ICIJ’s leak refinery

Mar Cabra - Data editor
International Consortium of Investigative Journalists (ICIJ)

25 years The Center for Public Integrity

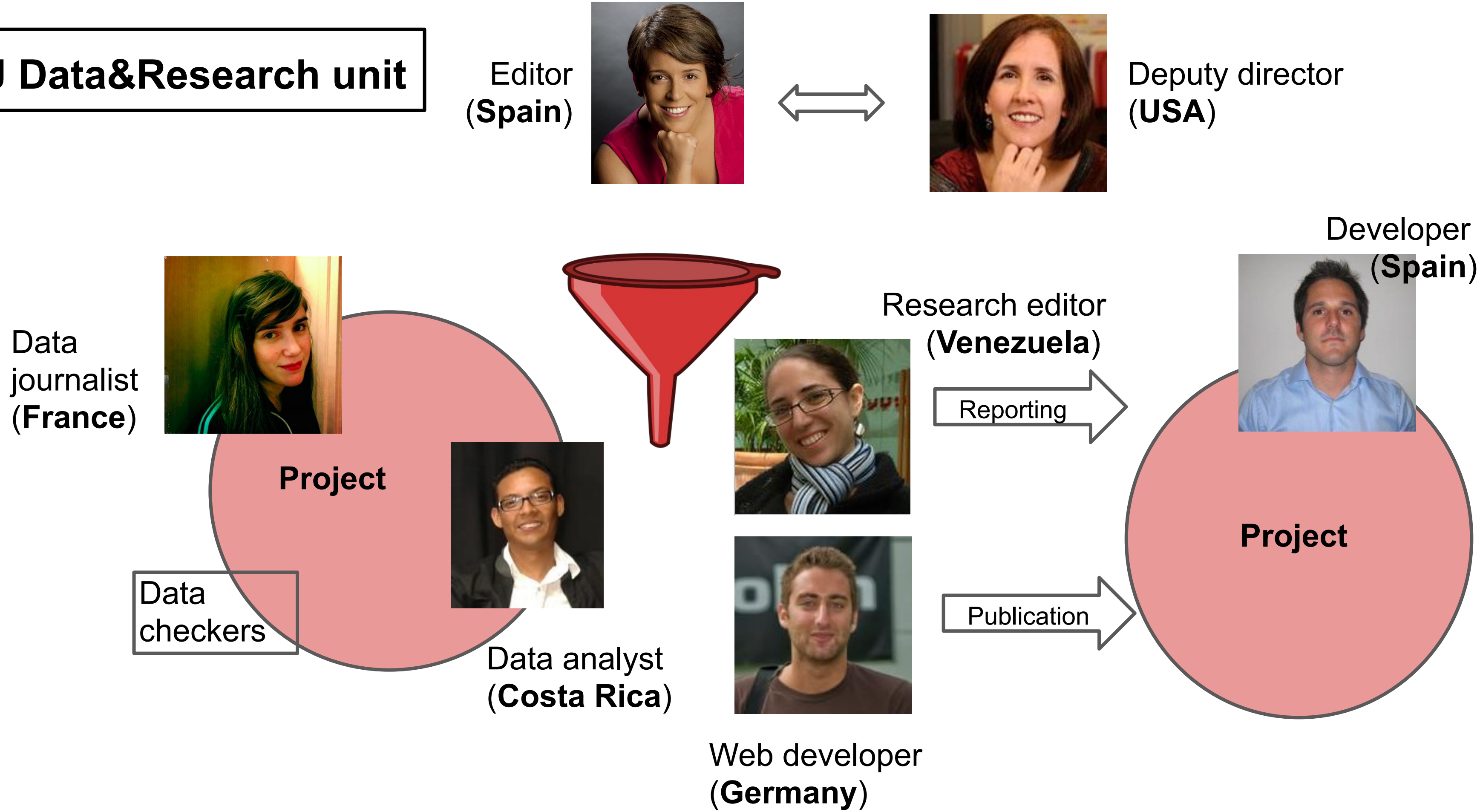
The World's Best Cross-Border Investigative Team

ICIJ The International Consortium of Investigative Journalists

+190 journalists in more than 65 countries

12 staff members (USA, Costa Rica, Venezuela, Germany, France, Spain)
50% of the team = Data & Research Unit

ICIJ Data&Research unit



Editor
(**Spain**)

Deputy director
(**USA**)

Data
journalist
(**France**)

Project

Data
checkers

Data analyst
(**Costa Rica**)

Research editor
(**Venezuela**)

Web developer
(**Germany**)

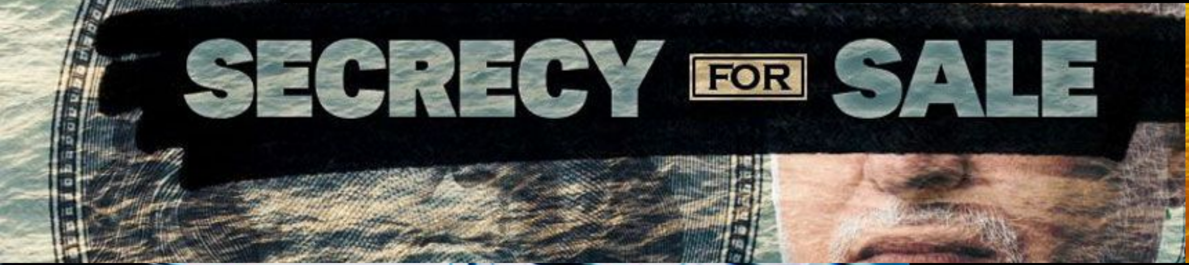
Reporting

Publication

Developer
(**Spain**)

