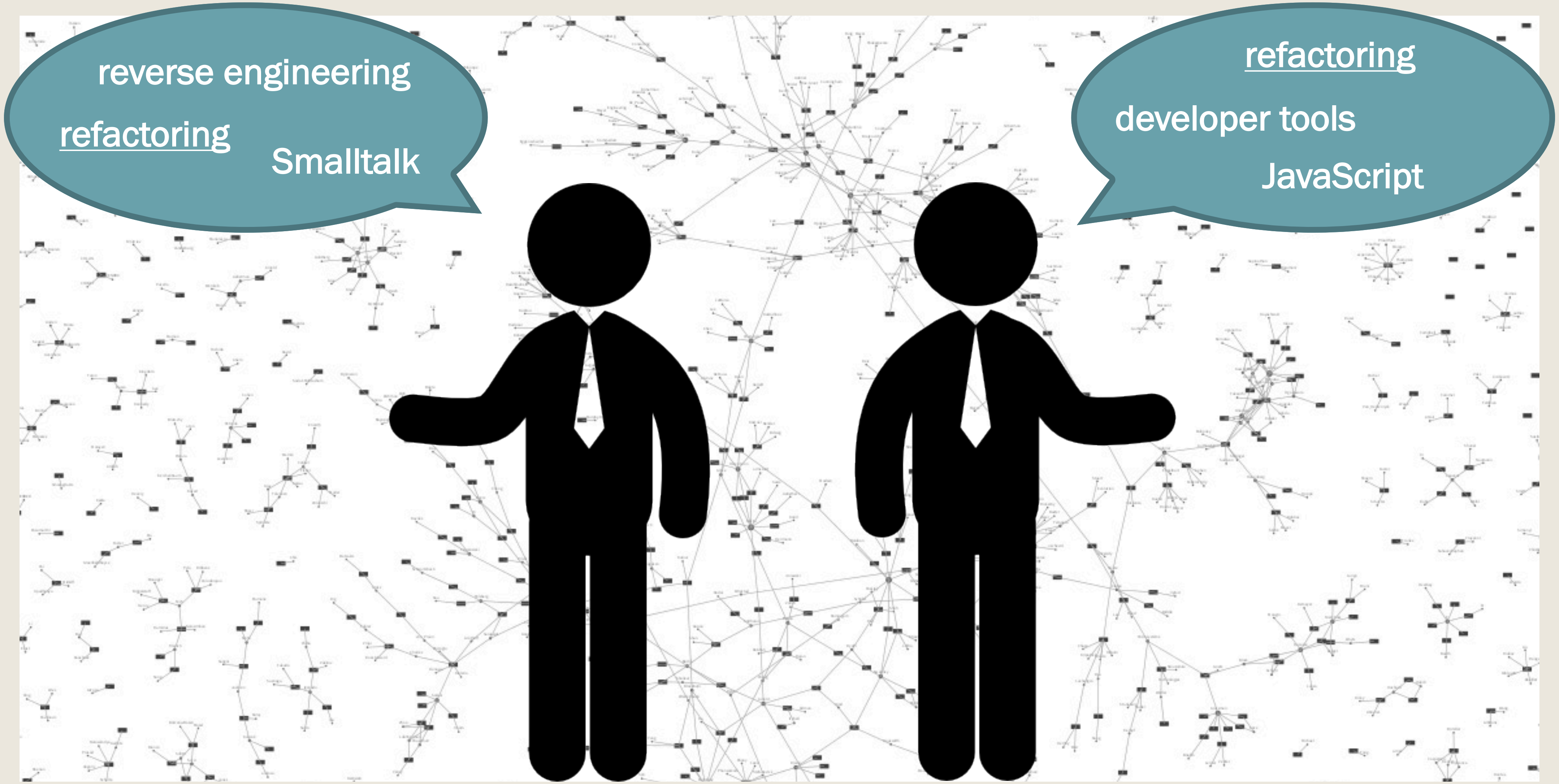
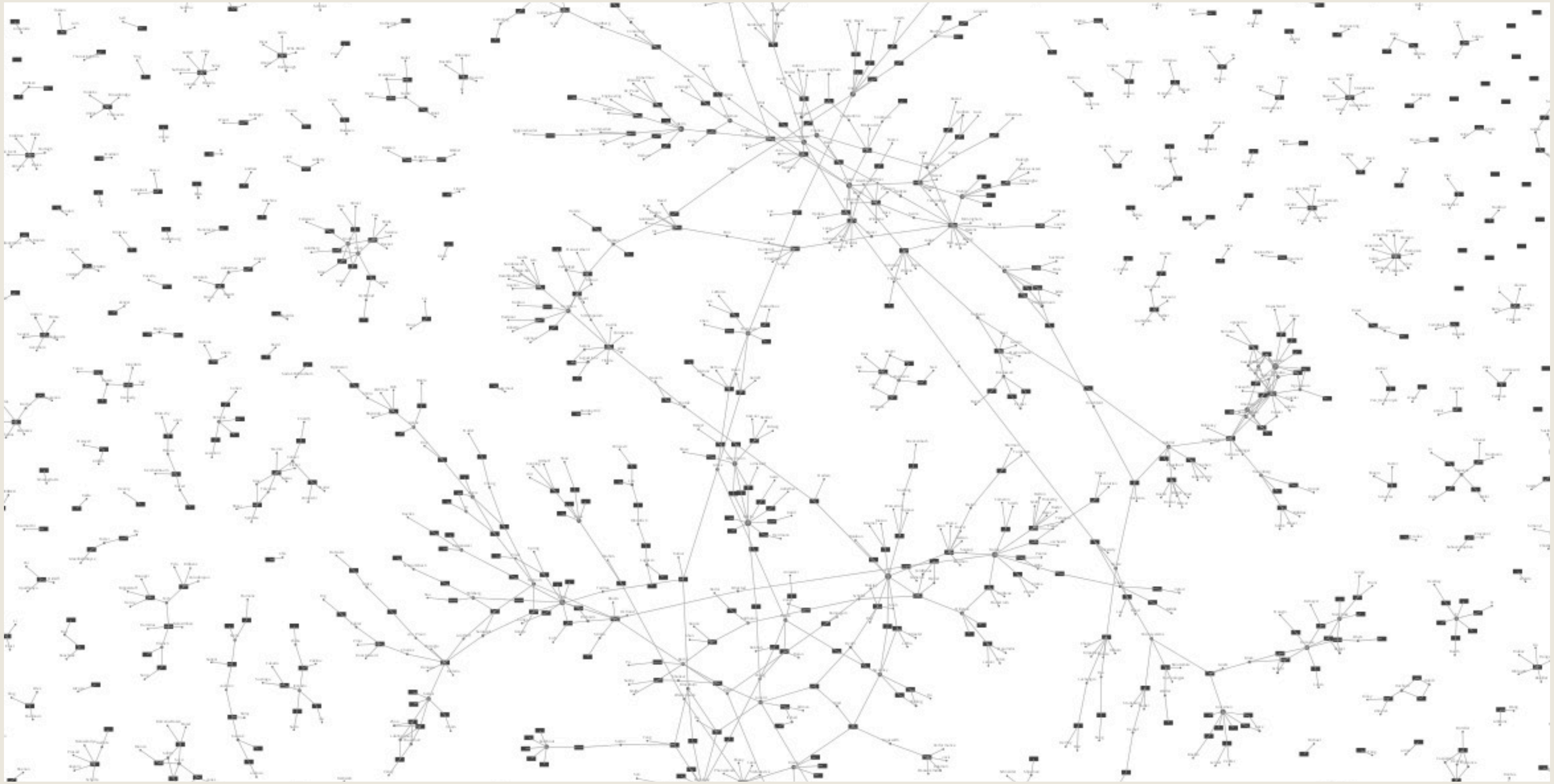
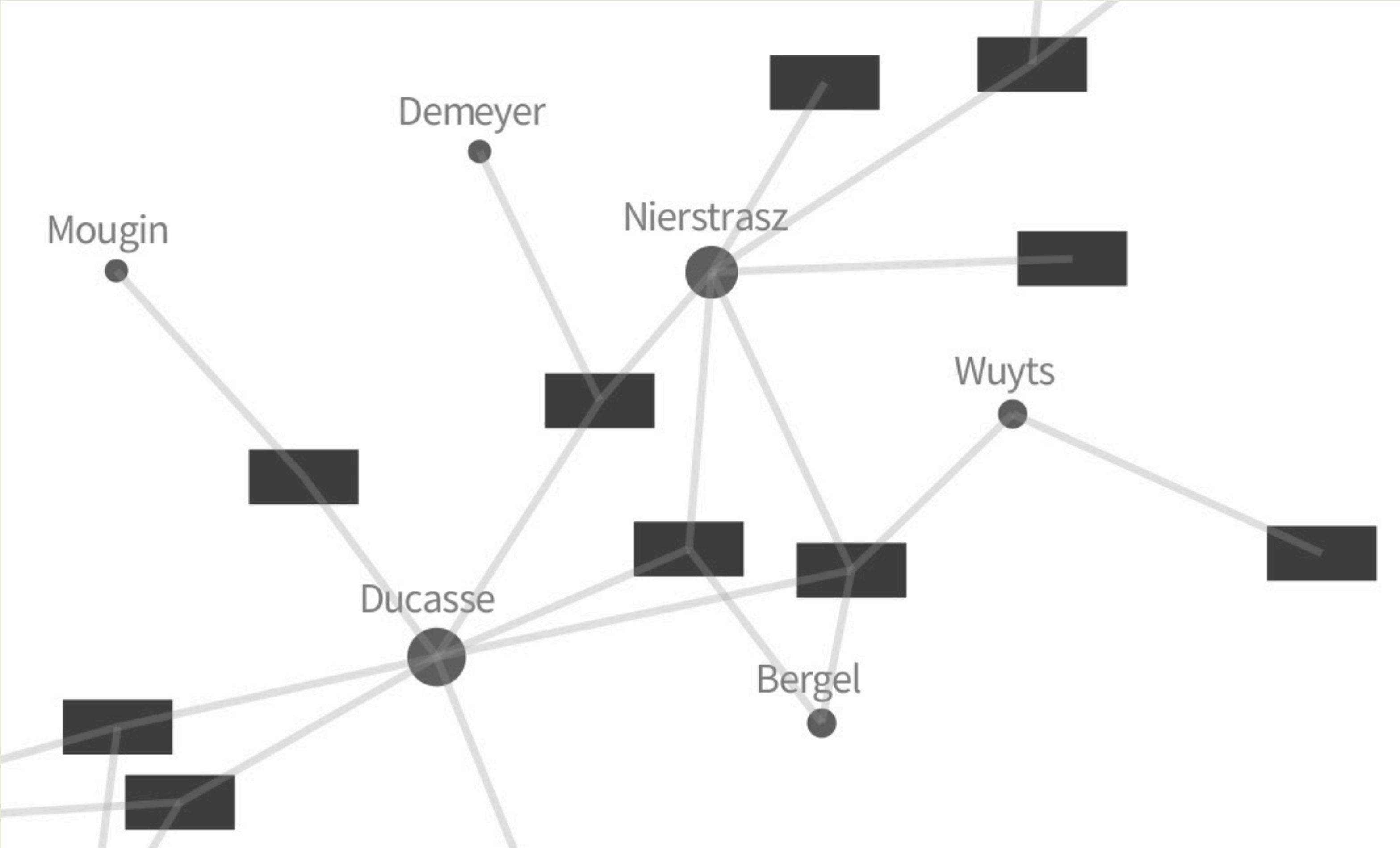


Image source: https://www.shareicon.net

A possible solution



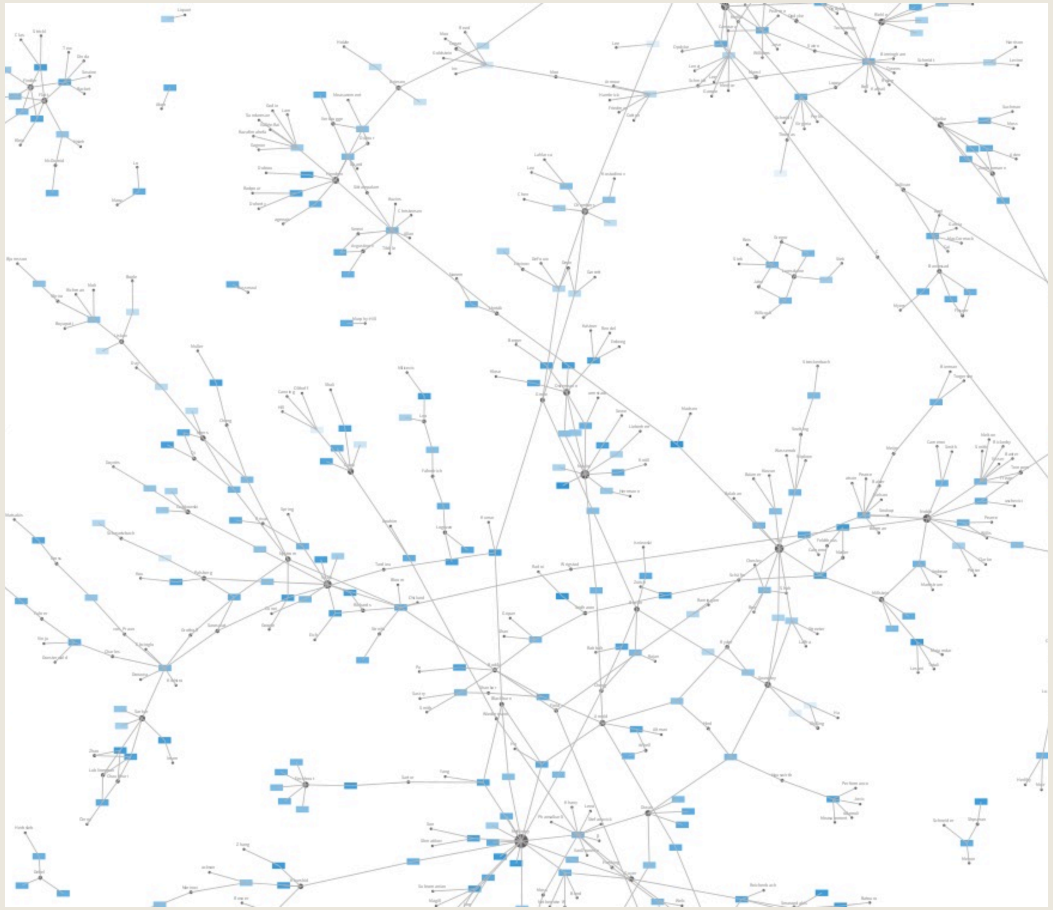
A possible solution



The visualization

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

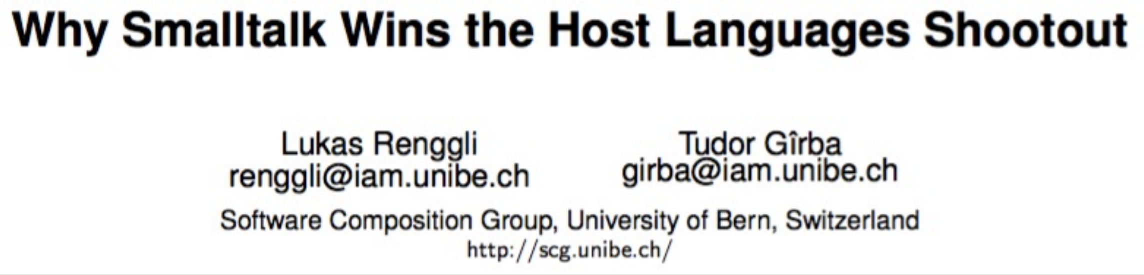
-> Live demo (corpus: 1’100 papers, published at OOPSLA, 1986-2015)



