



LUCENE.NET

THE RESURRECTION

Itamar Syn-Hershko
itamar@code972.com
@synhershko

WHAT IS LUCENE?

Op art 31, 179, 180, 192, 211
Orphism 13-14, 15-17, 33-6, 142
Oud, J. J. P. 84-5; 53
Ouspensky, Peter Demianovich 56-8
Ozenfant, Amédée 71

Palermo, Blinky 224-5
papiers collés 14, 64
Pasmore, Victor 136, 174
Pattern and Decoration (P&D) group
224

Pevsner, Antoine 103, 104, 109, 120
phenomenology 128
Picabia, Francis 13, 33, 36, 64, 142,
144; 21

Picasso, Pablo 11, 12, 13, 14-15,
72-3, 119, 146, 149, 161, 170, 197,
230; 2

Piène, Otto 180

Plastique 142

Poggioli, Renato 182

Poliakoff, Serge 125-6

Pollock, Jackson 128, 130, 135, 141,
144, 146, 148-54, 159, 163, 170,
173, 191, 192, 197; 97, 103, 107

Poons, Larry 192; 130

Pop art 173, 189, 197, 199, 212, 219

Popova, Lyubov 76, 80, 81; 48

Post-Modernism 219, 226

Rosenberg
Rosicrucia
Roszak, T
Rothko, M

113

Rozanova
Rubin, W
Russell, M
Ryman, I

St Ives 1

Salon des

Schlemm

Schmidt

Schoenb

Schoenn

Schopen

Schneid

Schultze

Schuma

Schwitt

104, 1

Scott, T

Scott, V

Serra, F

Seupho

Severin


Signac,

'Situati

THE INVERTED INDEX

User queries for “keeper”

1	The old night keeper keeps the keep in the town
2	In the big old house in the big old gown.
3	The house in the town had the big old keep
4	Where the old night keeper never did sleep.
5	The night keeper keeps the keep in the night
6	And keeps in the dark and sleeps in the light.

 6 documents to index

 The index:

Dictionary and
posting lists

Term	Documents
and	<6>
big	<2> <3>
dark	<6>
did	<4>
gown	<2>
had	<3>
house	<2> <3>
in	<1> <2> <3> <5> <6>
keep	<1> <3> <5>
keeper	<1> <4> <5>
keeps	<1> <5> <6>
light	<6>
never	<4>
night	<1> <4> <5>
old	<1> <2> <3> <4>
sleep	<4>
sleeps	<6>
the	<1> <2> <3> <4> <5> <6>
town	<1> <3>
where	<4>

TERM NORMALIZATION

- Lowercasing
- Stop words (grey)
 - Not best practice anymore
- Stemming
 - Porter stemmer
 - s-stemmer
- Relevance++
- SizeOnDisk--

Term	Documents
and	<6>
big	<2> <3>
dark	<6>
did	<4>
gown	<2>
had	<3>
house	<2> <3>
in	<1> <2> <3> <5> <6>
keep	<1> <3> <5>
keeper	<1> <4> <5>
keeps	<1> <5> <6>
light	<6>
never	<4>
night	<1> <4> <5>
old	<1> <2> <3> <4>
sleep	<4>
sleeps	<6>
the	<1> <2> <3> <4> <5> <6>
town	<1> <3>
where	<4>