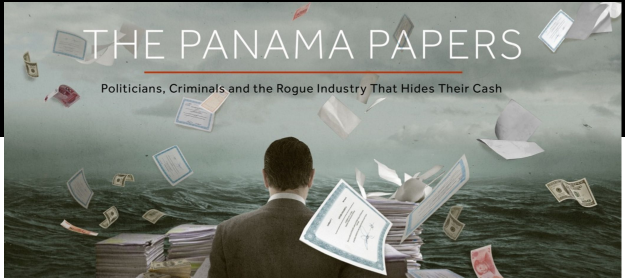
Project

260 GB - 100,000 companies



More than 100,000 HSBC clients
$100 Billion

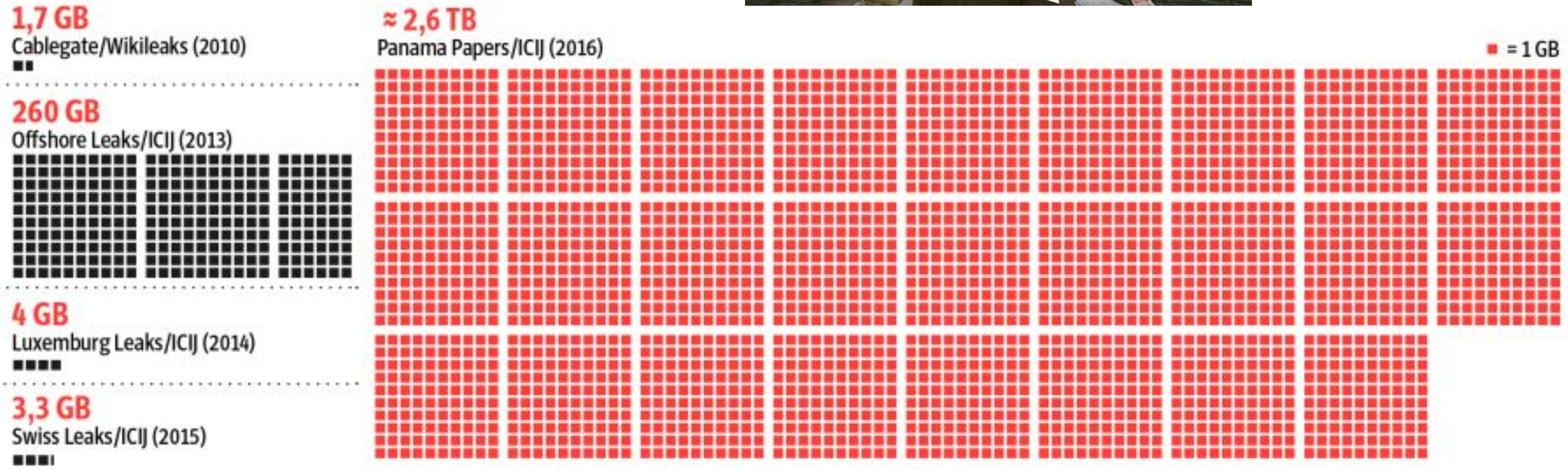
+500 secret tax agreements (PDFs)

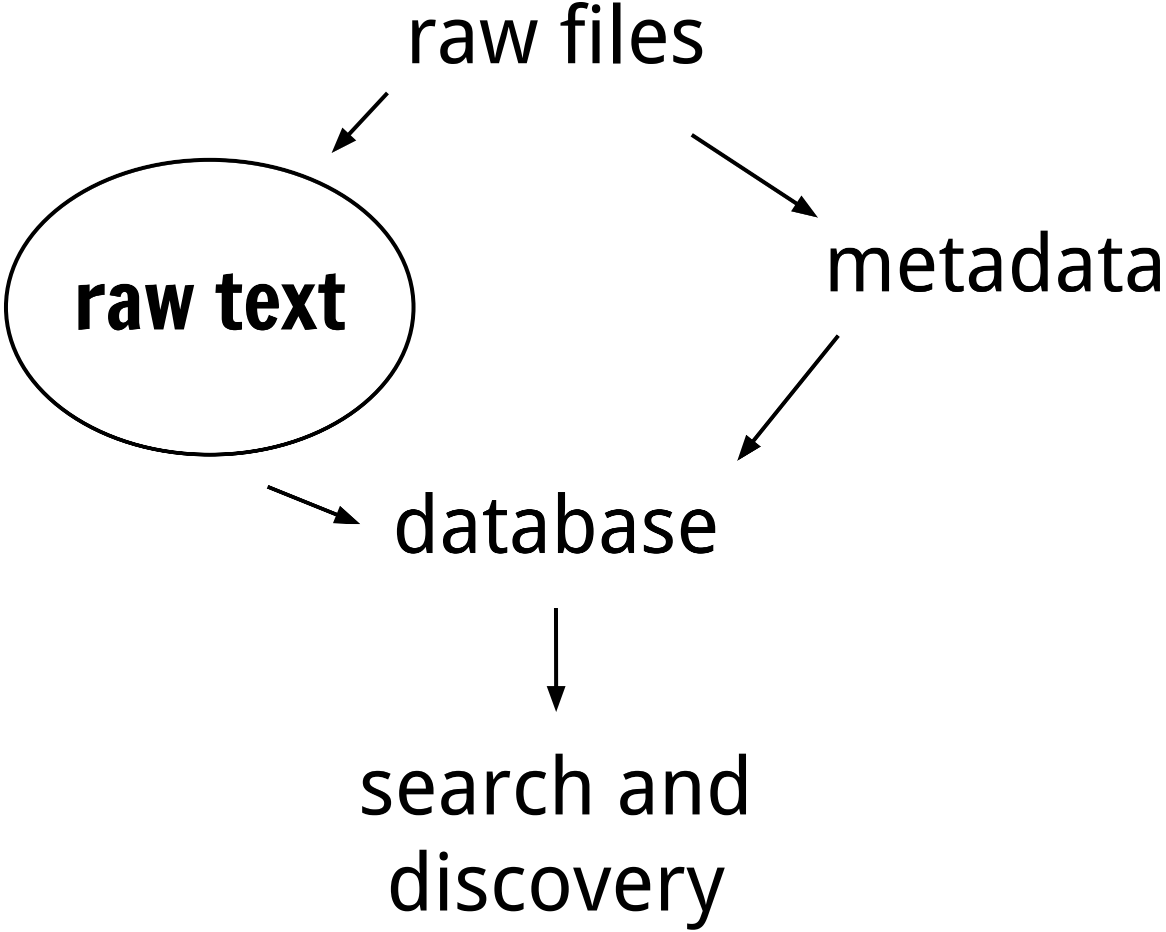


THE PANAMA PAPERS
Politicians, Criminals and the Rogue Industry That Hides Their Cash