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

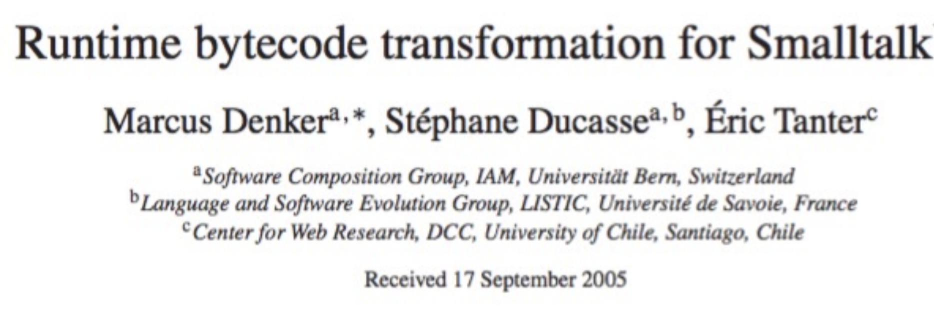


Why Smalltalk Wins the Host Languages Shootout

Lukas Renggli
renggli@iam.unibe.ch

Tudor Gîrba
girba@iam.unibe.ch

Software Composition Group, University of Bern, Switzerland
http://scg.unibe.ch/

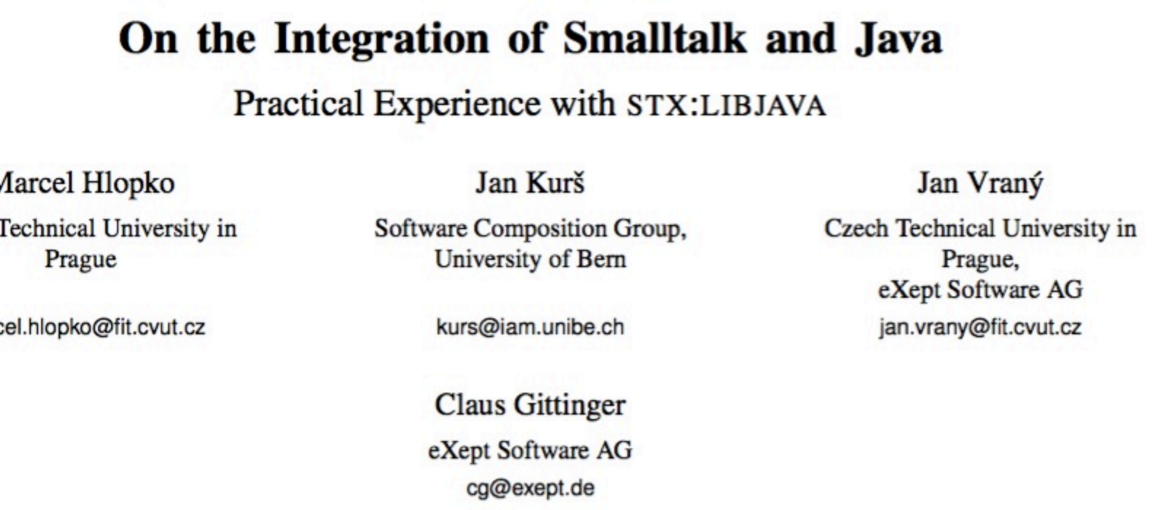


Runtime bytecode transformation for Smalltalk☆

Marcus Denker[a,*], Stéphane Ducasse[a,b], Éric Tanter[c]

[a] *Software Composition Group, IAM, Universität Bern, Switzerland*
[b] *Language and Software Evolution Group, LISTIC, Université de Savoie, France*
[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 Hlopko
Czech Technical University in Prague
marcel.hlopko@fit.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.vrany@fit.cvut.cz

Claus Gittinger
eXept Software AG
cg@exept.de

What we need

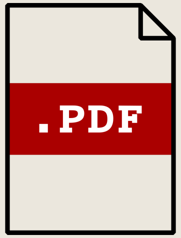
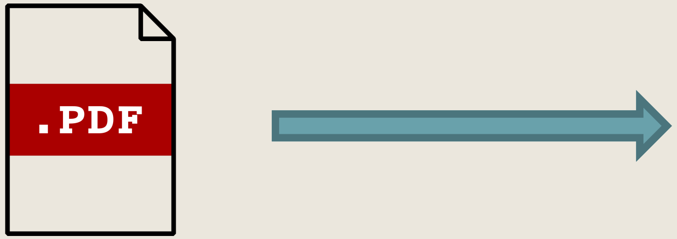
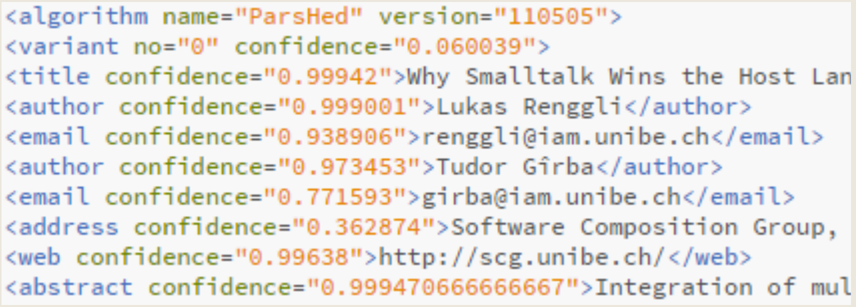


Image source: https://img.clipartfest.com

What we need

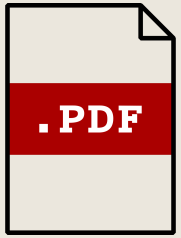


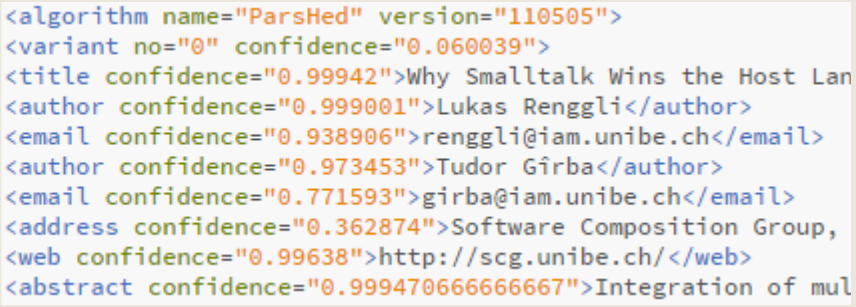


```
<algorithm name="ParsHed" version="110505">
<variant no="0" confidence="0.060039">
<title confidence="0.99942">Why Smalltalk Wins the Host Lan
<author confidence="0.999001">Lukas Renggli</author>
<email confidence="0.938906">renggli@iam.unibe.ch</email>
<author confidence="0.973453">Tudor Gîrba</author>
<email confidence="0.771593">girba@iam.unibe.ch</email>
<address confidence="0.362874">Software Composition Group,
<web confidence="0.99638">http://scg.unibe.ch/</web>
<abstract confidence="0.999470666666667">Integration of mul
```

Image source: https://img.clipartfest.com

What we need





```
<algorithm name="ParsHed" version="110505">
<variant no="0" confidence="0.060039">
<title confidence="0.99942">Why Smalltalk Wins the Host Lan
<author confidence="0.999001">Lukas Renggli</author>
<email confidence="0.938906">renggli@iam.unibe.ch</email>
<author confidence="0.973453">Tudor Gîrba</author>
<email confidence="0.771593">girba@iam.unibe.ch</email>
<address confidence="0.362874">Software Composition Group,
<web confidence="0.99638">http://scg.unibe.ch/</web>
<abstract confidence="0.999470666666667">Integration of mul
```

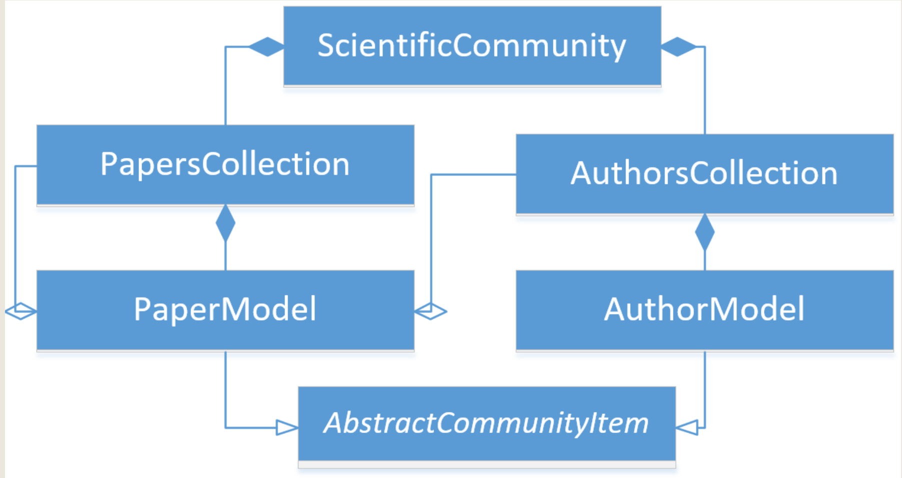
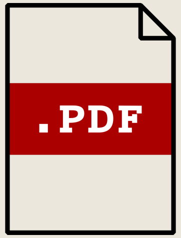
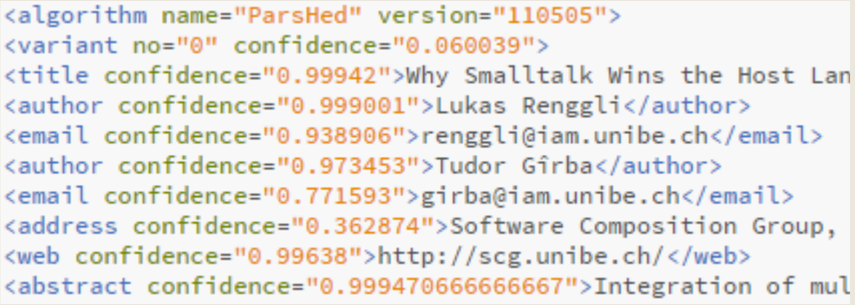


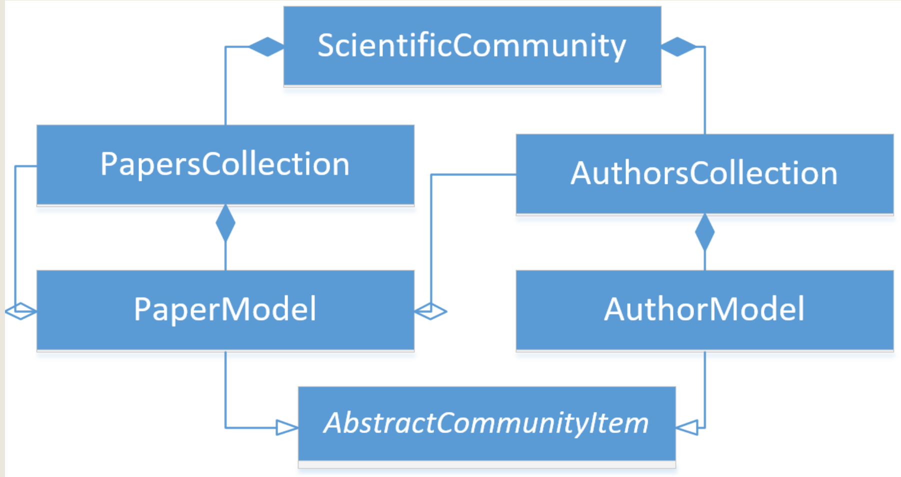
Image source: https://img.clipartfest.com

What we need





```
<algorithm name="ParsHed" version="110505">
<variant no="0" confidence="0.060039">
<title confidence="0.99942">Why Smalltalk Wins the Host Lan
<author confidence="0.999001">Lukas Renggli</author>
<email confidence="0.938906">renggli@iam.unibe.ch</email>
<author confidence="0.973453">Tudor Gîrba</author>
<email confidence="0.771593">girba@iam.unibe.ch</email>
<address confidence="0.362874">Software Composition Group,
<web confidence="0.99638">http://scg.unibe.ch/</web>
<abstract confidence="0.999470666666667">Integration of mul
```



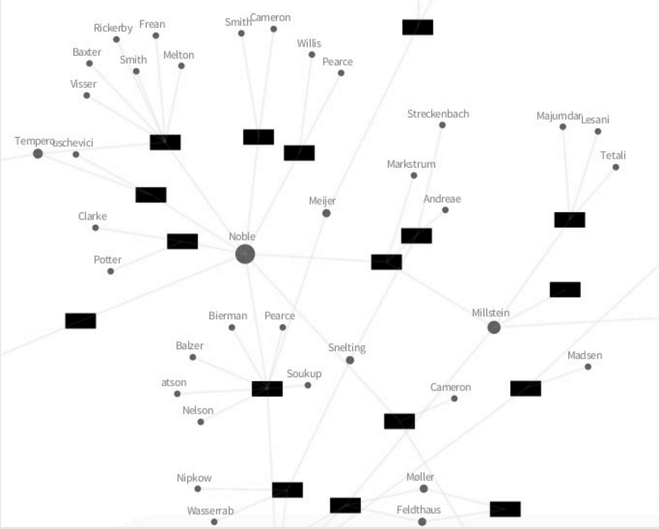
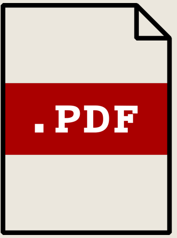


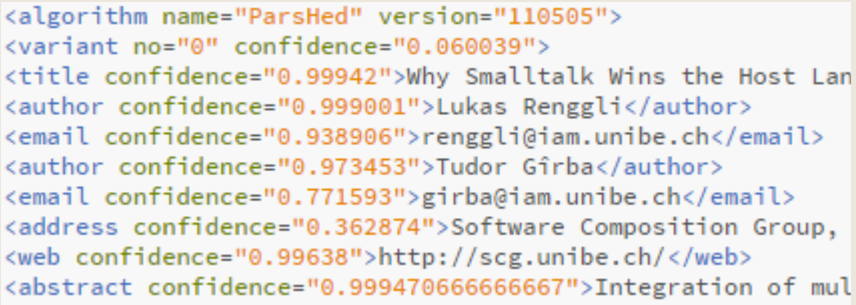
Image source: https://img.clipartfest.com

What we built

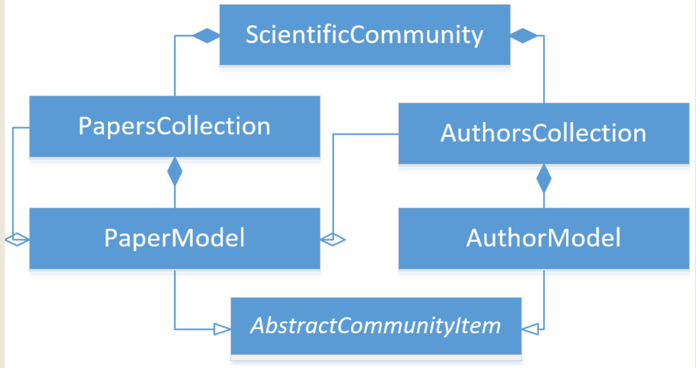
- A pipeline to convert PDFs into metadata-XML
- A query-able model of the underlying scientific community
- A graph and word cloud to visualize the model
- Builds on "EggShell", by Dominik Seliner

ExtendedEggShell





```
<algorithm name="ParsHed" version="110505">
<variant no="0" confidence="0.060039">
<title confidence="0.99942">Why Smalltalk Wins the Host Lan
<author confidence="0.999001">Lukas Renggli</author>
<email confidence="0.938906">renggli@iam.unibe.ch</email>
<author confidence="0.973453">Tudor Gîrba</author>
<email confidence="0.771593">girba@iam.unibe.ch</email>
<address confidence="0.362874">Software Composition Group,
<web confidence="0.99638">http://scg.unibe.ch/</web>
<abstract confidence="0.999470666666667">Integration of mul
```



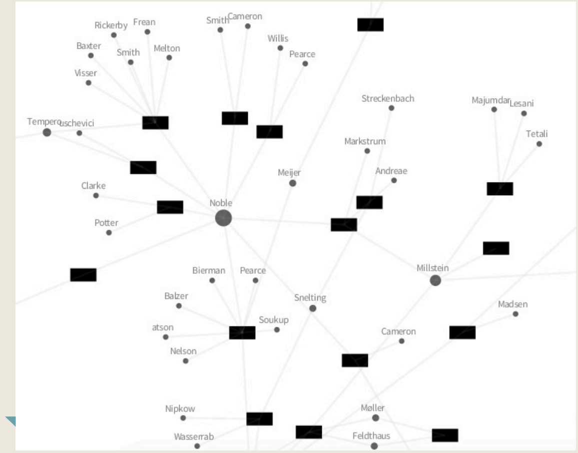
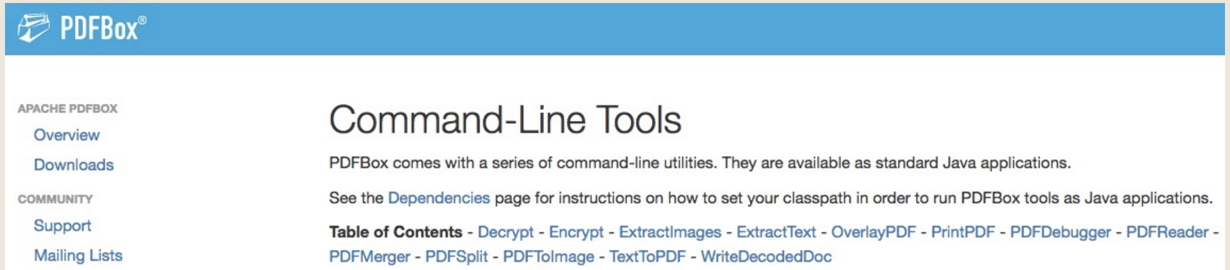


Image source: https://img.clipartfest.com

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

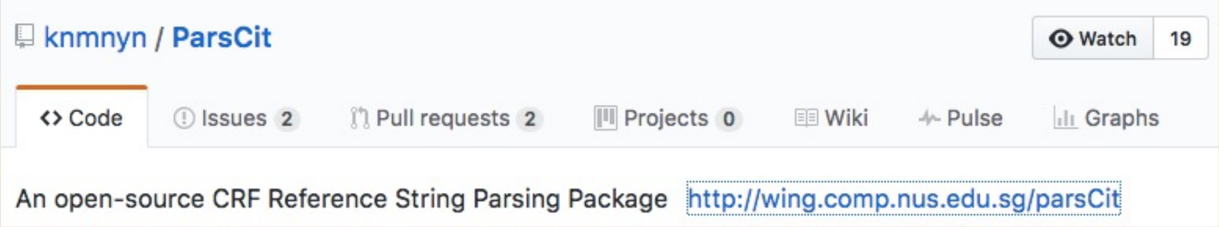


Command-Line Tools

PDFBox comes with a series of command-line utilities. They are available as standard Java applications.