NOT JUST SIMPLE SEARCH

Full-text search

Languages support

Fuzzy & regex search

Geo-spatial search

Search as you type functionality

Faceting

Record linkage

Classification, NLP



FacebookList input

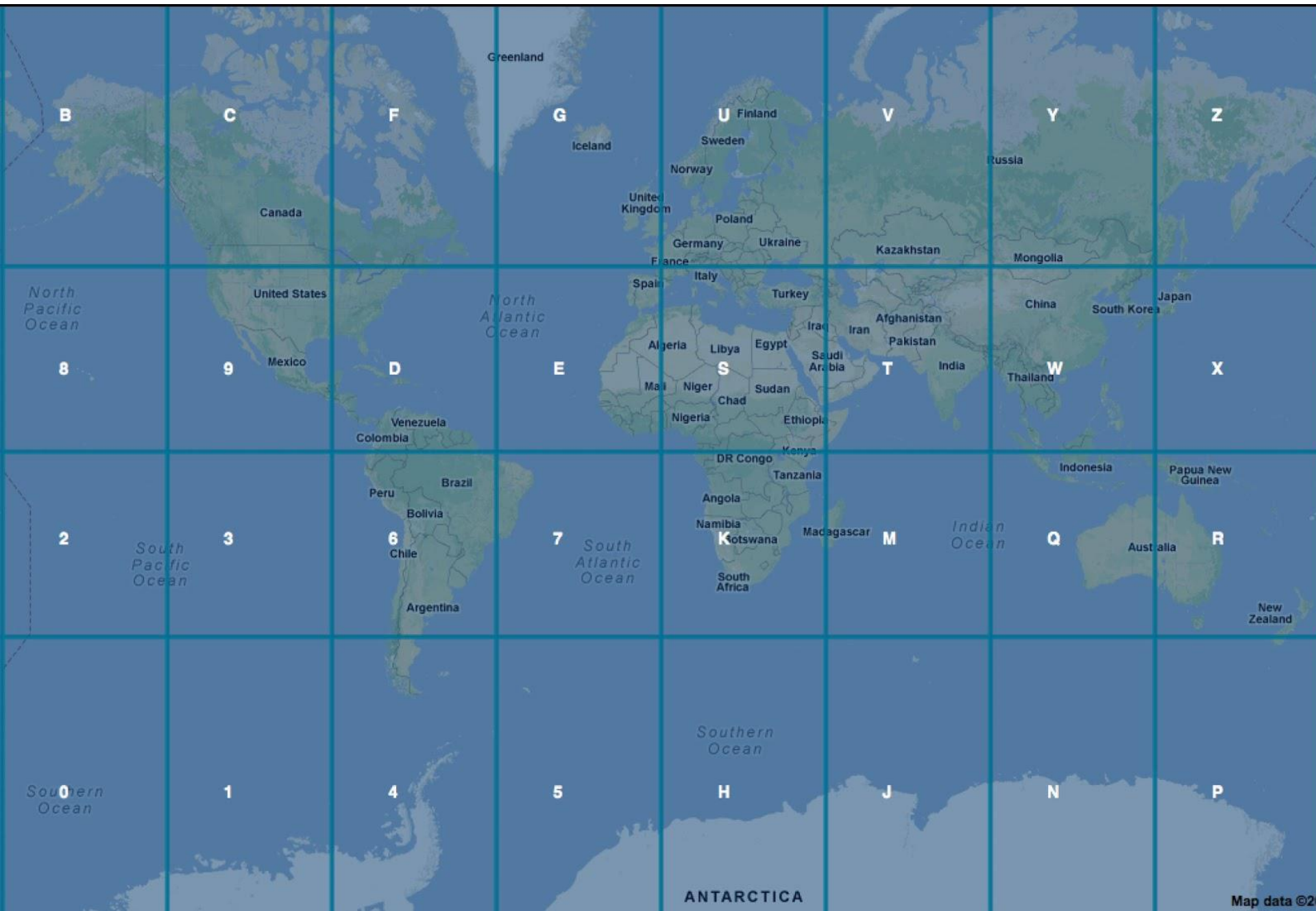
Jorge Luis Borges x Julio Cortazar x ma

Manuel Mujica Lainez

Hector German Oesterheld

Manuel Puig

Mario Rodriguez Cobos

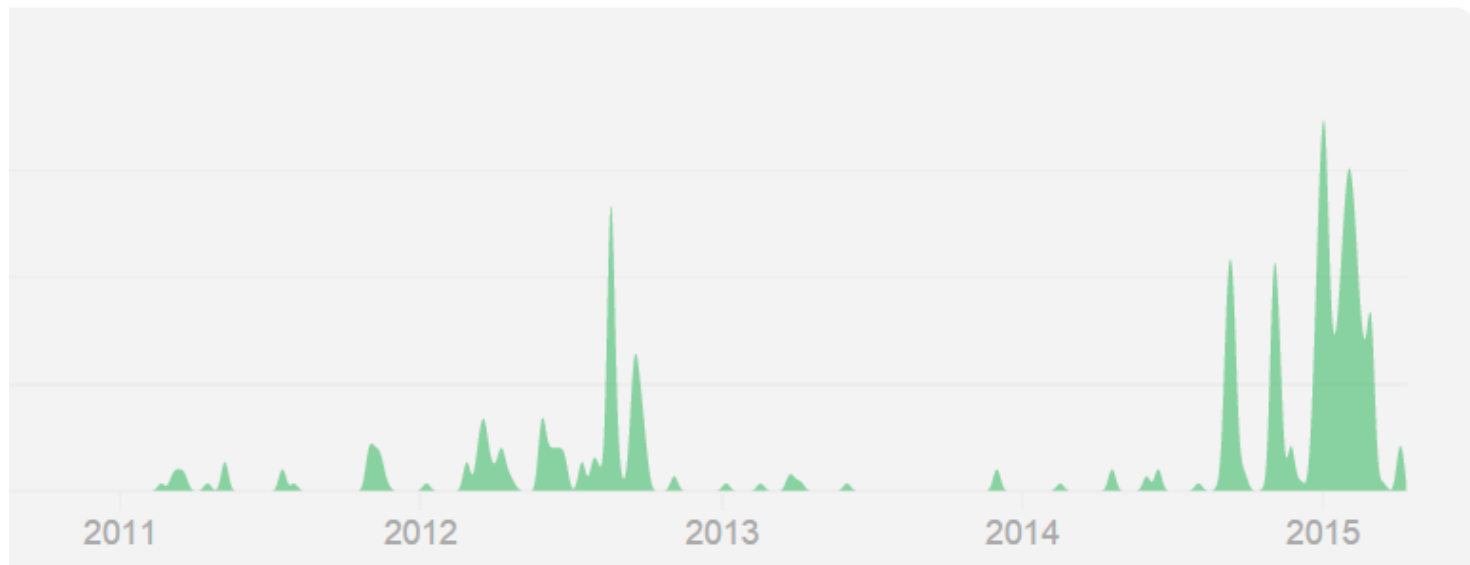


APACHE LUCENE.NET

Apache Software License

Lucene.Net took an estimated 296 years of effort (COCOMO model) starting with its first commit in November, 2005

Java Lucene took an estimated 136 years of effort (COCOMO model) starting with its first commit in March, 2000



SOME CODE

IndexWriter, Documents and Fields

IndexReader, IndexSearcher

The analysis chain

Filters, CachingWrapperFilter

NRT, reopening, SearcherManager

TermVectors, positions, highlighting and all that

APACHE LUCENE.NET

Over 700,000 downloads from nuget

143,000 for version 3.0.3

... which was released October 26, 2012 (!)

Being used by many. You probably used it today.

RavenDB, Orchard, SubText, Umbraco, Nuget, Chocolatey, Microsoft, Autodesk, BBC, ...

CURRENT STATE OF AFFAIRS

Many users

Big, recent, code contribution from Microsoft

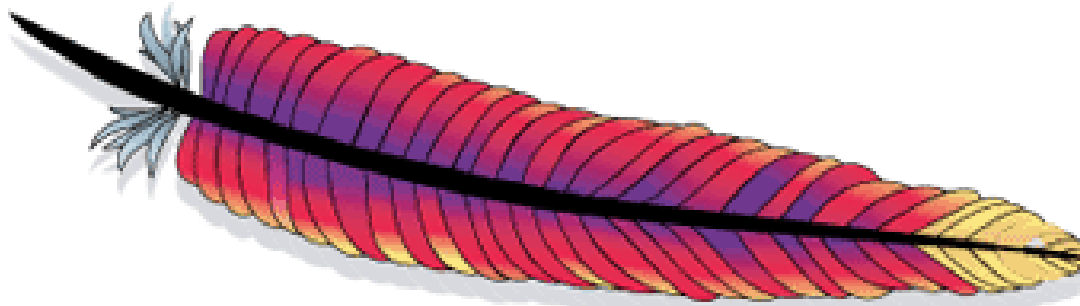
Not sponsored in any way

~30 failing tests in core

About 8 sub-projects still pending porting

~5 active contributors

THE APACHE WAY



Apache