The scale of the leak

Volume of data compared to previous leaks





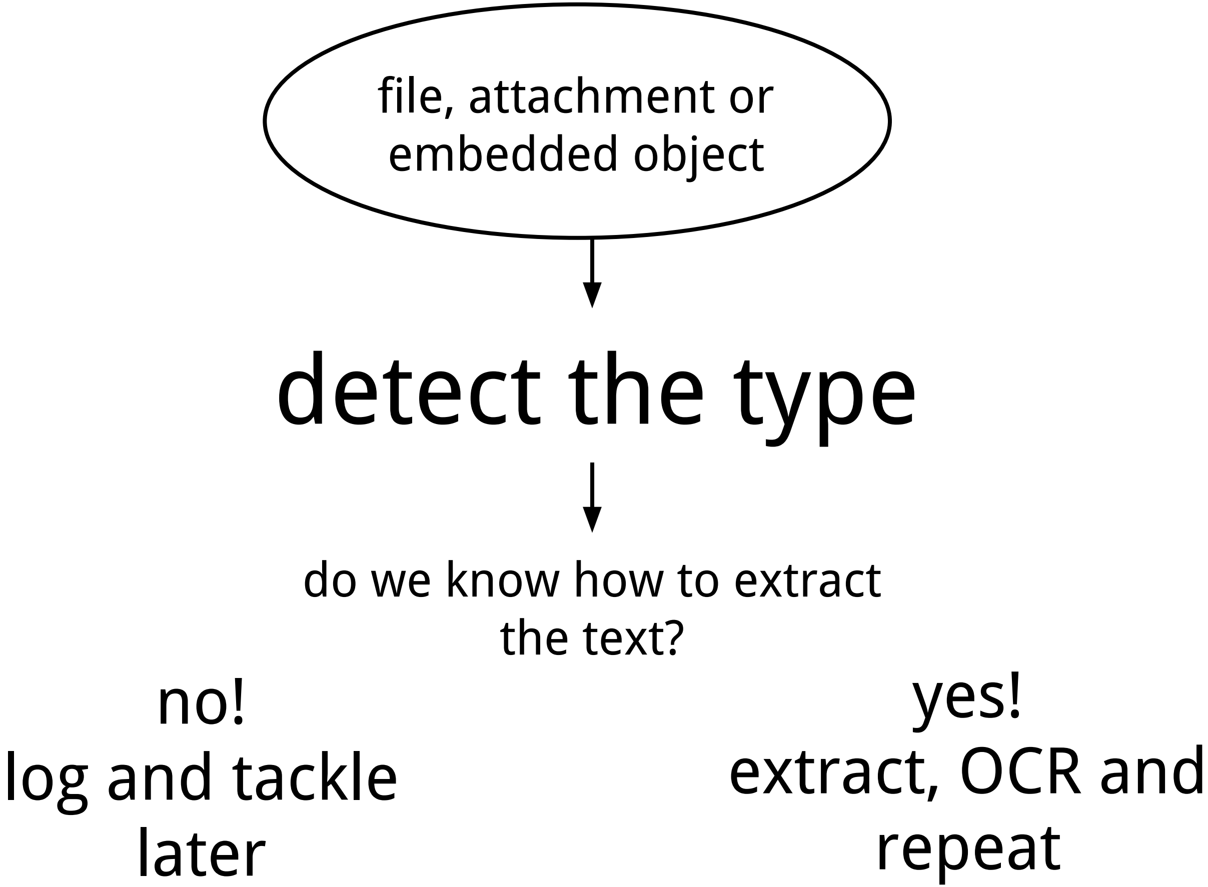
raw files

raw text

metadata

database

search and discovery



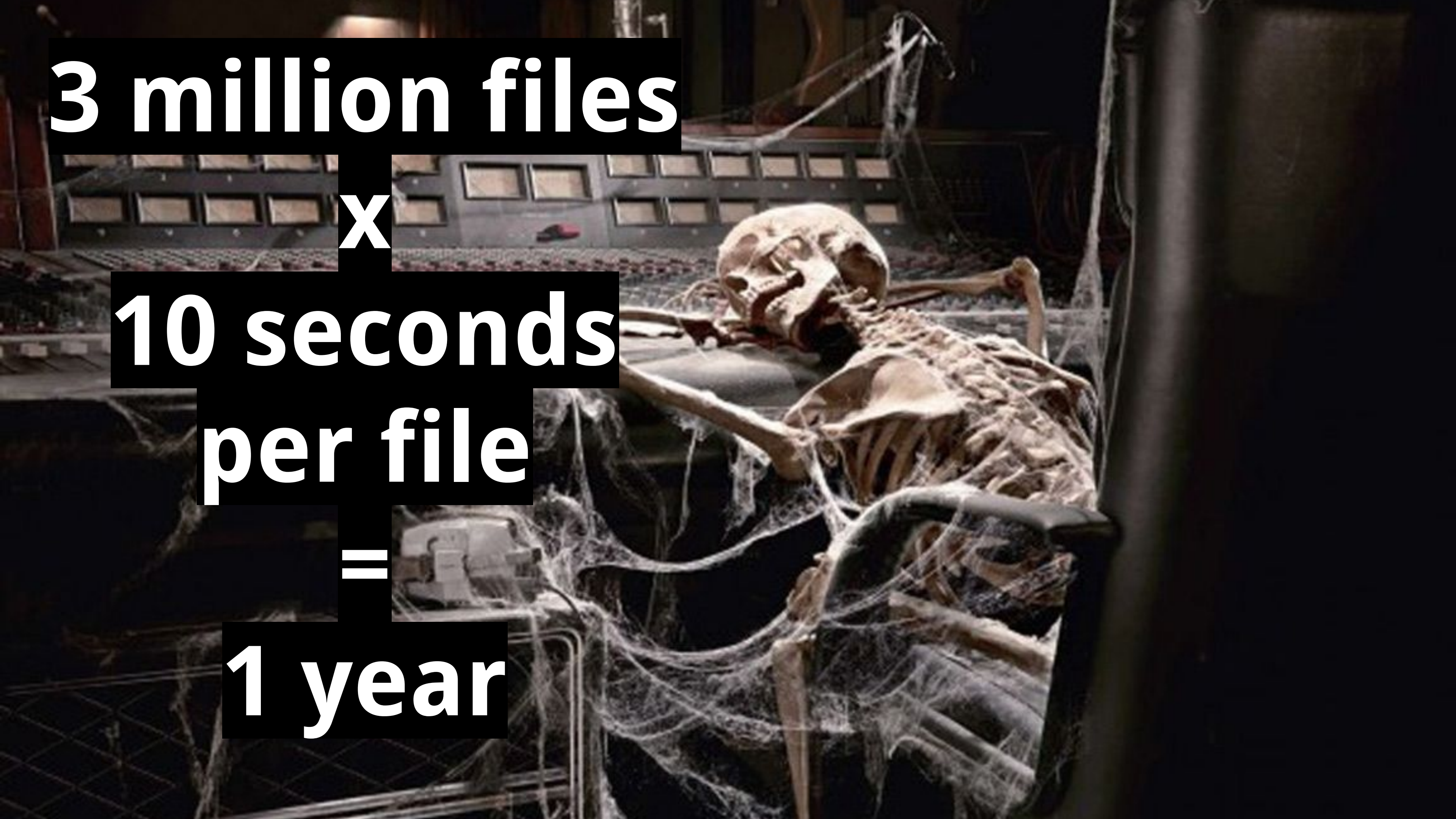
file, attachment or embedded object

detect the type

do we know how to extract the text?