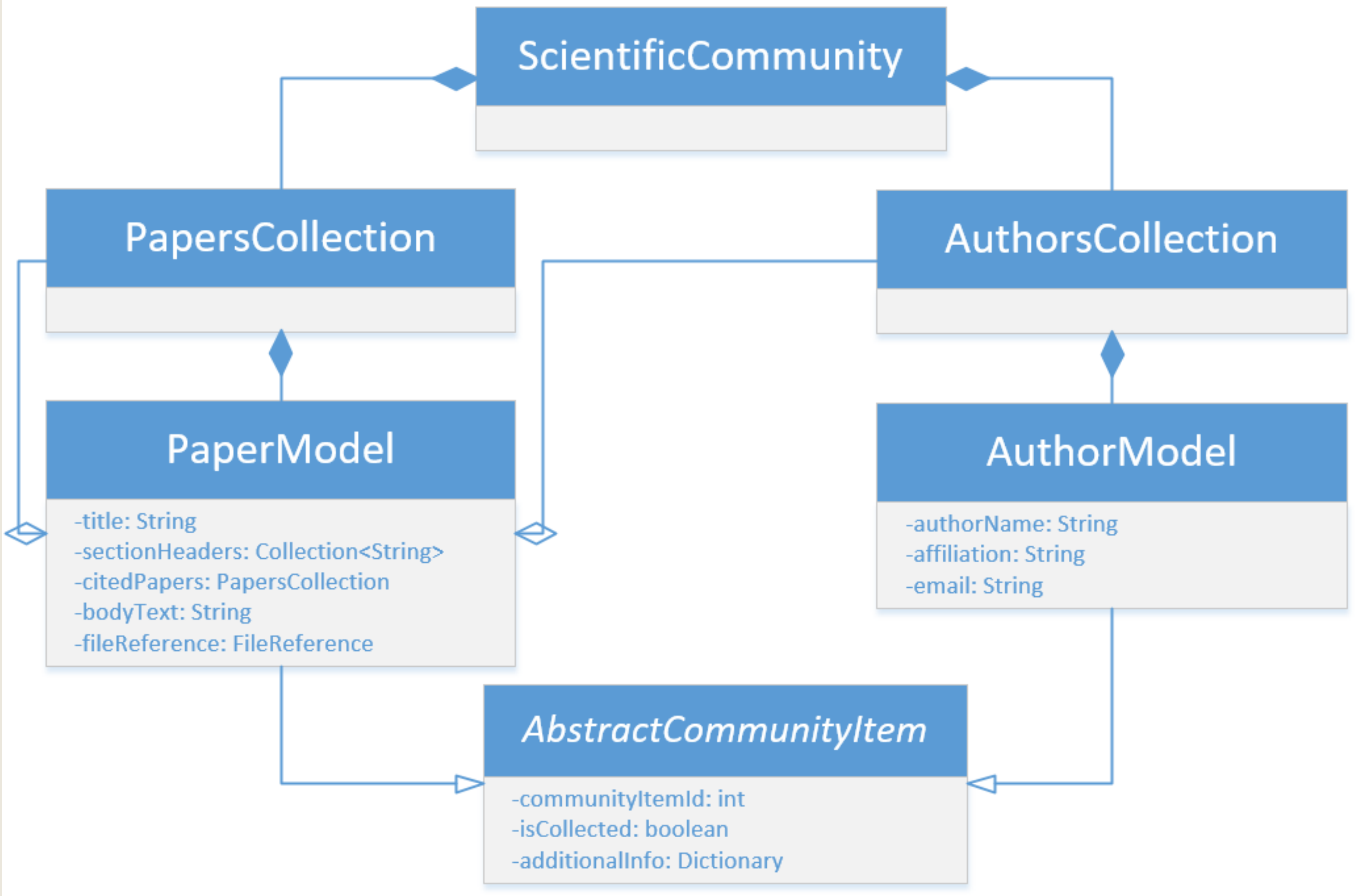
See the Dependencies page for instructions on how to set your classpath in order to run PDFBox tools as Java applications.

Table of Contents - Decrypt - Encrypt - ExtractImages - ExtractText - OverlayPDF - PrintPDF - PDFDebugger - PDFReader - PDFMerger - PDFSplit - PDFToImage - TextToPDF - WriteDecodedDoc



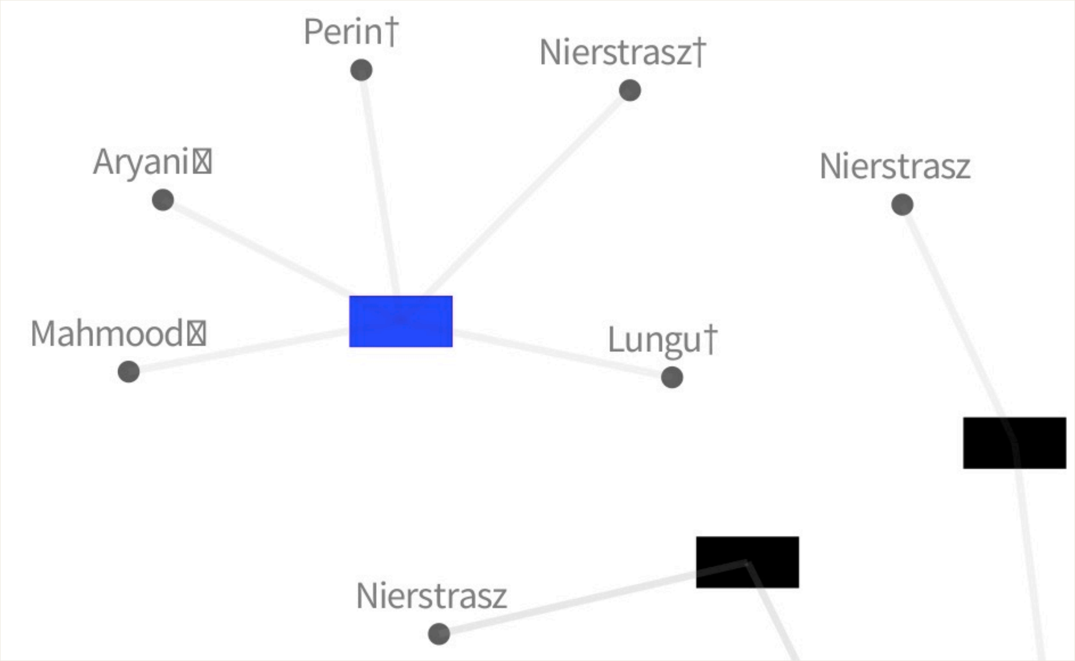
An open-source CRF Reference String Parsing Package http://wing.comp.nus.edu.sg/parsCit

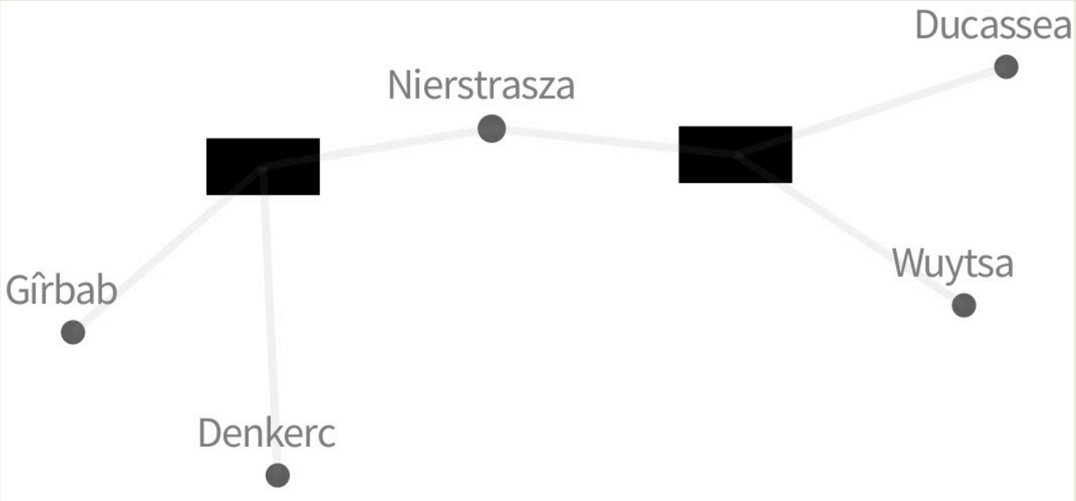
The Scientific Community model



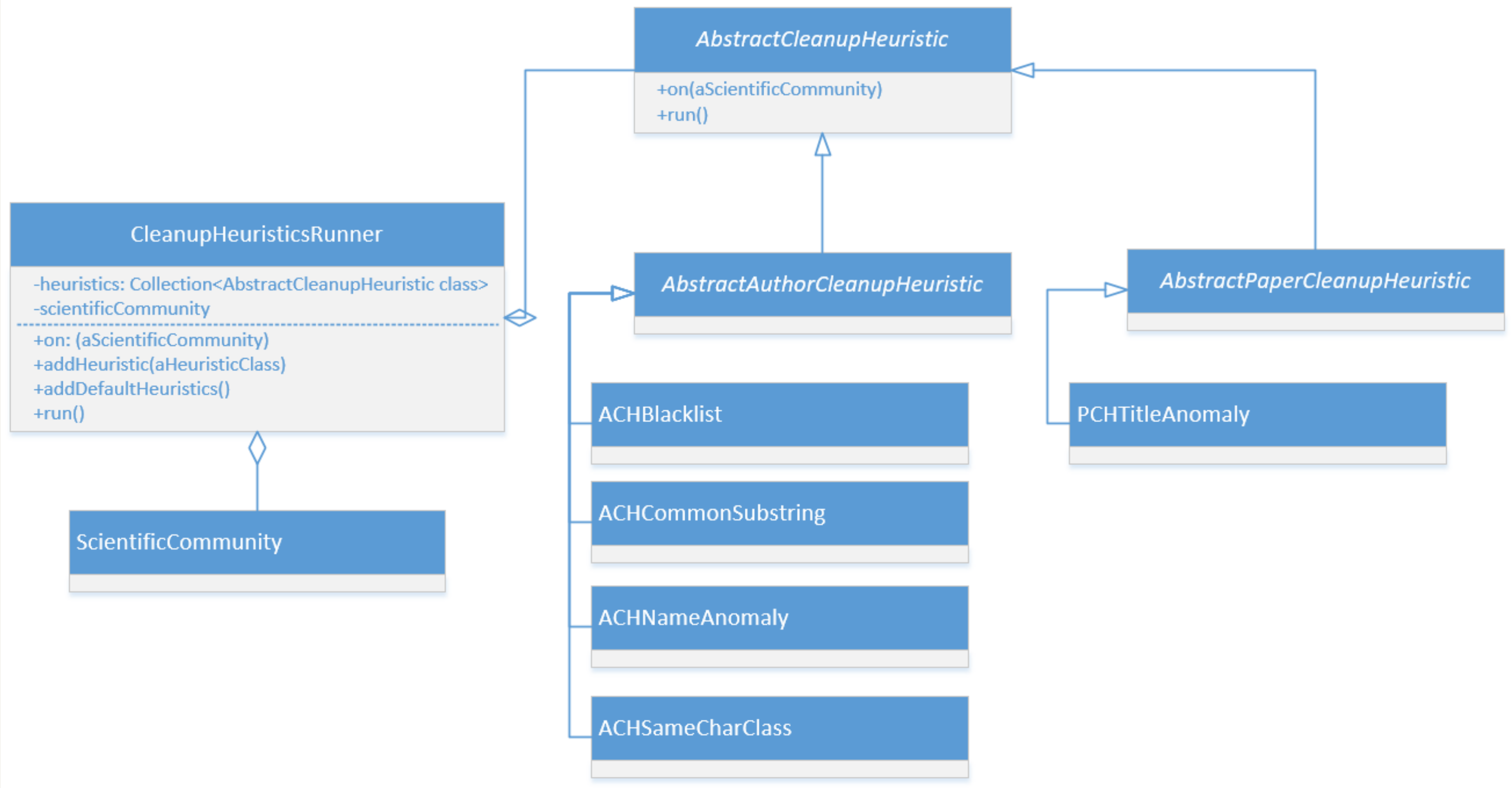
Initial model isn’t perfect

- Non-alphanumeric characters in author names
- Multiple spellings of the same name
- Bad name extraction