WE NEED A BIGGER COMMUNITY

OPEN SOURCE DEVELOPERS



What my friends think I do



What my spouse thinks I do



What the users think I should do



What enterprise companies think I do



What I think I do



What I actually do

WE NEED YOUR HELP!

Already a user? Join the conversation

Take current master for a spin

Help with fixing isolated failing tests = see
<http://markmail.org/message/sl2ipkbefub2cz6p>

Help with some porting work

Not a user? Tell us what resources are missing

MEANWHILE IN JAVA LUCENE...

Analysis (Snowball, many languages, ngrams, synonyms, ICU, phonetic,...)

Codecs

Facets are now a core functionality

Several highlighter implementations

MemoryIndex

Improved QueryParsers

Spatial search

Automata Suggesters

Lucene Expressions

State of the art testing framework, randomized testing

PLANS FOR THE FUTURE

Stable code base that is easy to continue porting with

Core + sub-projects

.NET API refactoring

Thorough testing, including randomized testing. Green over time

Performance testing

.NET specific development (CLR / Windows specific optimizations, Azure, PCL, async / await)

Strong naming

Tutorials and guidance

Tools to help further porting and keeping up to date

GETTING INVOLVED



<http://lucenenet.apache.org/>

<https://github.com/apache/lucenenet/>

List of isolated tests: <http://markmail.org/message/sl2ipkbefub2cz6p>



THANKS

Itamar Syn-Hershko
itamar@code972.com
@synhershko