no!
log and tackle later

yes!
extract, OCR and repeat



3 million files

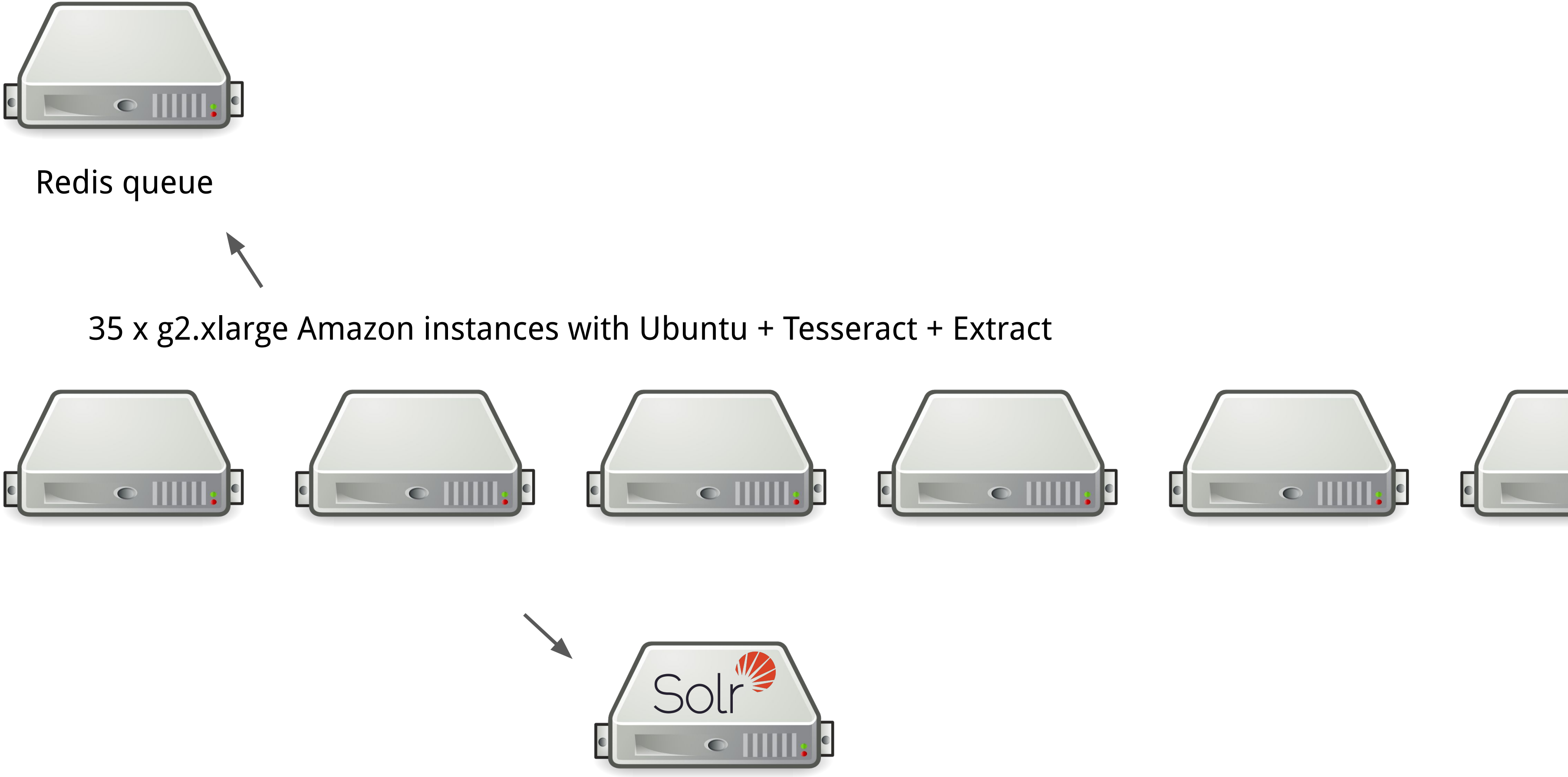
x

10 seconds

per file

=

1 year



Redis queue

35 x g2.xlarge Amazon instances with Ubuntu + Tesseract + Extract

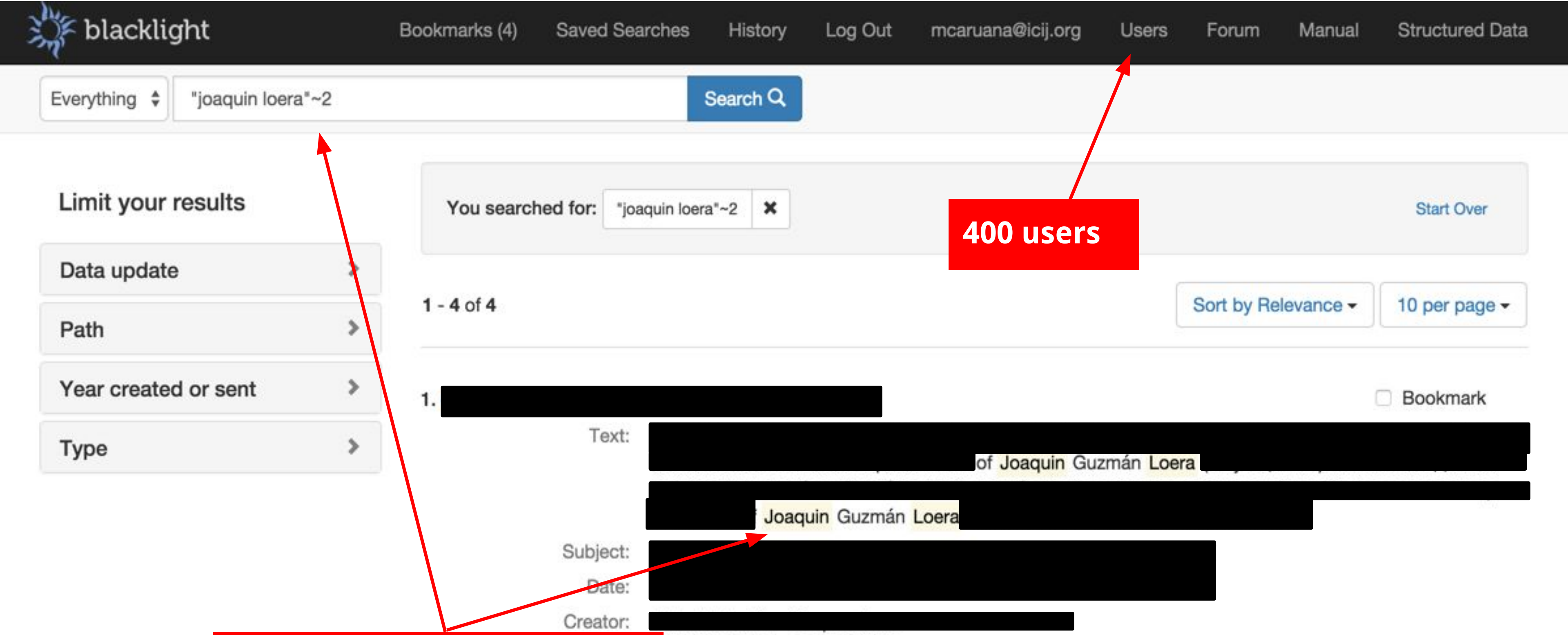
1 year

÷

35 machines

=

11 days



"joaquin loera"~2

You searched for: "joaquin loera"~2 ✖

1 - 4 of 4

1. [redacted]

Text: [redacted] of Joaquin Guzmán Loera [redacted] Joaquin Guzmán Loera [redacted]

Subject: [redacted]

Date:

Creator: [redacted]

400 users

Lucene syntax queries with proximity matching!

Solr

Stack

Unstructured data extraction

- ICIJ Extract (open source, Java: https://github.com/ICIJ/extract), leverages Apache Tika, Tesseract OCR and JBIG2-ImageIO.