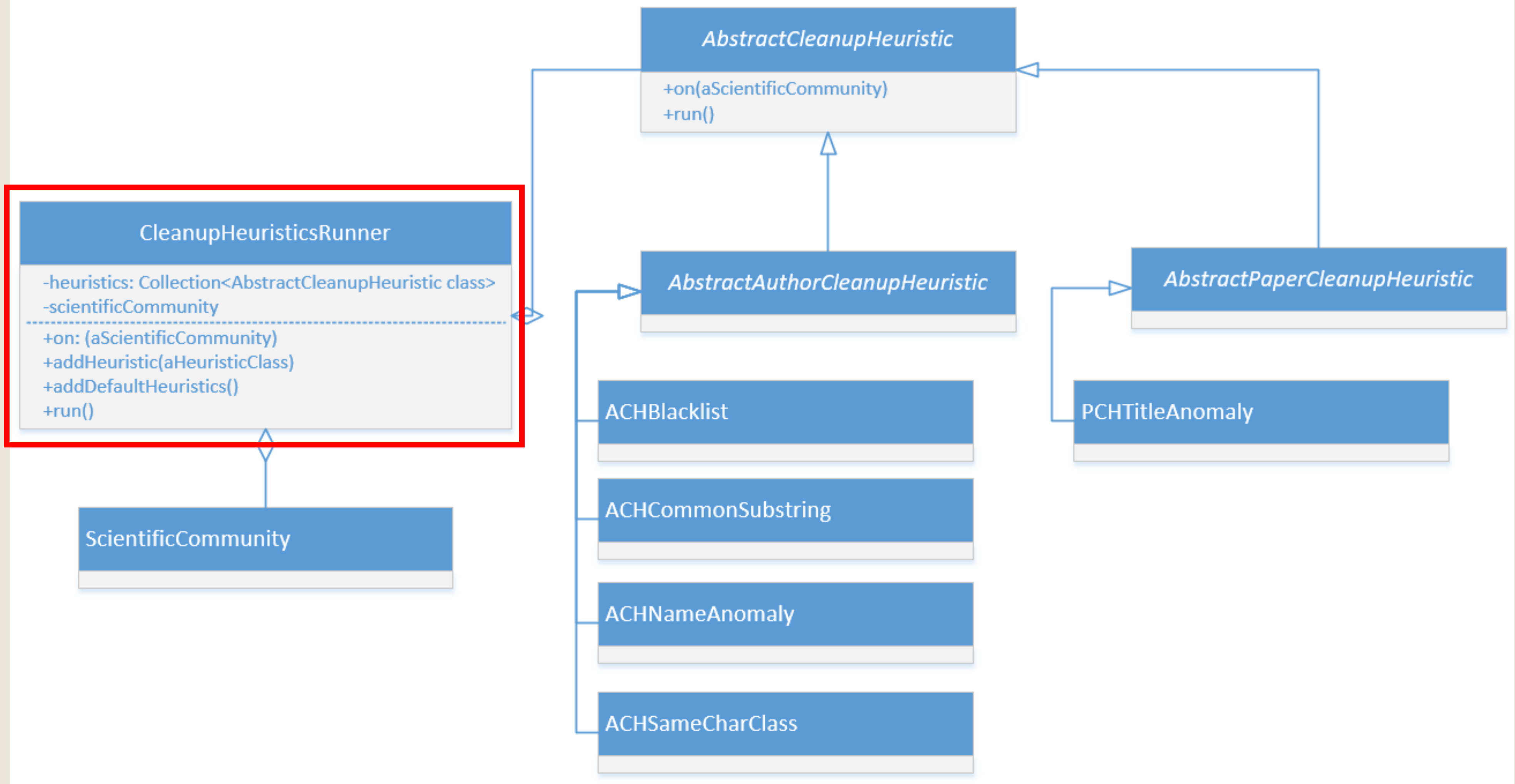




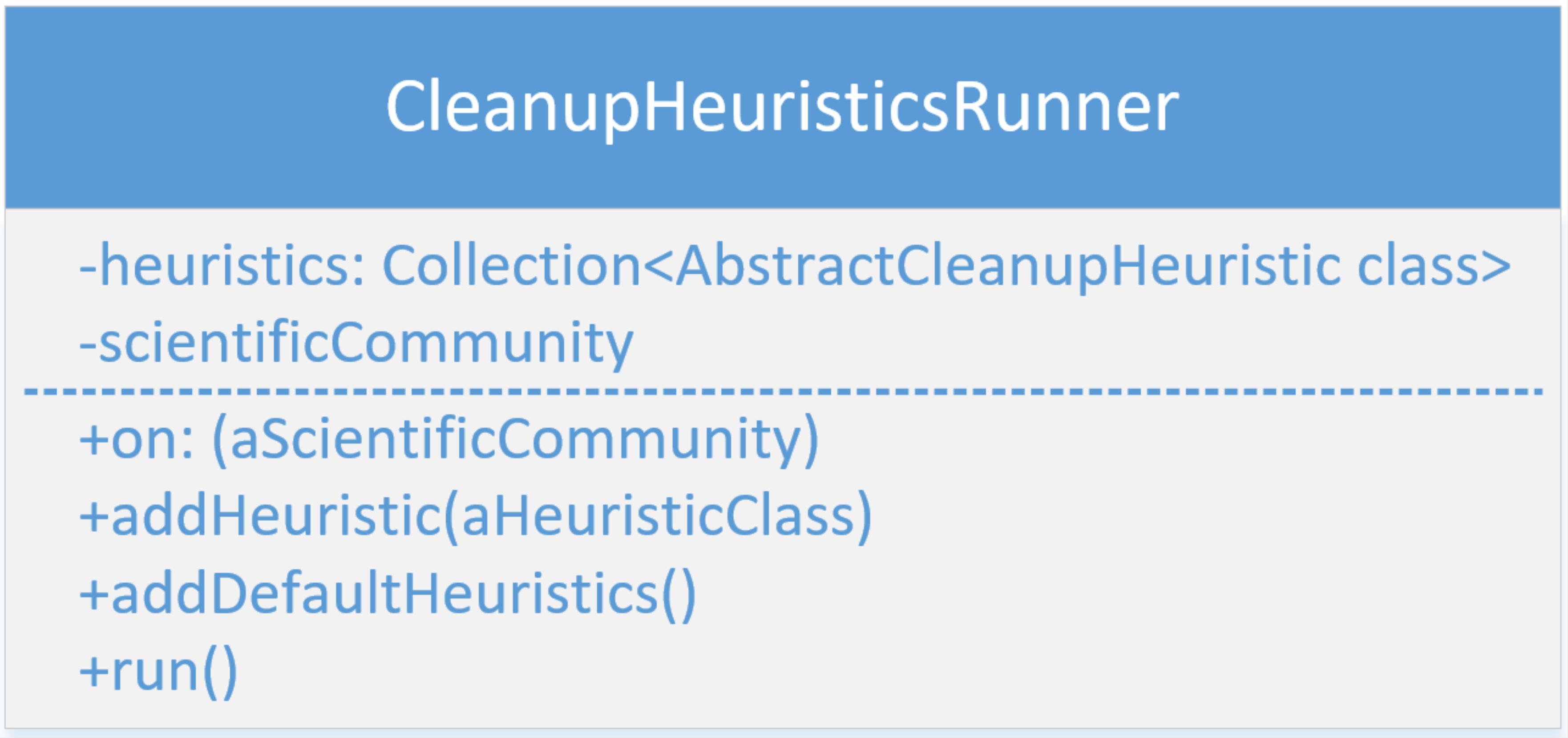
Model clean-up: heuristics framework


AbstractCleanupHeuristic
+on(aScientificCommunity)
+run()
CleanupHeuristicsRunner
-heuristics: Collection<AbstractCleanupHeuristic class>
-scientificCommunity
+on: (aScientificCommunity)
+addHeuristic(aHeuristicClass)
+addDefaultHeuristics()
+run()
ScientificCommunity
AbstractAuthorCleanupHeuristic
AbstractPaperCleanupHeuristic
ACHBlacklist
ACHCommonSubstring
ACHNameAnomaly
ACHSameCharClass
PCHTitleAnomaly

Model clean-up: heuristics framework


AbstractCleanupHeuristic
+on(aScientificCommunity)
+run()
CleanupHeuristicsRunner
-heuristics: Collection<AbstractCleanupHeuristic class>
-scientificCommunity
+on: (aScientificCommunity)
+addHeuristic(aHeuristicClass)
+addDefaultHeuristics()
+run()
AbstractAuthorCleanupHeuristic
AbstractPaperCleanupHeuristic
ACHBlacklist
PCHTitleAnomaly
ACHCommonSubstring
ScientificCommunity
ACHNameAnomaly
ACHSameCharClass

Model clean-up: heuristics runner

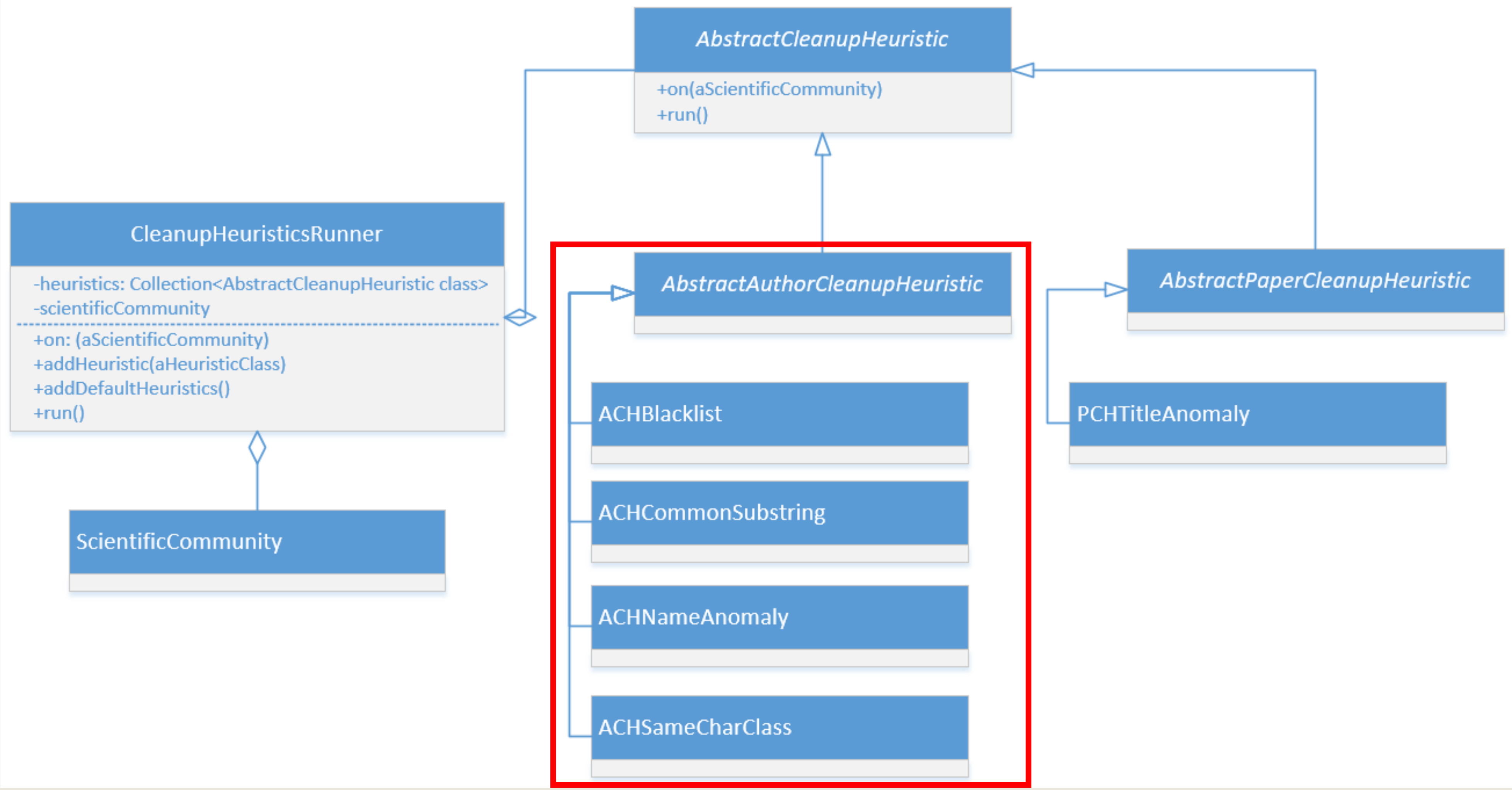


CleanupHeuristicsRunner

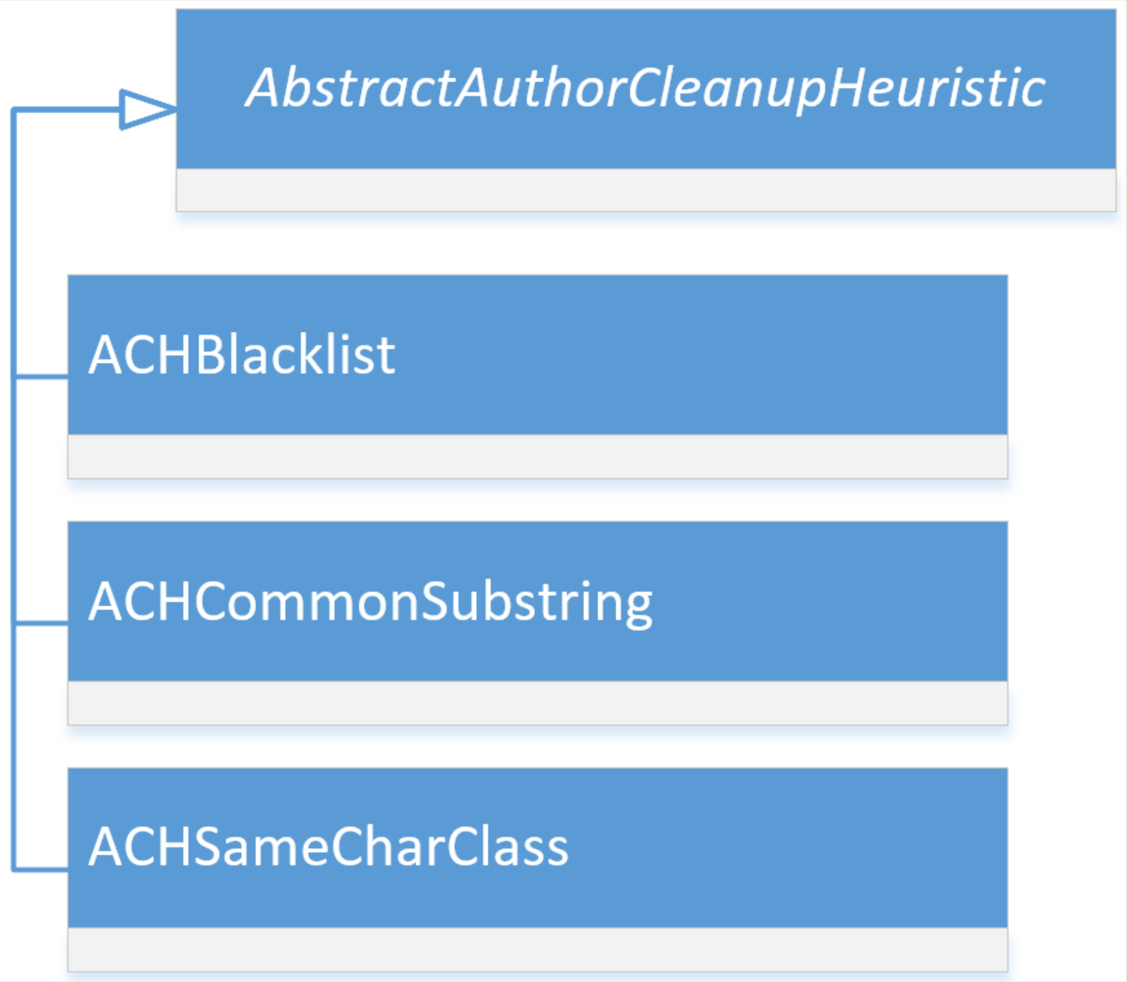
-heuristics: Collection<AbstractCleanupHeuristic class>
-scientificCommunity

+on: (aScientificCommunity)
+addHeuristic(aHeuristicClass)
+addDefaultHeuristics()
+run()

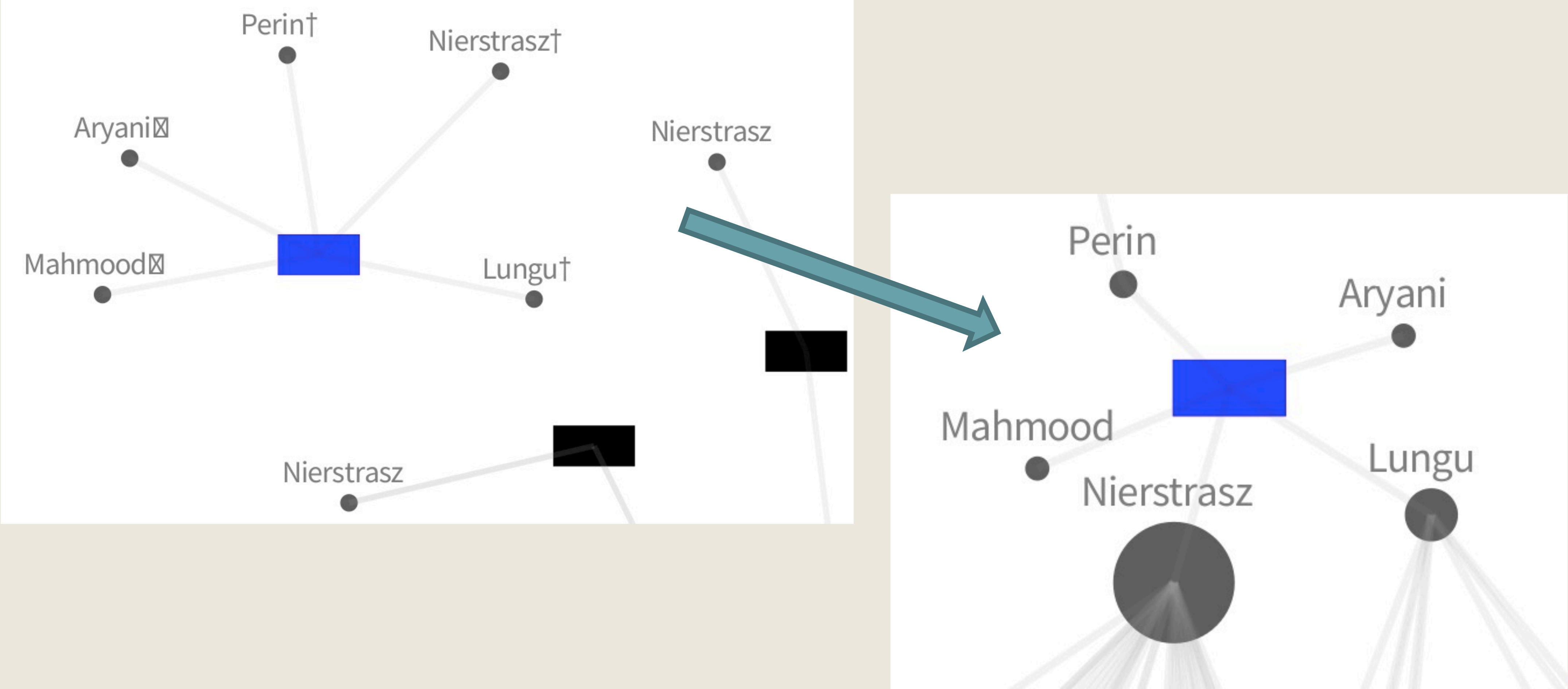
Model clean-up: heuristics framework


AbstractCleanupHeuristic
+on(aScientificCommunity)
+run()
CleanupHeuristicsRunner
-heuristics: Collection<AbstractCleanupHeuristic class>
-scientificCommunity
+on: (aScientificCommunity)
+addHeuristic(aHeuristicClass)
+addDefaultHeuristics()
+run()
ScientificCommunity
AbstractAuthorCleanupHeuristic
ACHBlacklist
ACHCommonSubstring
ACHNameAnomaly
ACHSameCharClass
AbstractPaperCleanupHeuristic
PCHTitleAnomaly

Model clean-up: example heuristics



Model clean-up: before and after



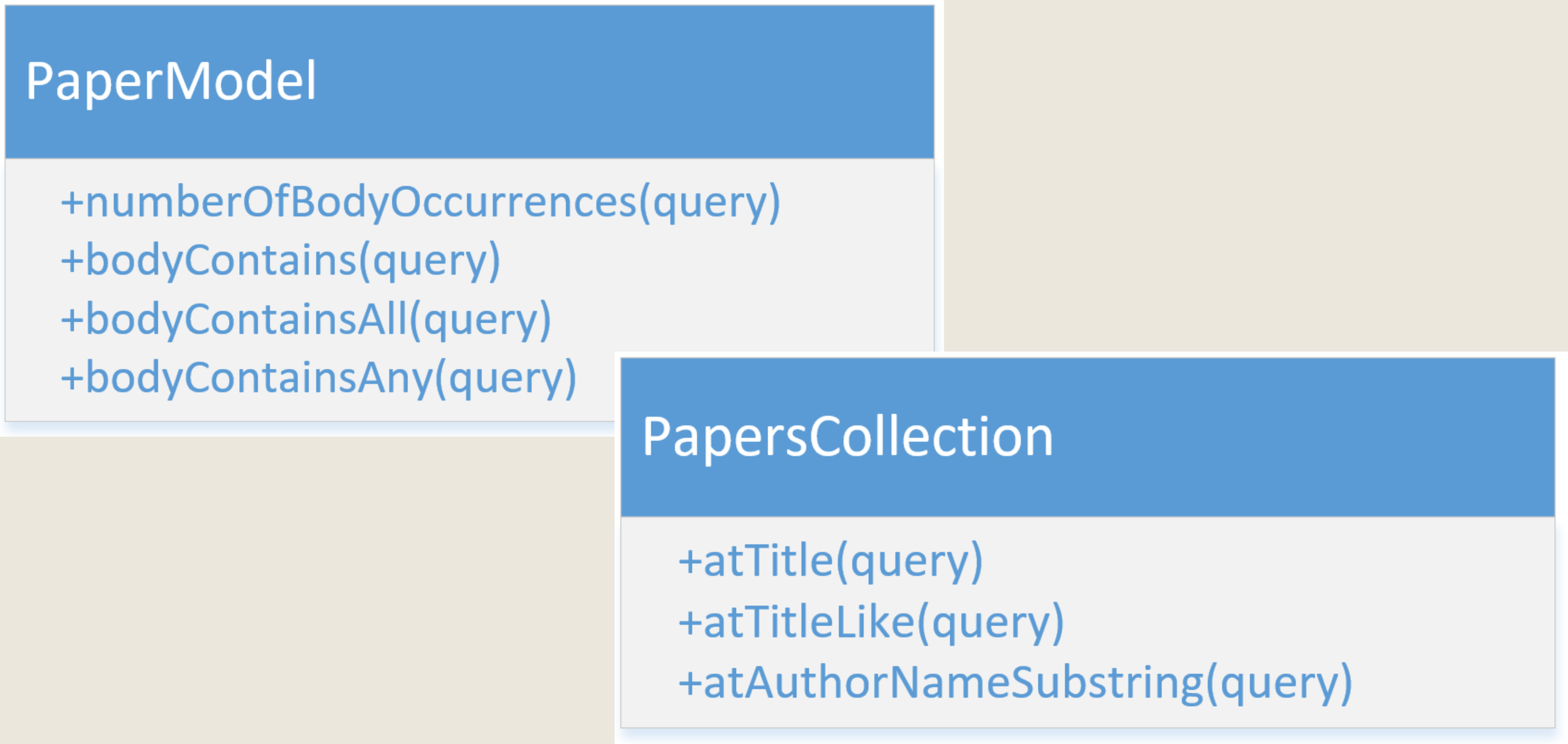
Conclusion

- 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



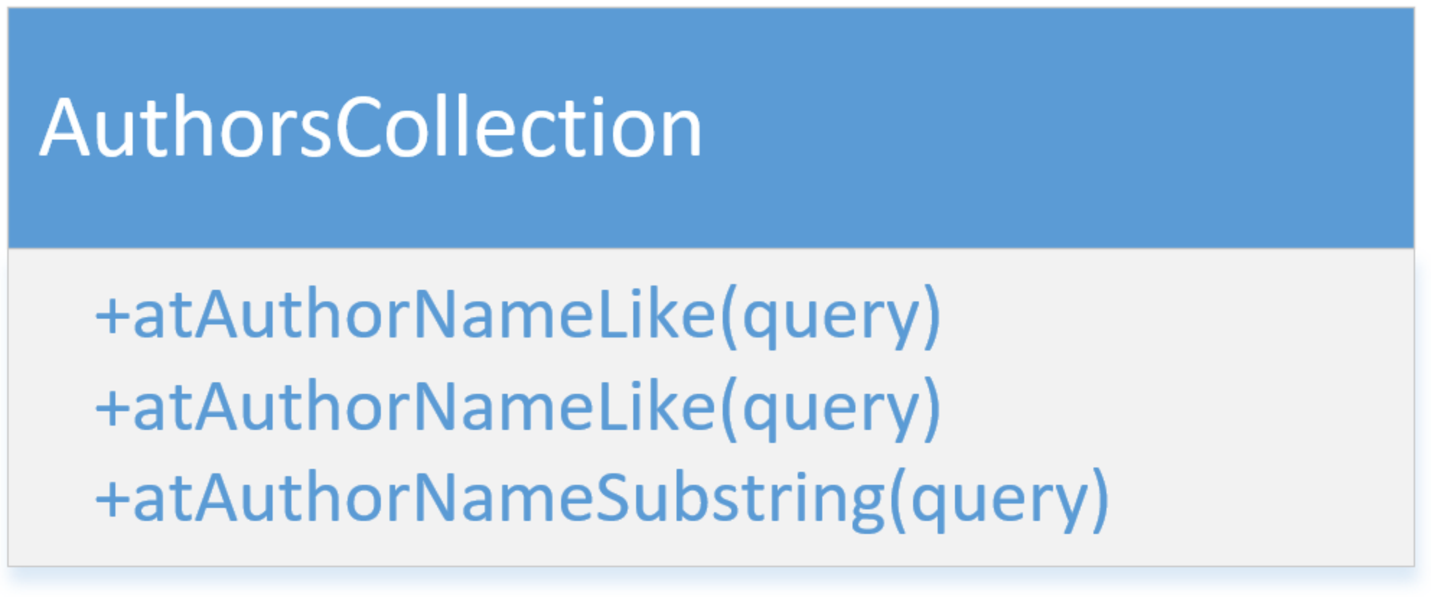
PaperModel

+numberOfBodyOccurrences(query)
+bodyContains(query)
+bodyContainsAll(query)
+bodyContainsAny(query)

PapersCollection

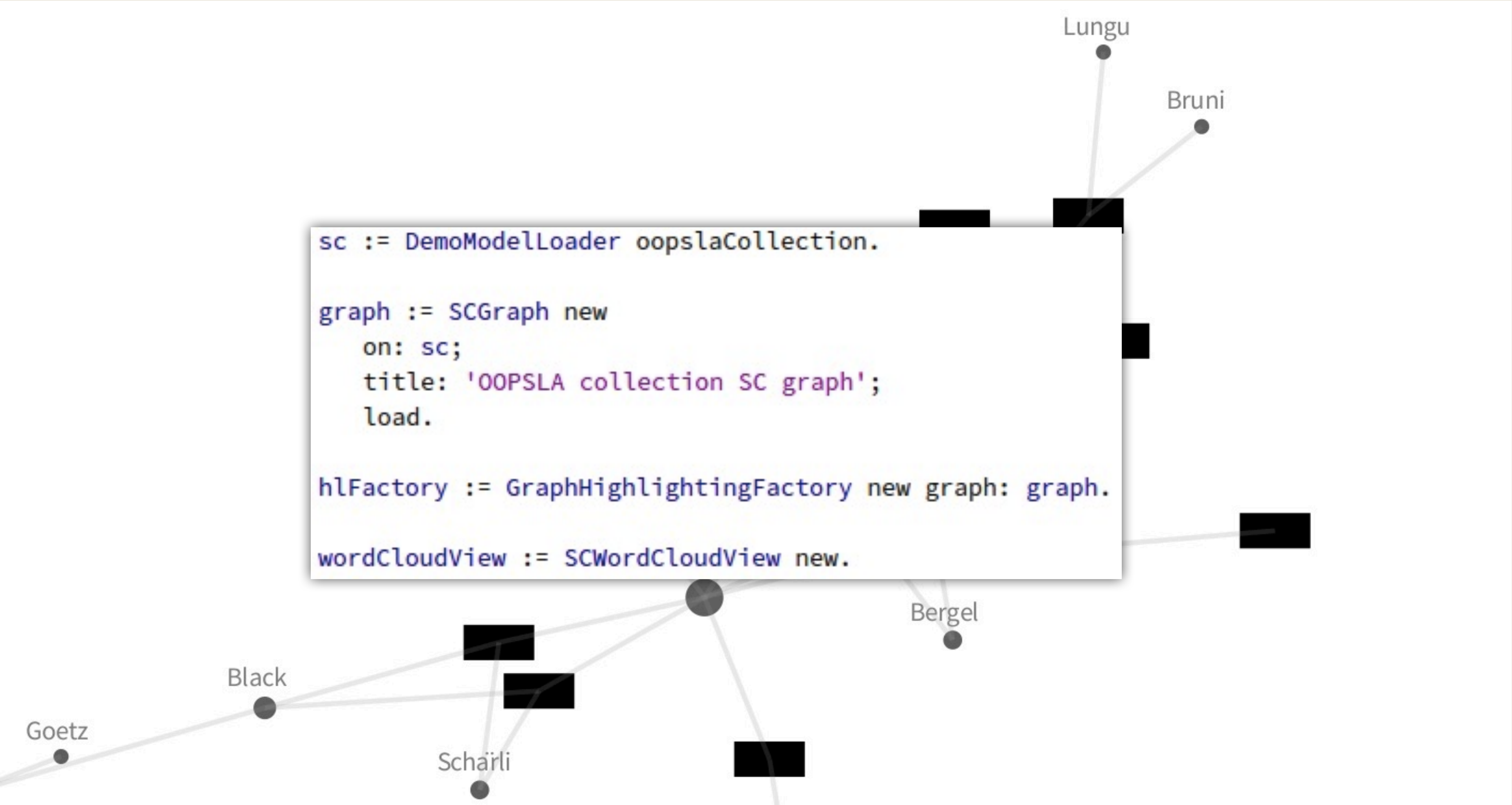
+atTitle(query)
+atTitleLike(query)
+atAuthorNameSubstring(query)