Structured data extraction

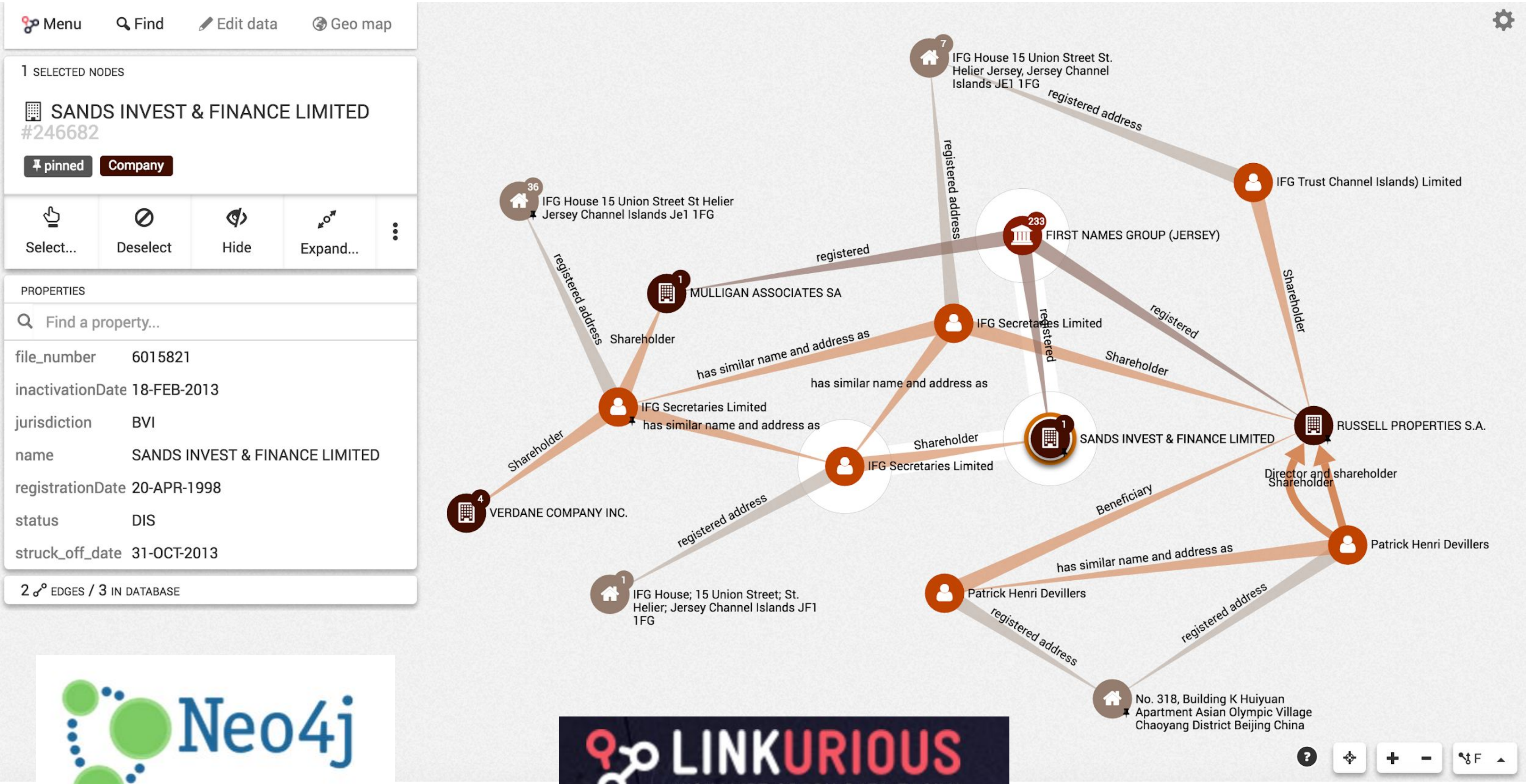
- A bunch of Python

Database

- Apache Solr (open source, Java)
- Redis (open source, C)
- Neo4j (open source, Java)

App

- Blacklight (open source, Rails)
- Linkurious (closed source, JS)



Menu Find Edit data Geo map

1 SELECTED NODES

SANDS INVEST & FINANCE LIMITED
#246682

pinned Company

Select... Deselect Hide Expand...

PROPERTIES

Find a property...

file_number	6015821
inactivationDate	18-FEB-2013
jurisdiction	BVI
name	SANDS INVEST & FINANCE LIMITED
registrationDate	20-APR-1998
status	DIS
struck_off_date	31-OCT-2013

2 EDGES / 3 IN DATABASE

7
IFG House 15 Union Street St. Helier Jersey, Jersey Channel Islands JE1 1FG
registered address
registered address
IFG Trust Channel Islands) Limited
36
IFG House 15 Union Street St Helier Jersey Channel Islands Je1 1FG
233
FIRST NAMES GROUP (JERSEY)
registered
Shareholder
registered address
1
MULLIGAN ASSOCIATES SA
IFG Secretaries Limited
registered
registered
Shareholder
has similar name and address as
Shareholder
has similar name and address as
IFG Secretaries Limited
has similar name and address as
Shareholder
1
SANDS INVEST & FINANCE LIMITED
RUSSELL PROPERTIES S.A.
Shareholder
IFG Secretaries Limited
Director and shareholder
Shareholder
4
VERDANE COMPANY INC.
Beneficiary
registered address
Patrick Henri Devillers
has similar name and address as
1
IFG House; 15 Union Street; St. Helier; Jersey Channel Islands JF1 1FG
Patrick Henri Devillers
registered address
registered address
No. 318, Building K Huiyuan Apartment Asian Olympic Village Chaoyang District Beijing China

Neo4j

LINKURIOUS
VISUALIZE GRAPH DATA EASILY

Our platforms by the numbers

- + 14 million documents, 20 formats
- 500 users* - 100 active each week, close to 200 per month
- 1 full-time programmer for improvements