Some handy model query methods

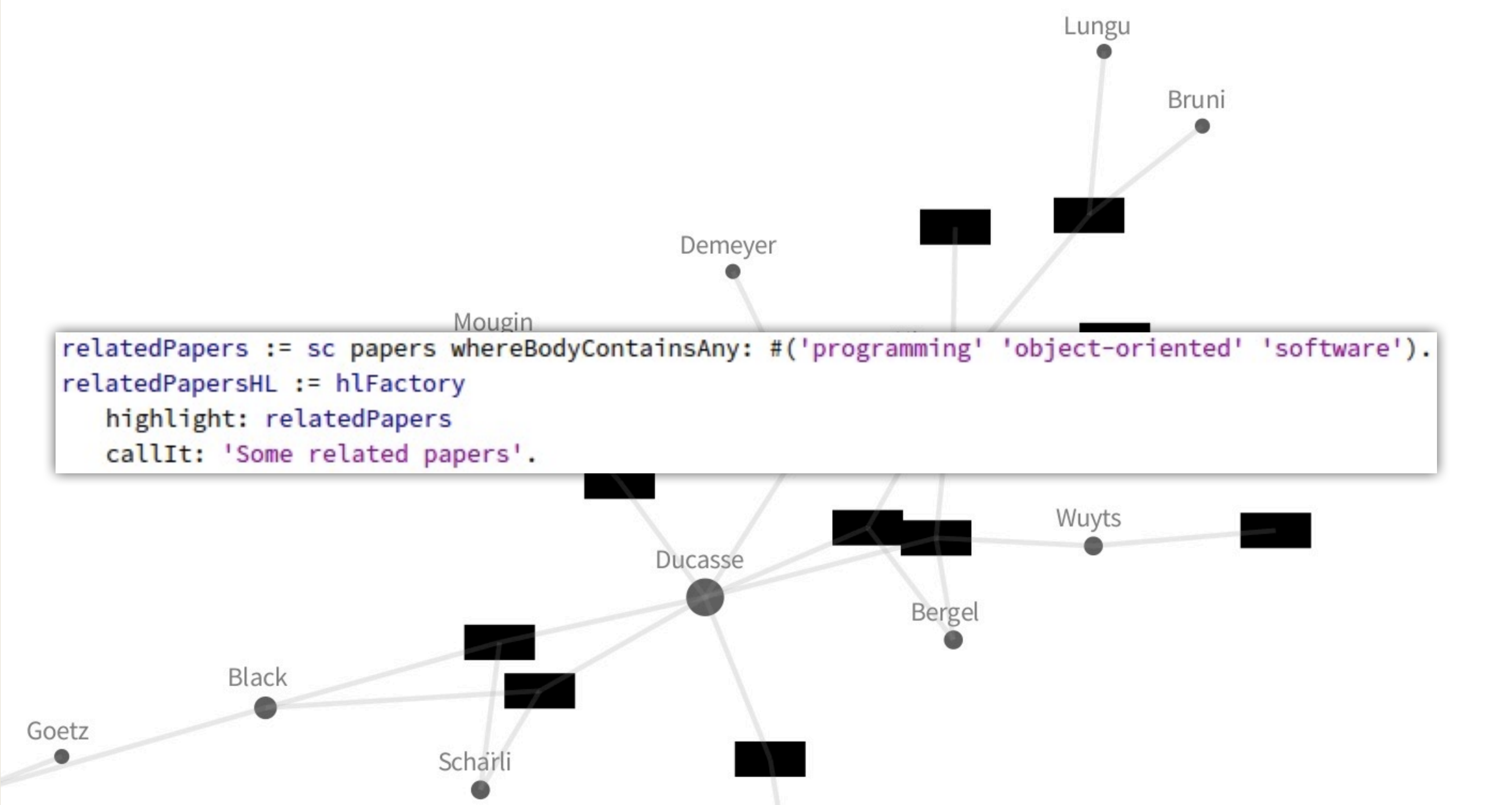


AuthorsCollection

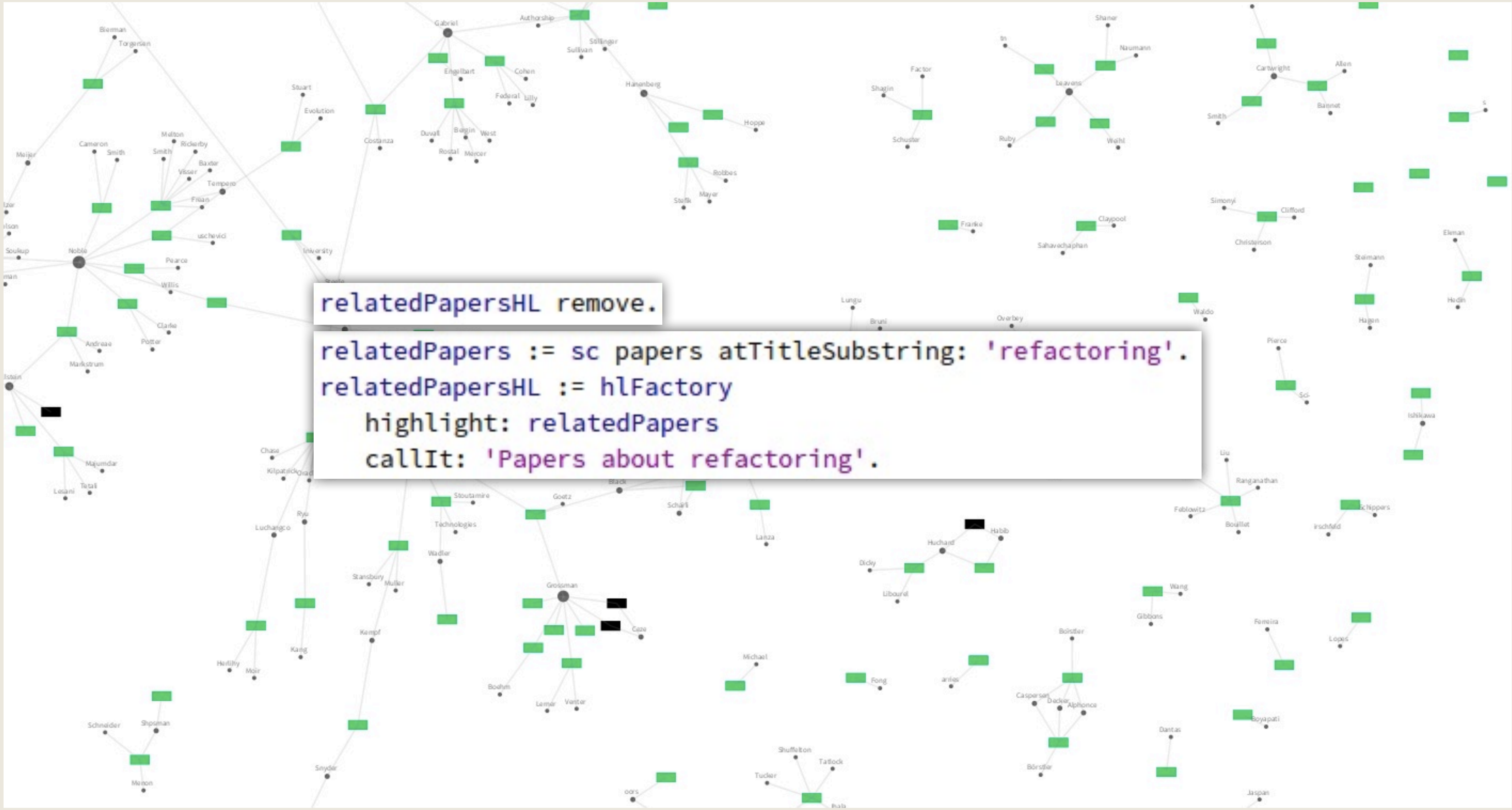
+atAuthorNameLike(query)
+atAuthorNameLike(query)
+atAuthorNameSubstring(query)



```
sc := DemoModelLoader oopslaCollection.

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

hlFactory := GraphHighlightingFactory new graph: graph.

wordCloudView := SCWordCloudView new.
```

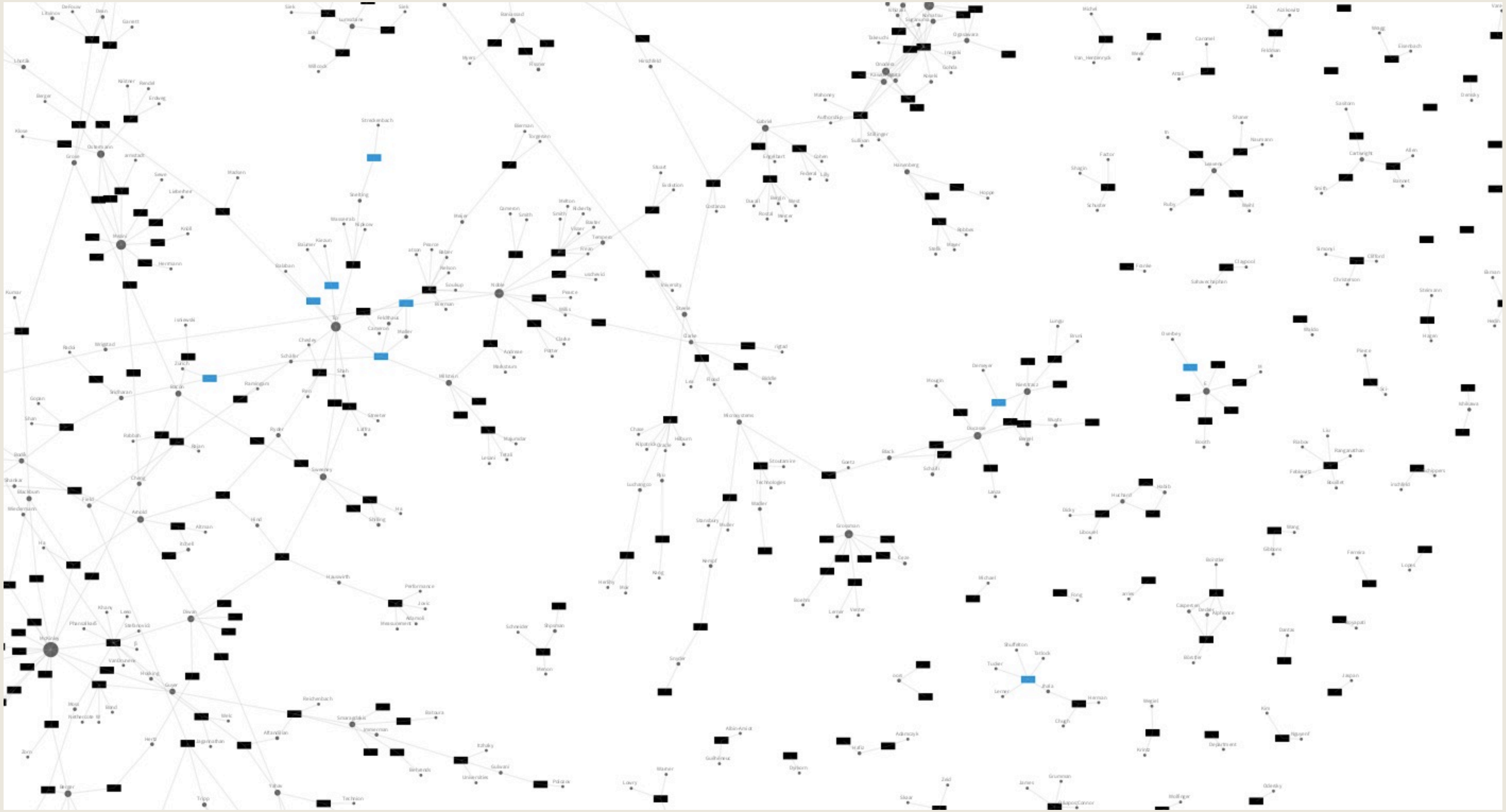


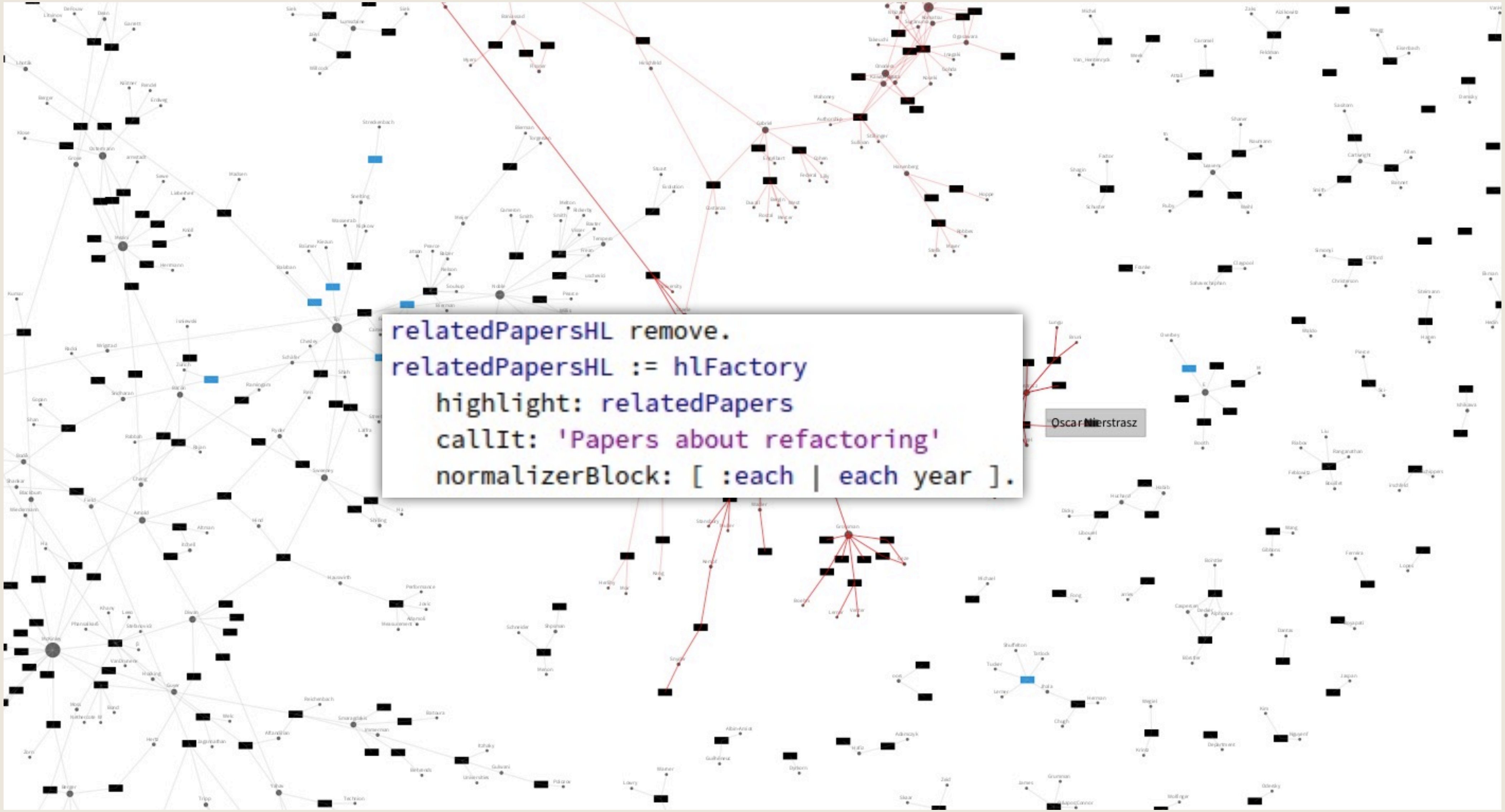
```
relatedPapers := sc papers whereBodyContainsAny: #('programming' 'object-oriented' 'software').
relatedPapersHL := hlFactory
   highlight: relatedPapers
   callIt: 'Some related papers'.
```



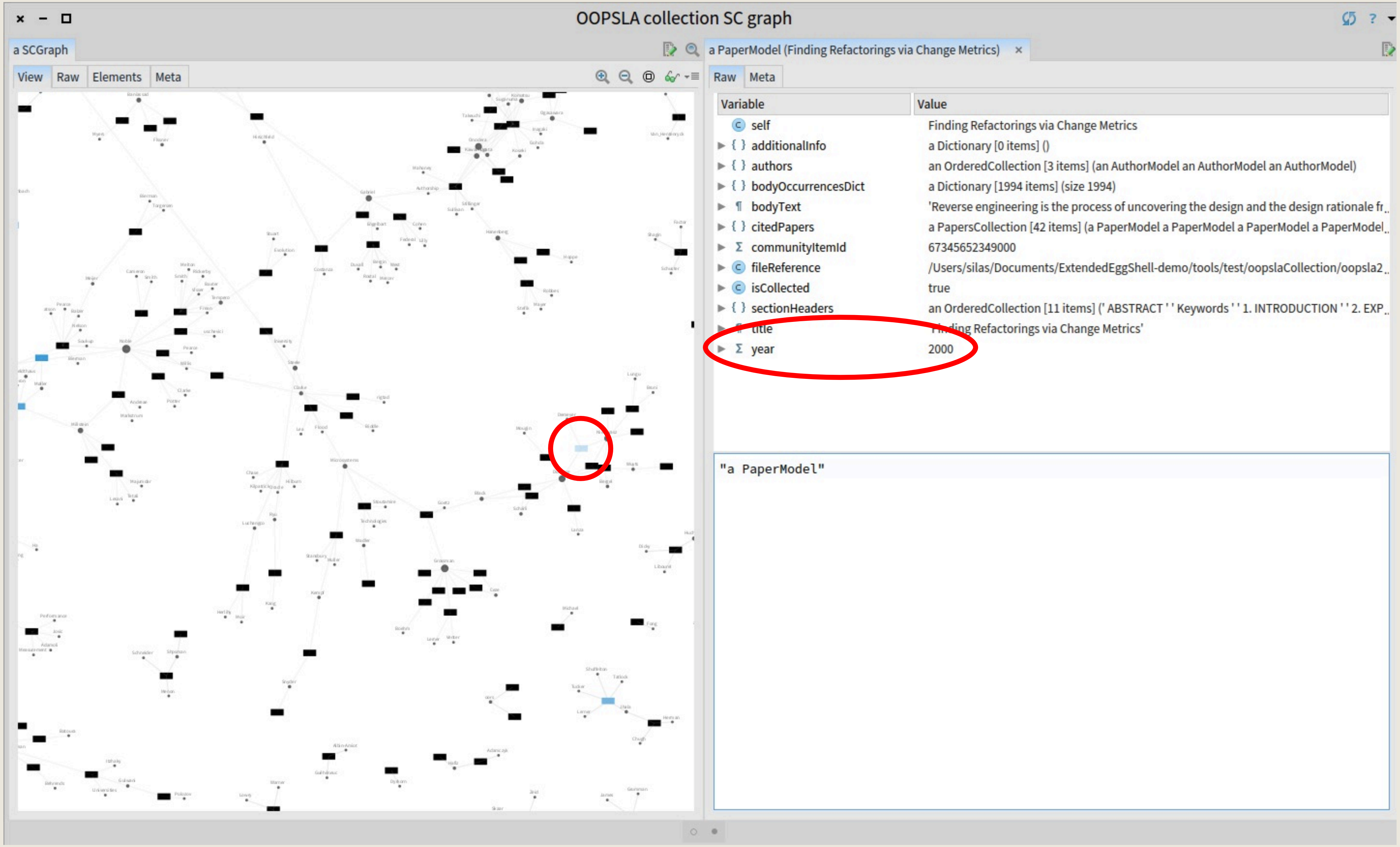
```
relatedPapersHL remove.
```

```
relatedPapers := sc papers atTitleSubstring: 'refactoring'.
relatedPapersHL := hlFactory
   highlight: relatedPapers
   callIt: 'Papers about refactoring'.
```