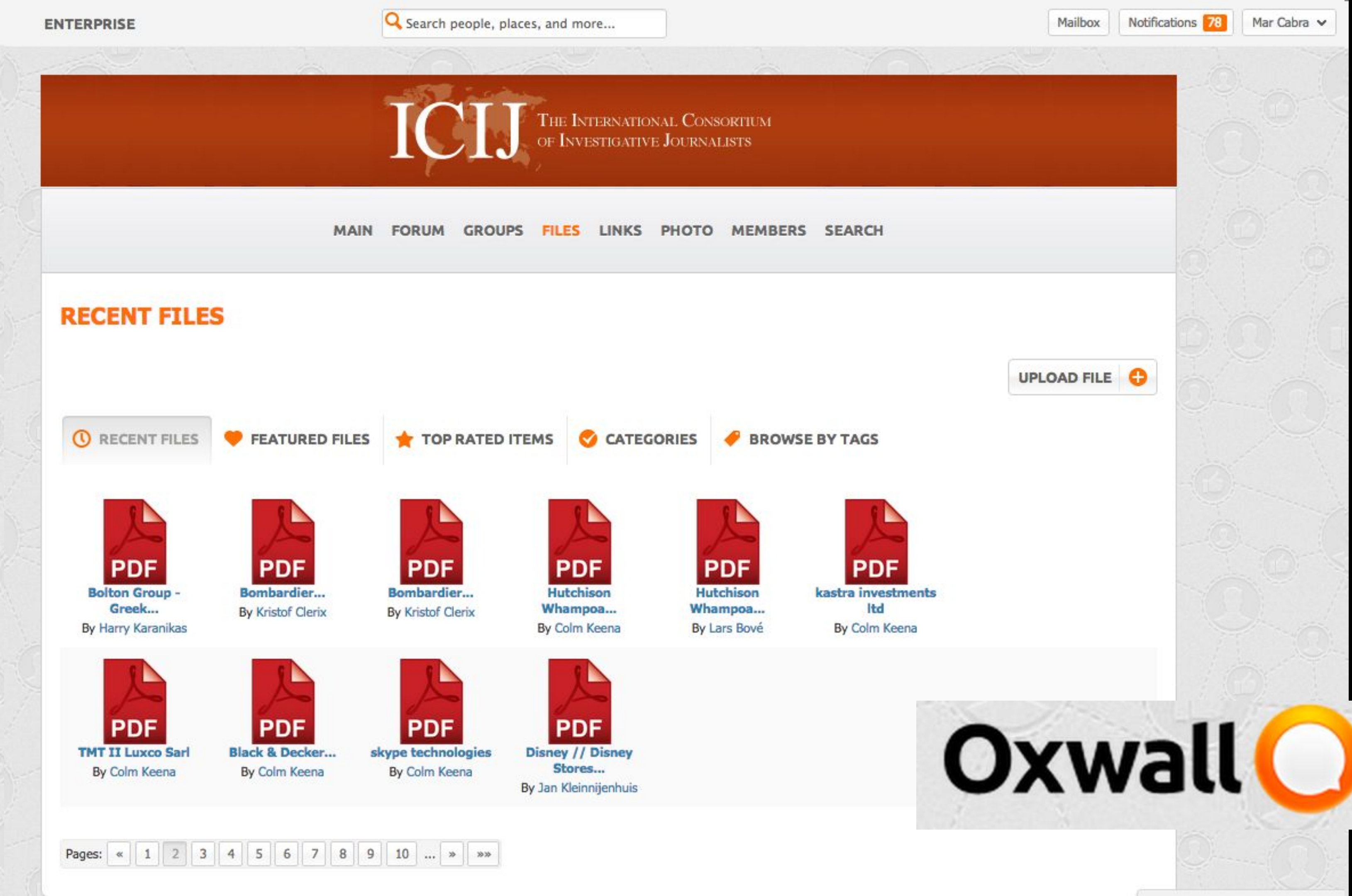
Open-source + exclusive expertise**

*not unique users

*except graph database

Not just a network…

… but a community



ENTERPRISE

Search people, places, and more...

Mailbox

Notifications 78

Mar Cabra

ICIJ The International Consortium of Investigative Journalists

MAIN FORUM GROUPS FILES LINKS PHOTO MEMBERS SEARCH

RECENT FILES

UPLOAD FILE

RECENT FILES | FEATURED FILES | TOP RATED ITEMS | CATEGORIES | BROWSE BY TAGS

PDF Bolton Group - Greek... By Harry Karanikas

PDF Bombardier... By Kristof Clerix

PDF Bombardier... By Kristof Clerix

PDF Hutchison Whampoa... By Colm Keena

PDF Hutchison Whampoa... By Lars Bové

PDF kastra investments ltd By Colm Keena

PDF TMT II Luxco Sarl By Colm Keena

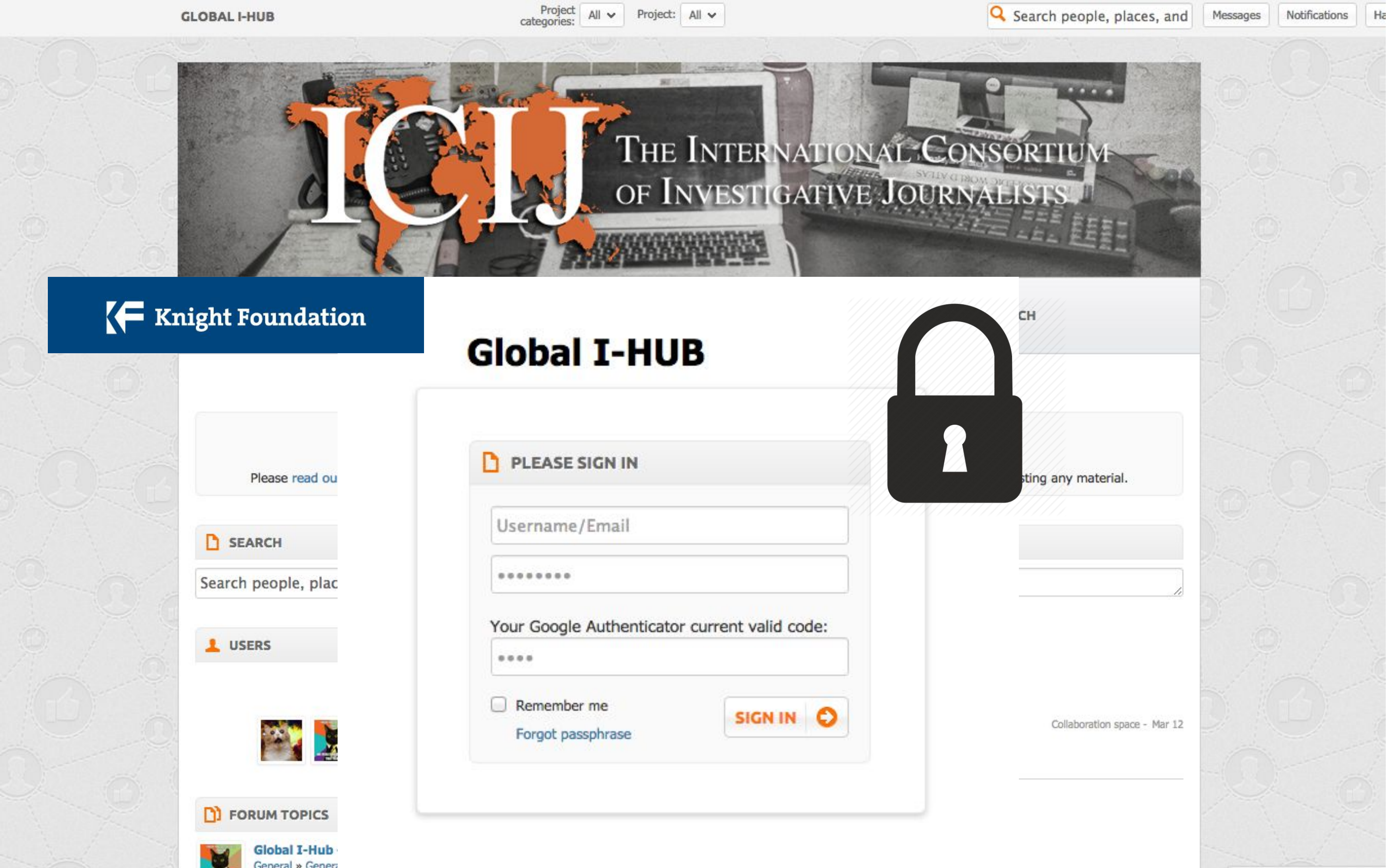
PDF Black & Decker... By Colm Keena

PDF skype technologies By Colm Keena

PDF Disney // Disney Stores... By Jan Kleinnijenhuis

Pages: « 1 2 3 4 5 6 7 8 9 10 ... » »»

Oxwall



GLOBAL I-HUB

Project categories: All

Project: All

Search people, places, and

Messages

Notifications

Ha

ICIJ

The International Consortium of Investigative Journalists

Knight Foundation

Global I-HUB

Please read ou

sting any material.