DeFouw
Litvinov
Dein
Garrett
Siek
Lumsdaine
Jalyi
Willcock
Bonassad
Myers
Lhoták
Kästner
Rendel
Berger
Erdweg
Klose
Ostermann
Grose
armstadt
Madsen
Sewe
Liebenher
Kröll
Mezini
Hermann
Kumar
Streckenbach
Snelting
Biemann
Wasserrab
Nipkow
Bolsner
Klesun
Balaban
Pearce
Nelson
Soukup
Noble
Cameron
Smith
Feldthaus
Gabriel
Mahoney
Stillinger
Sullivan
Heinesberg
Hoppe
Robbes
Steele
Clarke
Microsystems
Black
Grossman
Shaver
Neumann
Factor
Clifford
Christensen
Booth
Ranganathan
Performance
Reichenbach
Technion
Schneider
Snyder
Department
Fong
Jacques



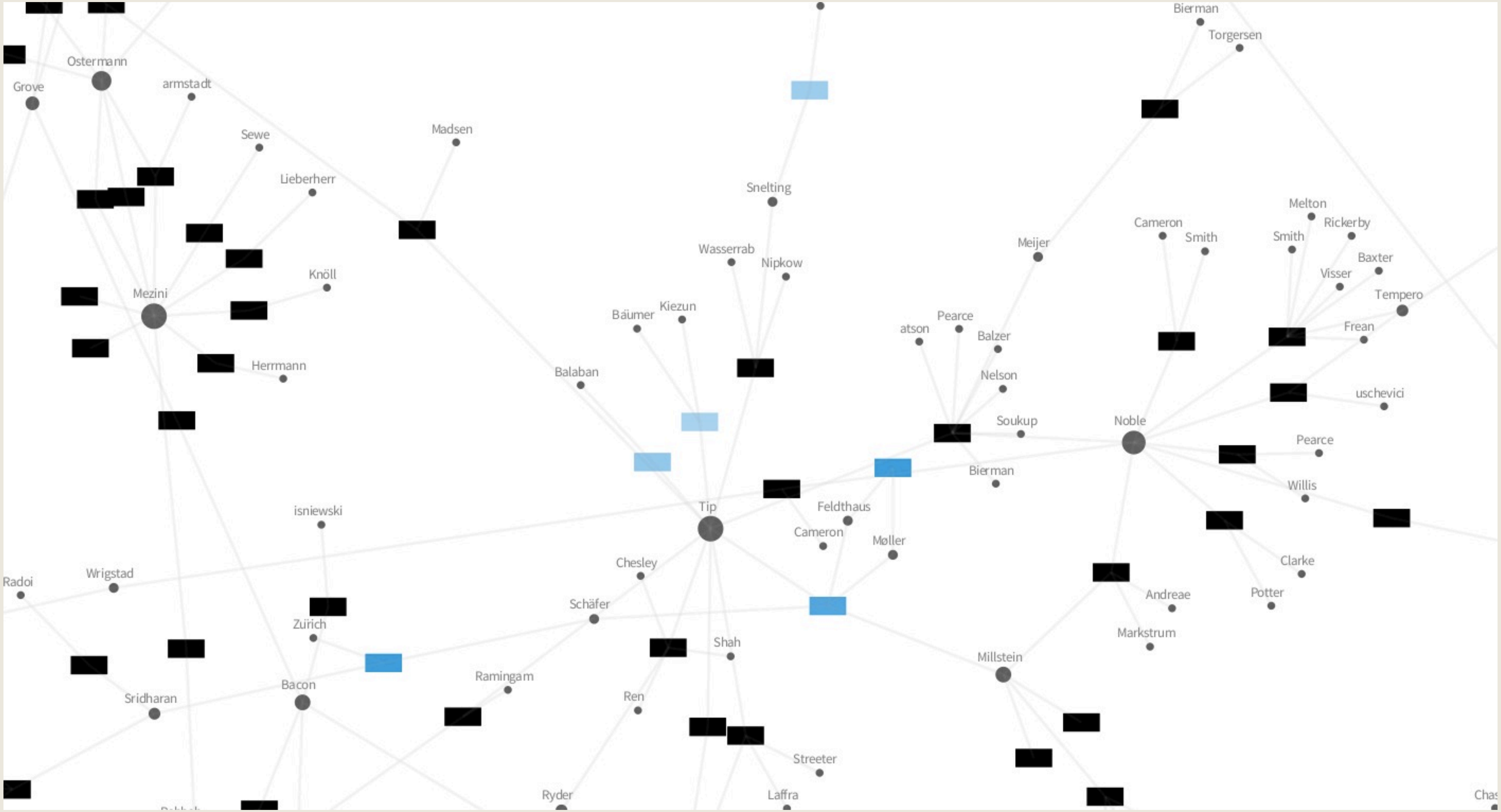
```
relatedPapersHL remove.
relatedPapersHL := hlFactory
    highlight: relatedPapers
    callIt: 'Papers about refactoring'
    normalizerBlock: [ :each | each year ].
```

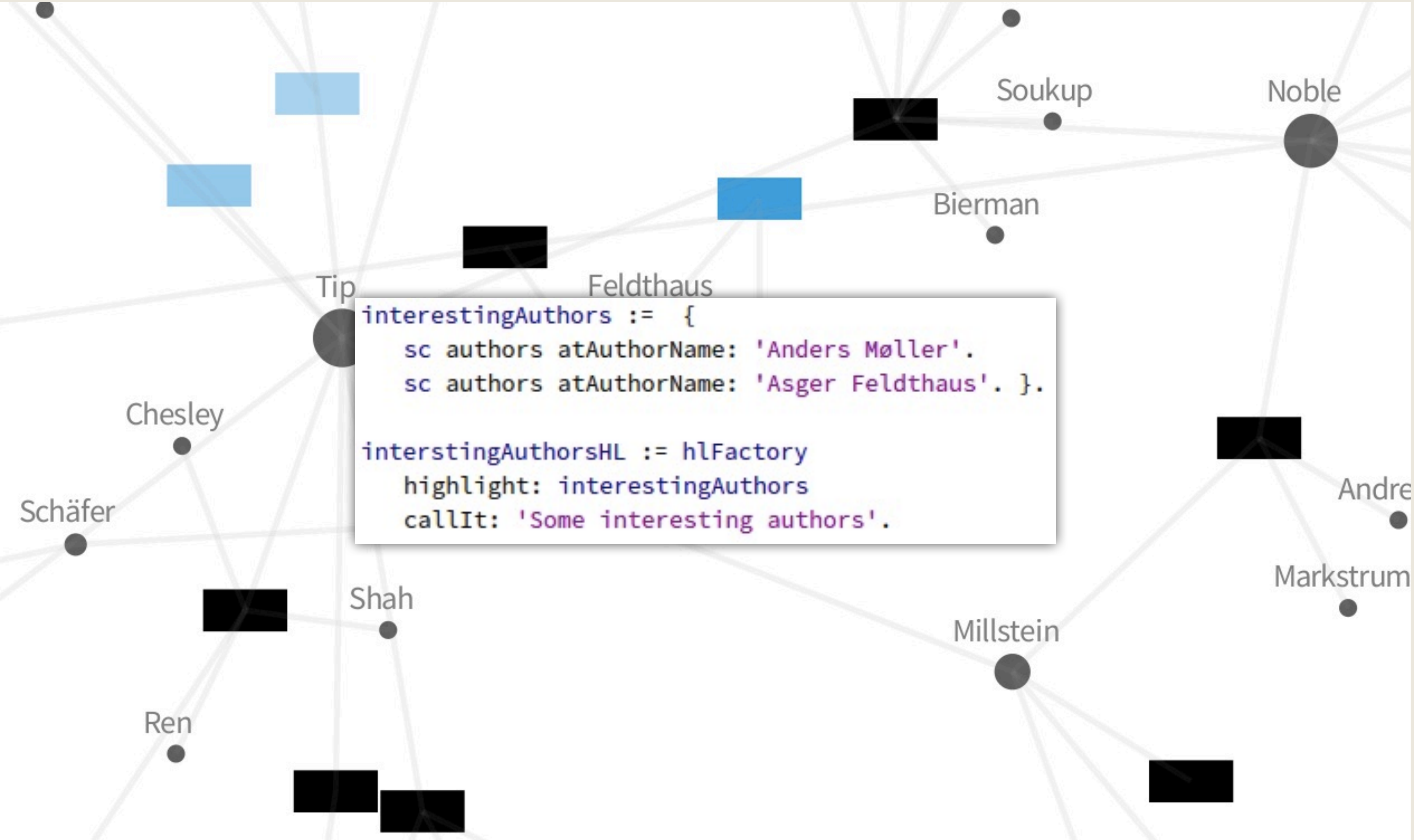


OOPSLA collection SC graph
a SCGraph
View
Raw
Elements
Meta
a PaperModel (Finding Refactorings via Change Metrics)
Raw
Meta

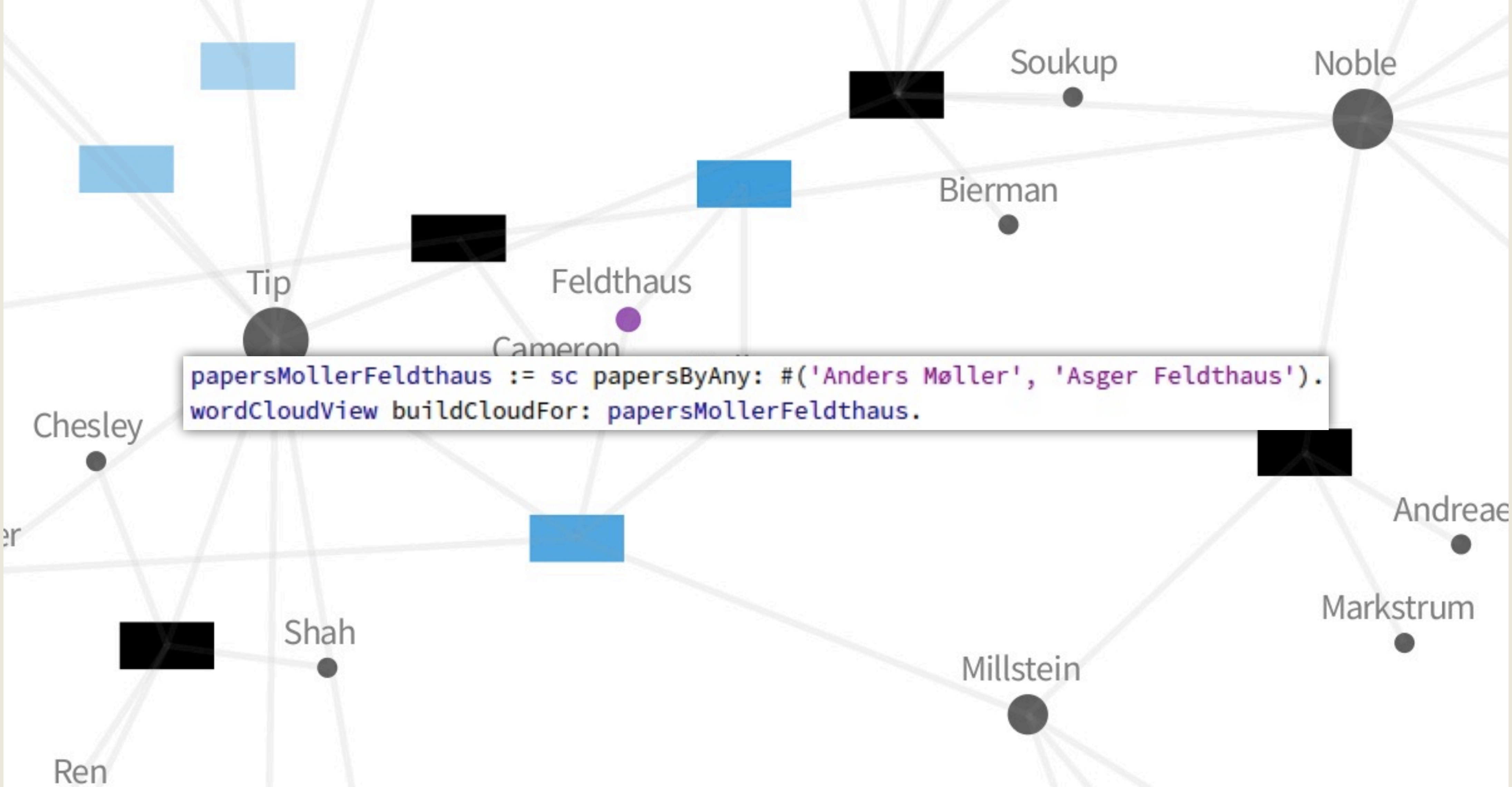
Variable	Value
self	Finding Refactorings via Change Metrics
additionalInfo	a Dictionary [0 items] ()
authors	an OrderedCollection [3 items] (an AuthorModel an AuthorModel an AuthorModel)
bodyOccurrencesDict	a Dictionary [1994 items] (size 1994)
bodyText	'Reverse engineering is the process of uncovering the design and the design rationale fr…
citedPapers	a PapersCollection [42 items] (a PaperModel a PaperModel a PaperModel a PaperModel…
communityItemId	67345652349000
fileReference	/Users/silas/Documents/ExtendedEggShell-demo/tools/test/oopslaCollection/oopsla2…
isCollected	true
sectionHeaders	an OrderedCollection [11 items] (' ABSTRACT ' ' Keywords ' ' 1. INTRODUCTION ' ' 2. EXP…
title	'Finding Refactorings via Change Metrics'
year	2000

```
"a PaperModel"
```

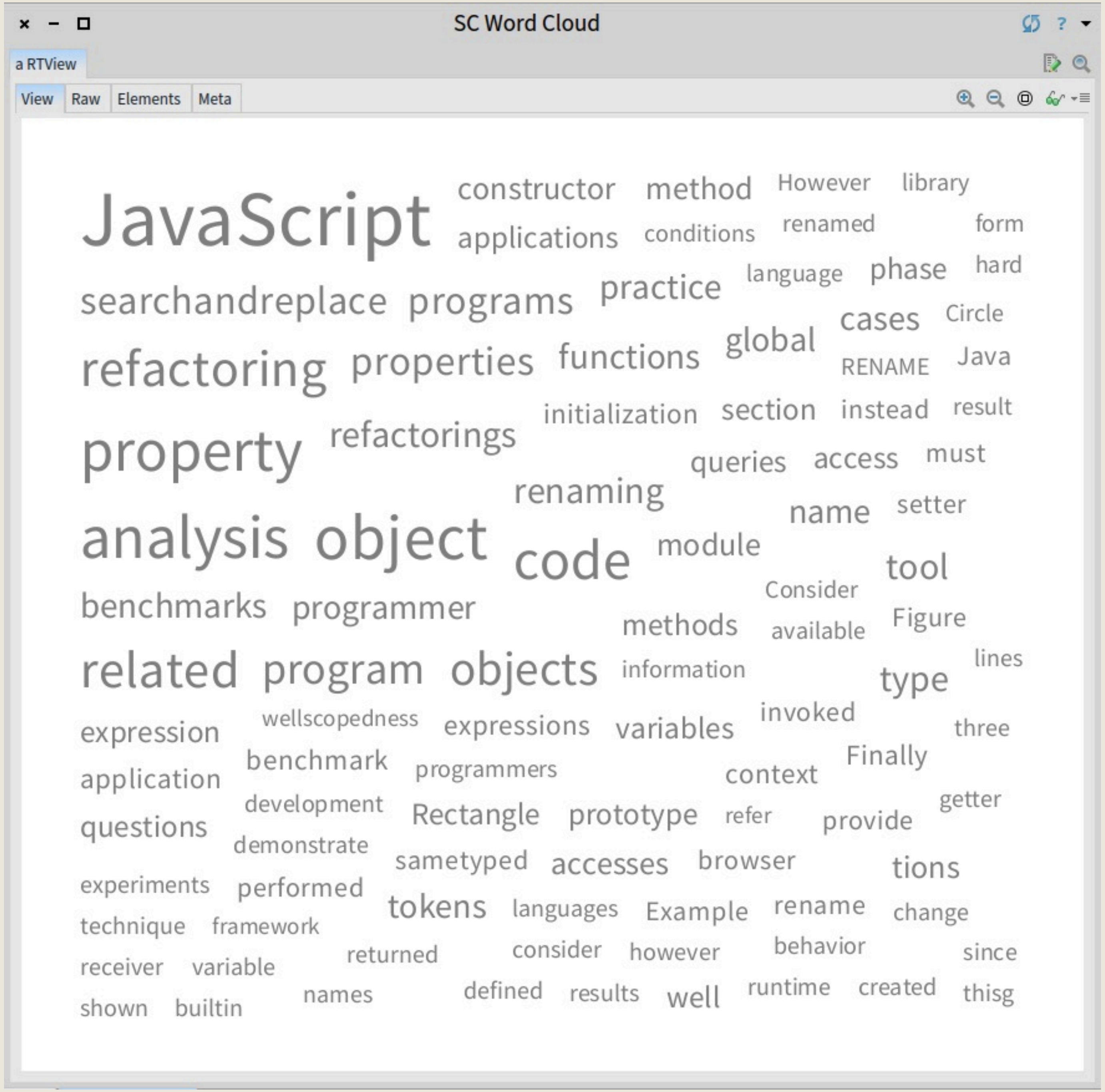

Bierman
Torgersen
Ostermann
Grove
armstadt
Sewe
Madsen
Lieberherr
Snelting
Melton
Cameron
Rickerby
Smith
Smith
Meijer
Wasserrab
Nipkow
Baxter
Visser
Tempero
Knöll
Mezini
Kiezun
Bäumer
Pearce
atson
Balzer
Frean
Herrmann
Balaban
Nelson
uschevici
Soukup
Noble
Pearce
Bierman
Willis
Tip
Feldthaus
isniewski
Cameron
Møller
Clarke
Chesley
Radoi
Wrigstad
Andreae
Potter
Schäfer
Zürich
Markstrum
Shah
Millstein
Ramingam
Bacon
Sridharan
Ren
Streeter
Ryder
Laffra
Chas