SEARCH

Search people, plac

USERS

FORUM TOPICS

Global I-Hub

General » Genera

CH

PLEASE SIGN IN

Username/Email

••••••••

Your Google Authenticator current valid code:

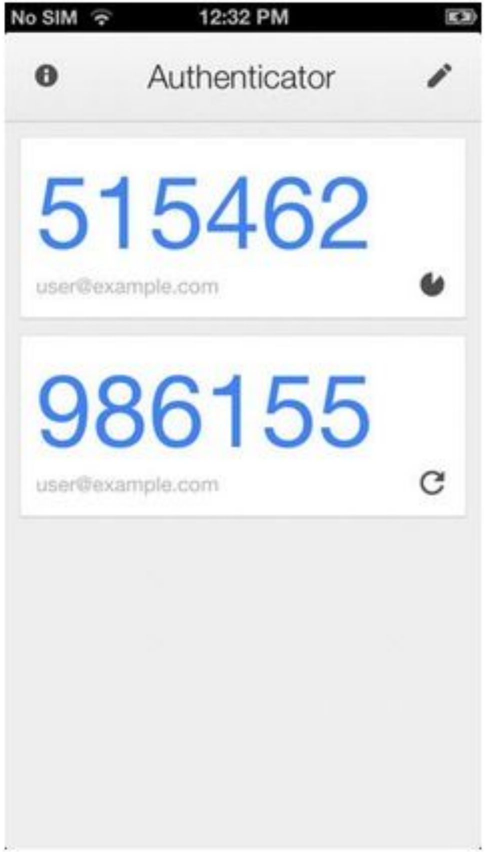
••••

□ Remember me

Forgot passphrase

SIGN IN

Collaboration space - Mar 12



No SIM 12:32 PM

Authenticator

515462

user@example.com

986155

user@example.com

Our community by the numbers

- +600 users*
- Shared status in +21,000 occasions**
- 1,500 forum topics with 5,400 posts**
- Uploaded +600 files**

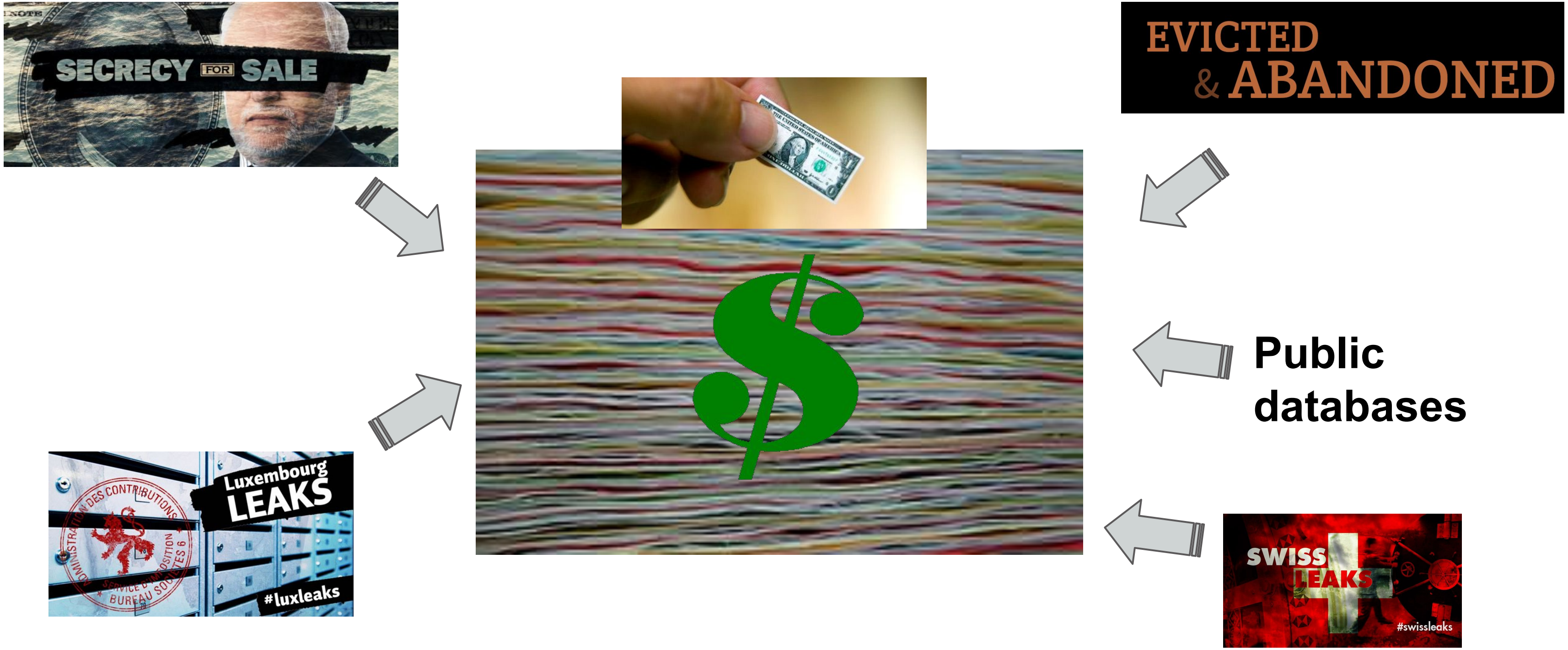
*not unique users

**October 2015

the next steps

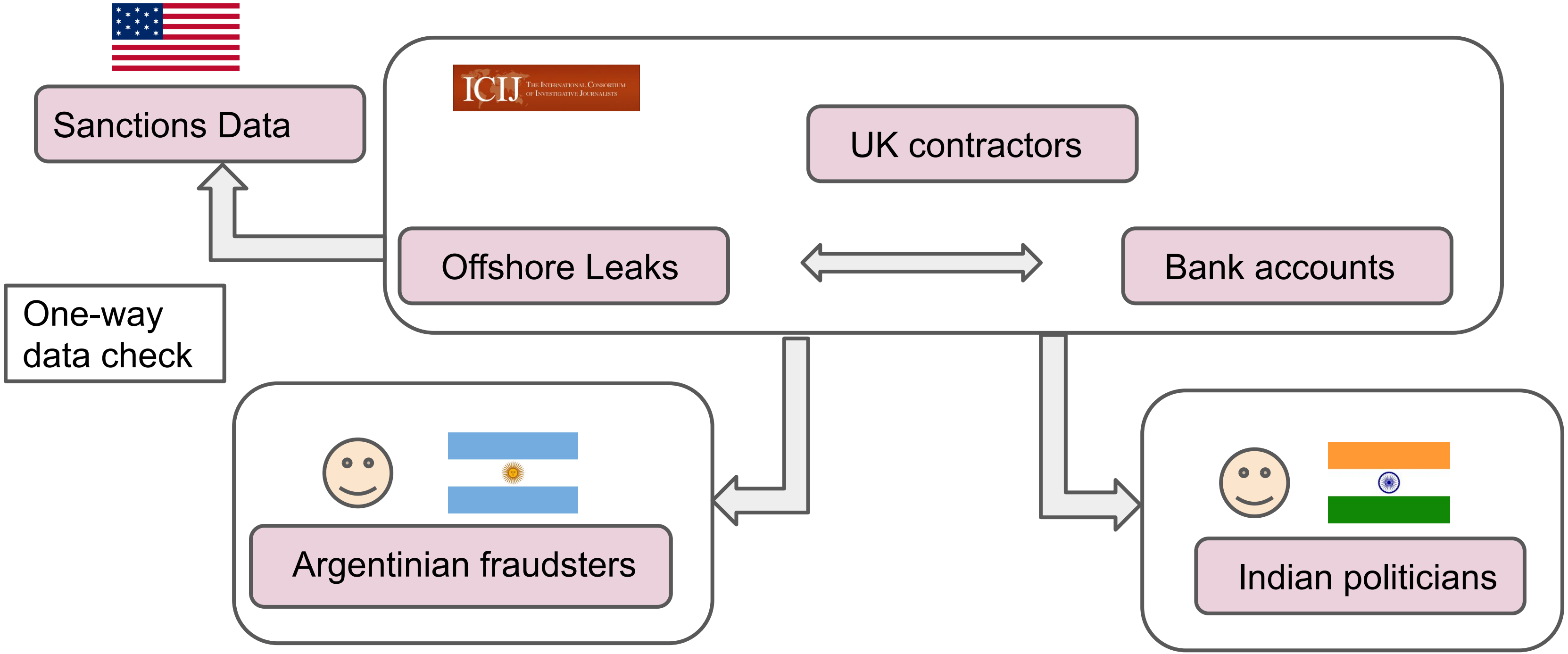
- entity extraction
- making our data silos talk