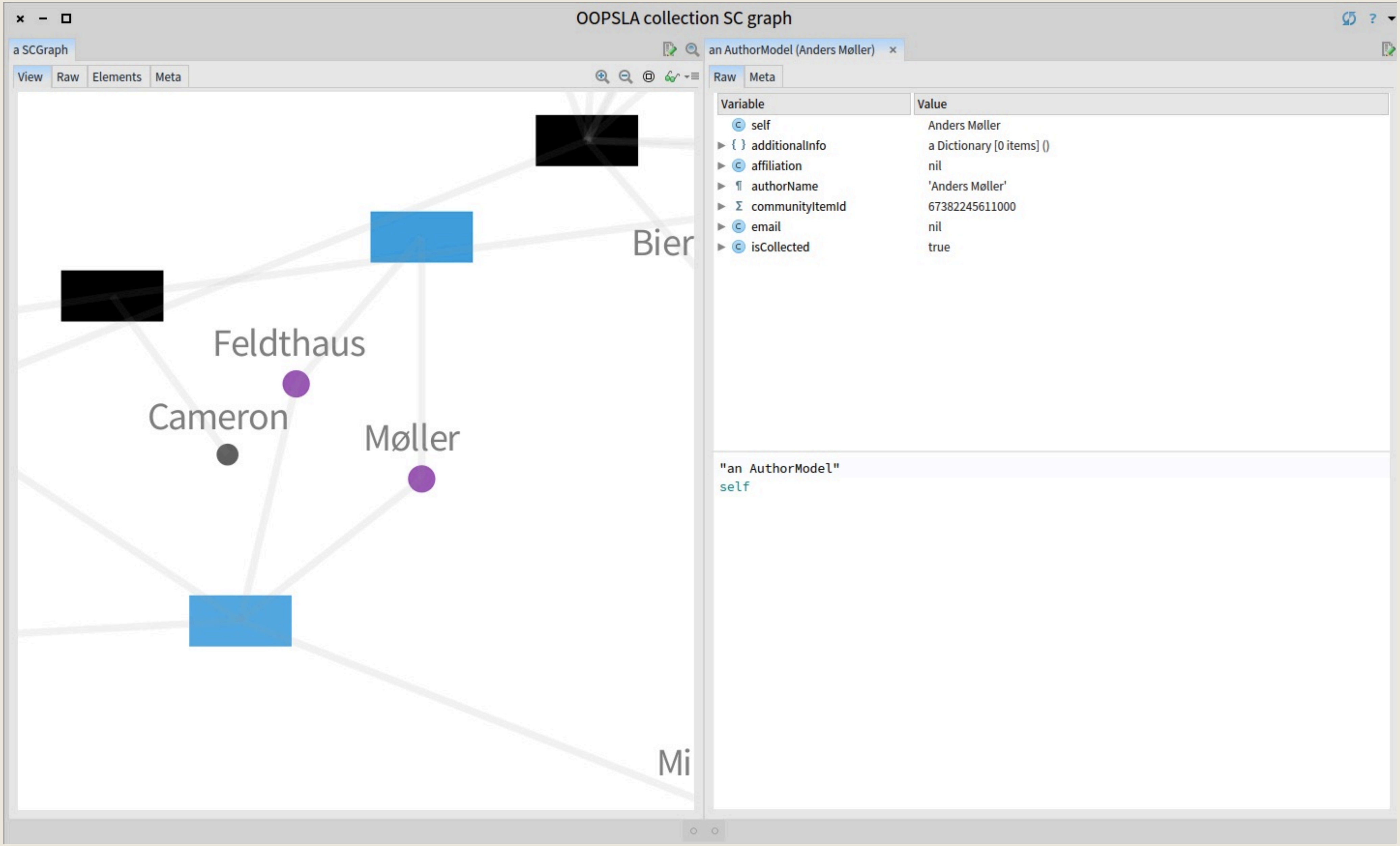


```
interestingAuthors :=  {
   sc authors atAuthorName: 'Anders Møller'.
   sc authors atAuthorName: 'Asger Feldthaus'. }.

interstingAuthorsHL := hlFactory
   highlight: interestingAuthors
   callIt: 'Some interesting authors'.
```



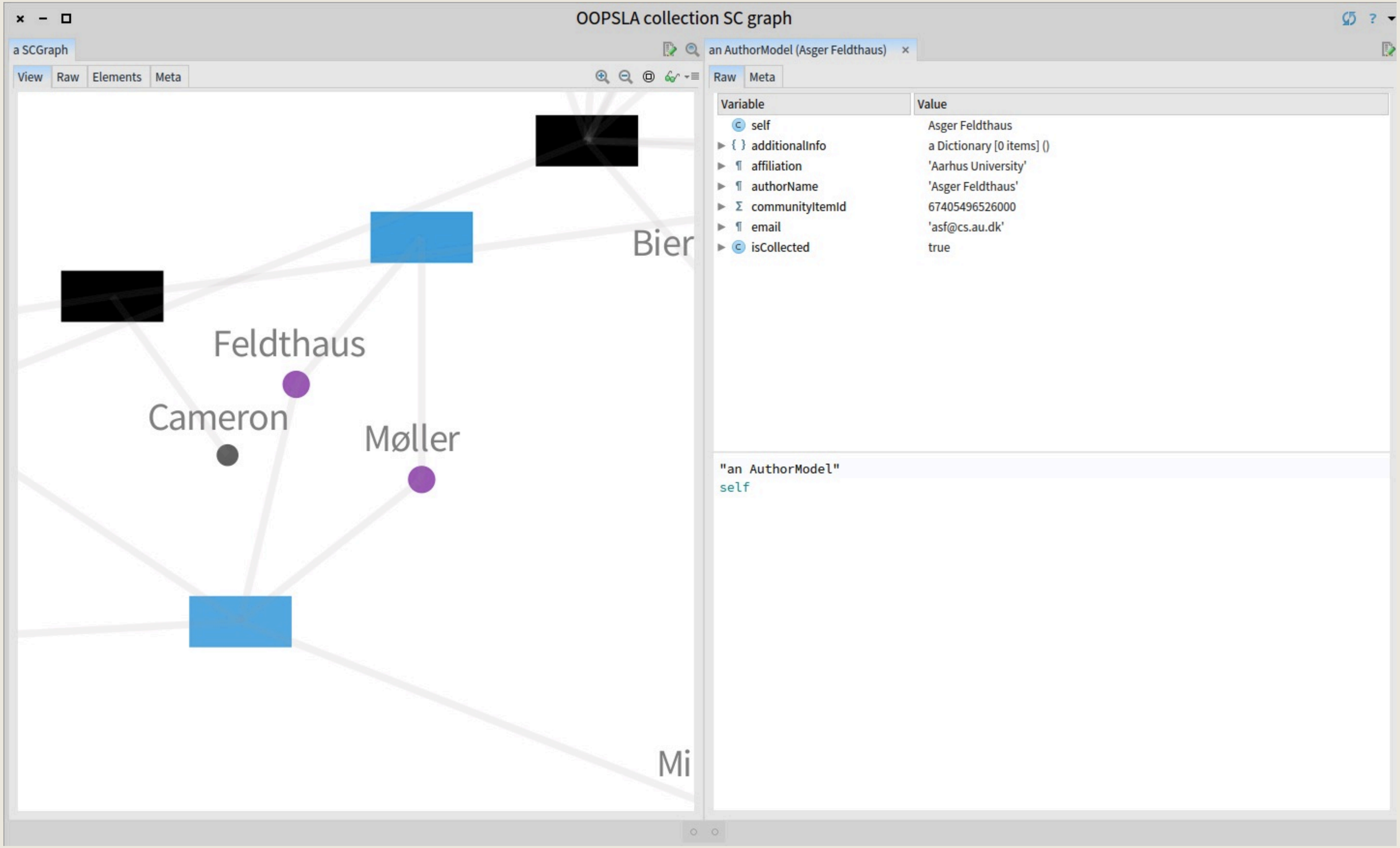
```
papersMollerFeldthaus := sc papersByAny: #('Anders Møller', 'Asger Feldthaus').
wordCloudView buildCloudFor: papersMollerFeldthaus.
```


SC Word Cloud
a RTView
View
Raw
Elements
Meta
JavaScript
constructor
method
However
library
applications
conditions
renamed
form
searchandreplace
programs
practice
language
phase
hard
refactoring
properties
functions
global
cases
Circle
RENAME
Java
initialization
section
instead
result
property
refactorings
queries
access
must
renaming
name
setter
analysis
object
code
module
tool
Consider
benchmarks
programmer
methods
available
Figure
related
program
objects
information
type
lines
expression
wellscopedness
expressions
variables
invoked
three
application
benchmark
programmers
context
Finally
questions
development
Rectangle
prototype
refer
provide
getter
demonstrate
sametyped
accesses
browser
tions
experiments
performed
tokens
languages
Example
rename
change
technique
framework
receiver
variable
returned
consider
however
behavior
since
names
defined
results
well
runtime
created
thisg
shown
builtin



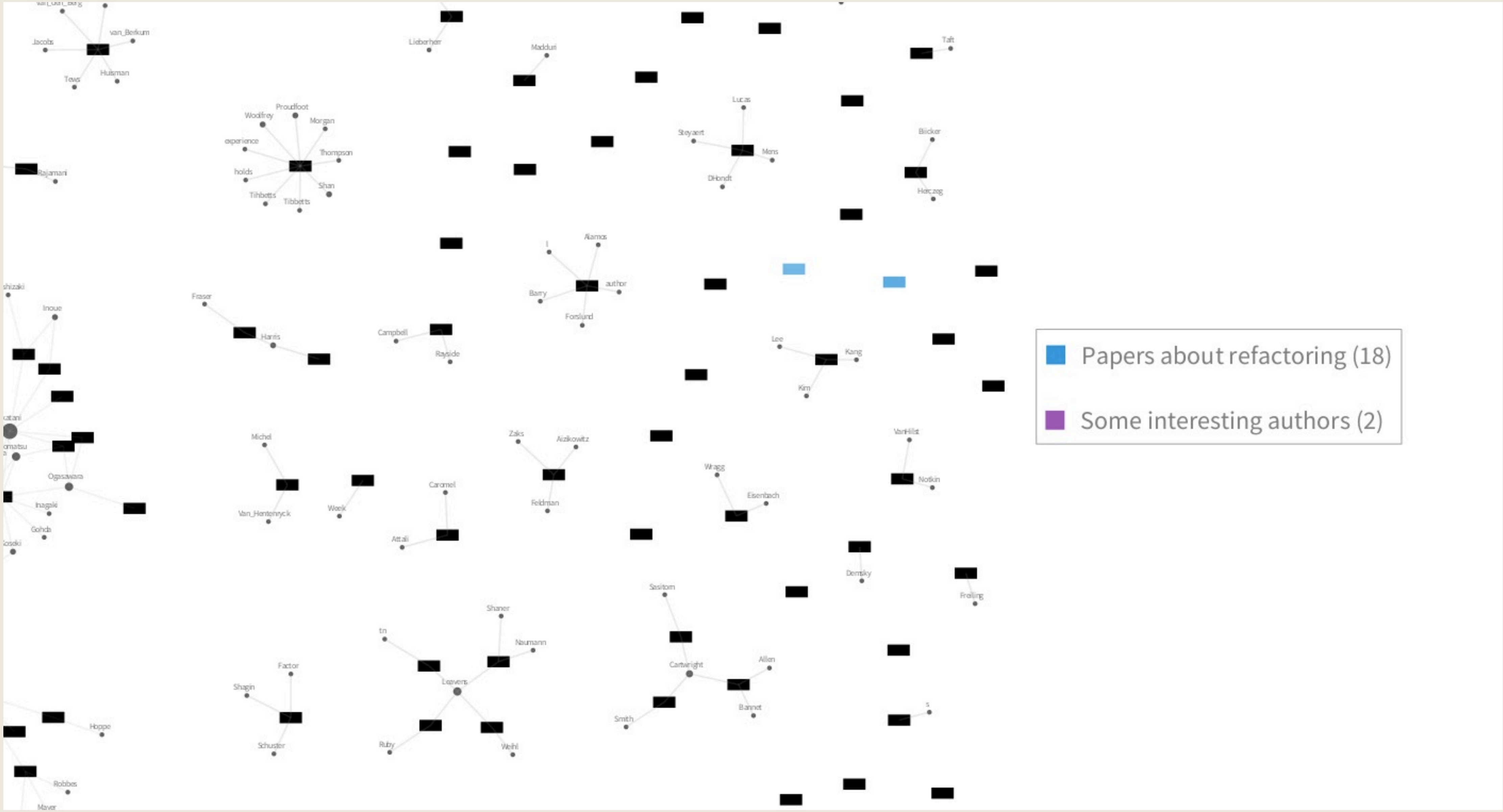
Variable	Value
self	Anders Møller
additionalInfo	a Dictionary [0 items] ()
affiliation	nil
authorName	'Anders Møller'
communityItemId	67382245611000
email	nil
isCollected	true

```
"an AuthorModel"
self
```



Variable	Value
self	Asger Feldthaus
additionalInfo	a Dictionary [0 items] ()
affiliation	'Aarhus University'
authorName	'Asger Feldthaus'
communityItemId	67405496526000
email	'asf@cs.au.dk'
isCollected	true

```
"an AuthorModel"
self
```


Jacobs
van_Berkum
Tews
Huisman
Lieberherr
Madduri
Taft
Lucas
Steyaert
Mens
DHondt
Bicker
Herczeg
Proudfoot
Woolfrey
Morgan
experience
Thompson
holds
Shan
Tihbetts
Tibbetts
Rajamani
Alamos
author
Barry
Forslund
shizaki
Inoue
Fraser
Harris
Campbell
Rayside
Lee
Kang
Kim
Papers about refactoring (18)
Some interesting authors (2)
Michel
Zaks
Aizikowitz
Feldman
VanHilst
Notkin
Ogasawara
Inagaki
Gohda
Van_Hentenryck
Week
Caromel
Attali
Wragg
Eisenbach
Demsky
Frolling
Sasitorn
Shaner
Naumann
tn
Factor
Shagin
Schuster
Leavens
Ruby
Weihl
Cartwright
Allen
Smith
Bannet
s
Hoppe
Robbes
Mayer