Next: ICIJ Knowledge Center



Public databases

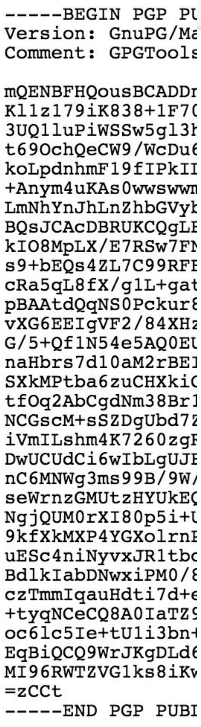
Our current challenge *(global data sharing)*



Encrypted email (easy)

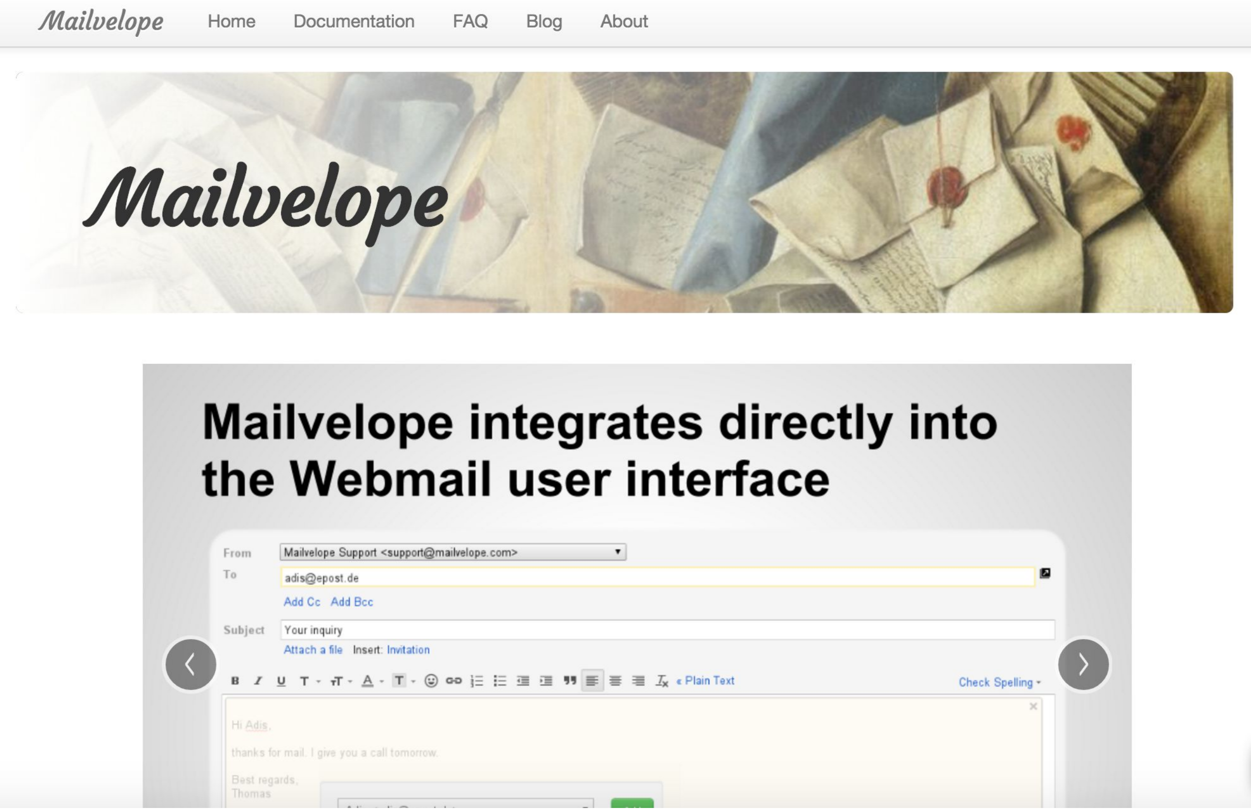
Review of tools

Encrypted email: PGP



```
-----BEGIN PGP PU
Version: GnuPG/Ma
Comment: GPGTools

mQENBFHQousBCADDr
Kl1z179iK838+1F7C
3UQ1luPiWSSw5gl3h
t69OchQeCW9/WcDu6
koLpdnhmF19fIPkII
+Anym4uKAs0wwswwm
LmNhYnJhLnZhbGVyb
BQsJCAcDBRUKCQgLE
kIO8MpLX/E7RSw7FN
s9+bEQs4ZL7C99RFE
cRa5qL8fX/g1L+gat
pBAAtdQqNS0Pckur8
vXG6EEIgVF2/84XHz
G/5+Qf1N54e5AQ0EU
naHbrs7d10aM2rBEI
SXkMPtba6zuCHXkiC
tfOq2AbCgdNm38Br1
NCGscM+sSZDgUbd7Z
iVmILshm4K7260zgF
DwUCUdCi6wIbLgUJE
nC6MNWg3ms99B/9W/
seWrnzGMUtzHYUkEQ
NgjQUM0rXI80p5i+U
9kfXkMXP4YGXolrnE
uESc4niNyvxJR1tbc
BdlkIabDNwxiPM0/8
czTmmIqauHdti7d+e
+tyqNCeCQ8A0IaTZ9
oc6lc5Ie+tU1i3bn+
EqBiQCQ9WrJKgDLd6
MI96RWTZVG1ks8iKw
=zCCt
-----END PGP PUBI
```



Mailvelope integrates directly into the Webmail user interface

From: Mailvelope Support <support@mailvelope.com>

To: adis@epost.de

Subject: Your inquiry

Hi Adis,

thanks for mail. I give you a call tomorrow.

Best regards,
Thomas

Questions?

Thanks!

mcabra@icij.org