# Elastic Search

Inutile de chercher ailleurs

# Who's who

- Gérald Croës
  - Thèse sur les bases de données fragmentées avec indexations multiples et réparties.
  - Expert en analyse comportementale dans les processus de "recherche synchrone et asynchrone".
- Julien Salleyron
  - Consultant moteurs de recherche depuis 15 ans.
  - Lead du projet "Alpha Search", moteur de recherche en 4D.

# Who's who

- Gérald Croës
  - Thèse sur les bases de données fragmentées avec indexations multiples et réparties.
  - Expert en analyse comportementale dans les processus de "recherche synchrone et asynchrone".
- Julien Salleyron
  - Consultant moteurs de recherche depuis 15 ans.
  - Lead du projet "Alpha Search", moteur de recherche en 4D.

# Who's who

## Gérald Croës

@geraldcroes

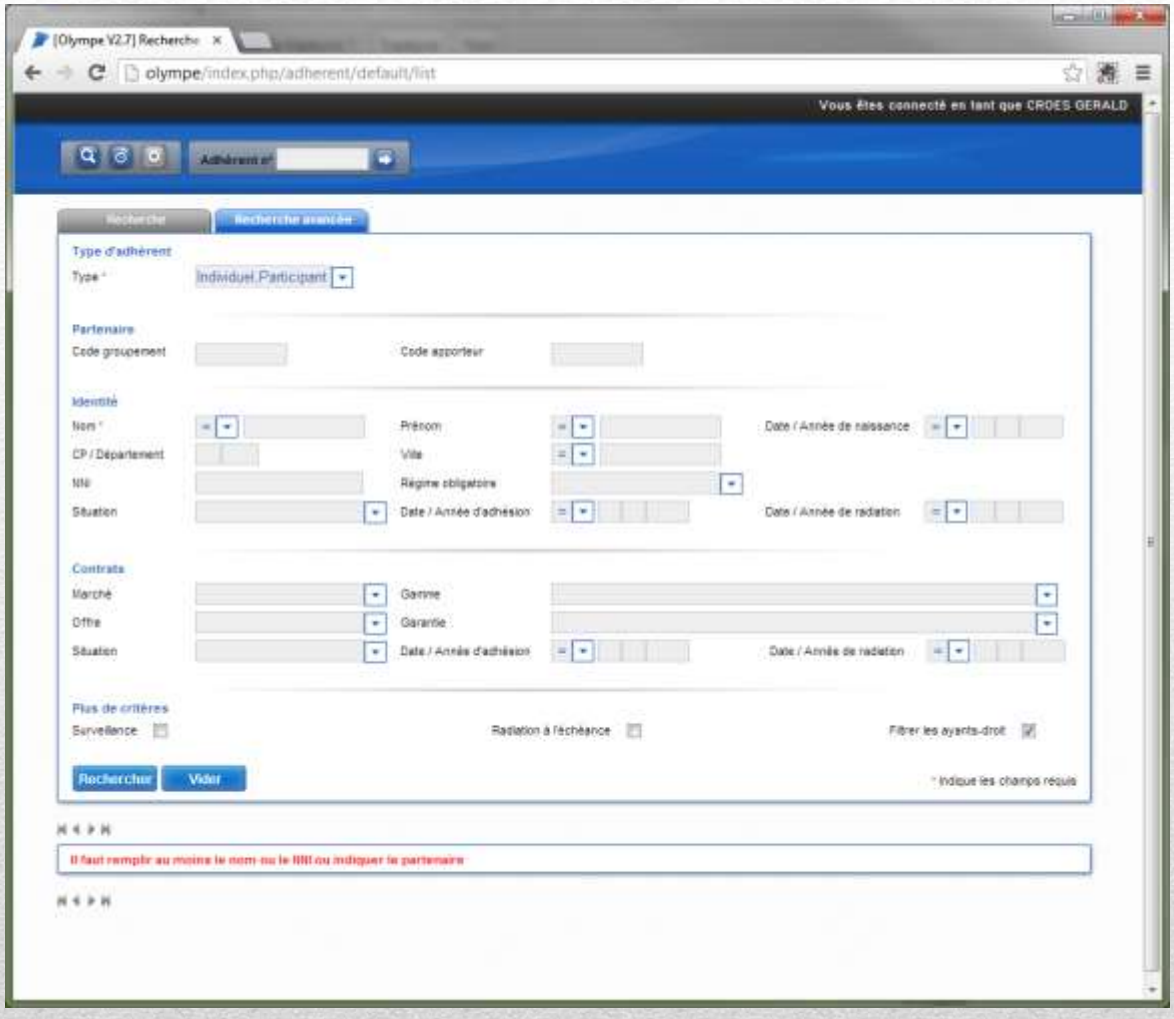
www.croes.org/gerald/

gerald@php.net

## Julien Salleyron

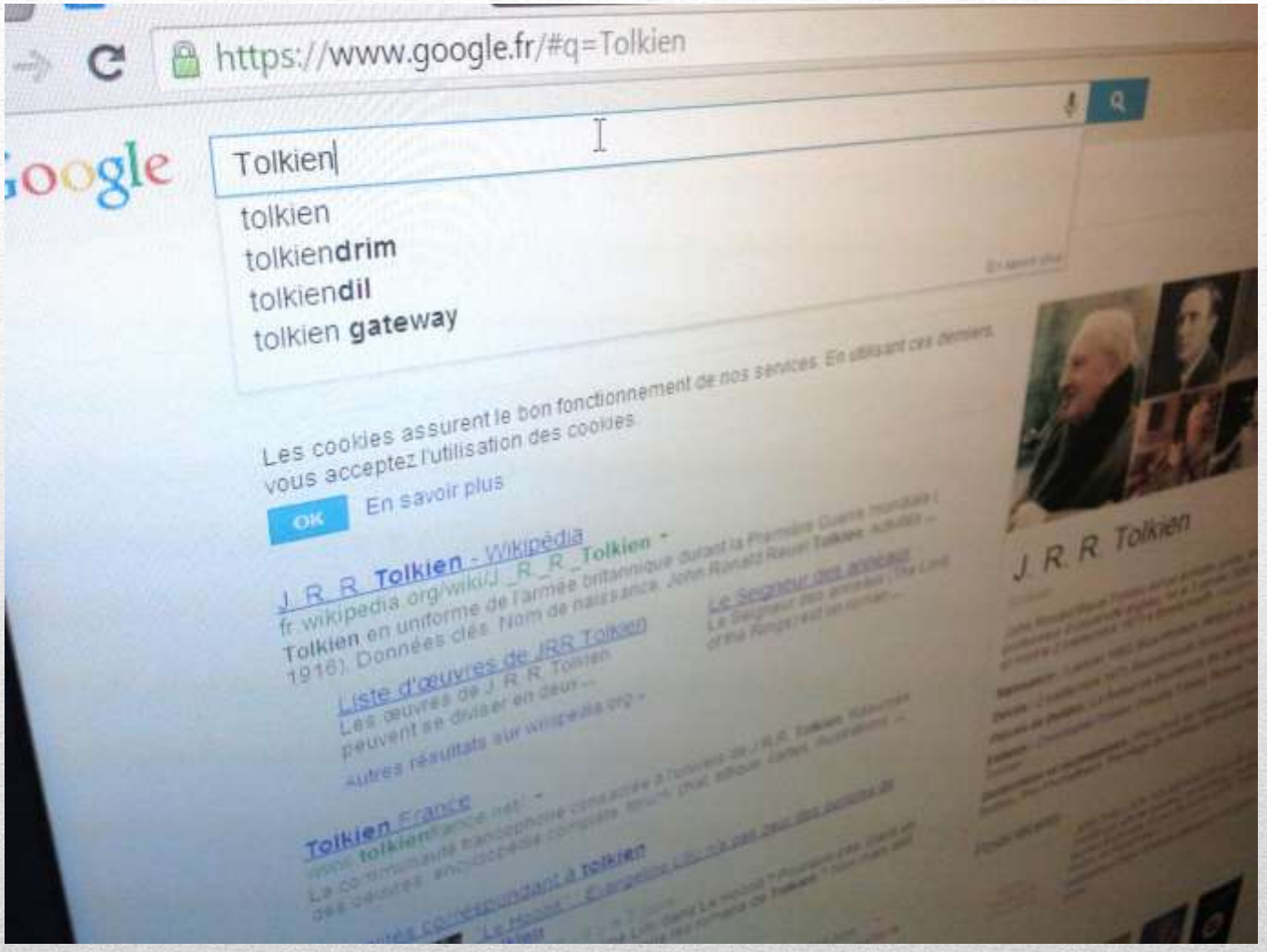
@juguul

juliens@php.net



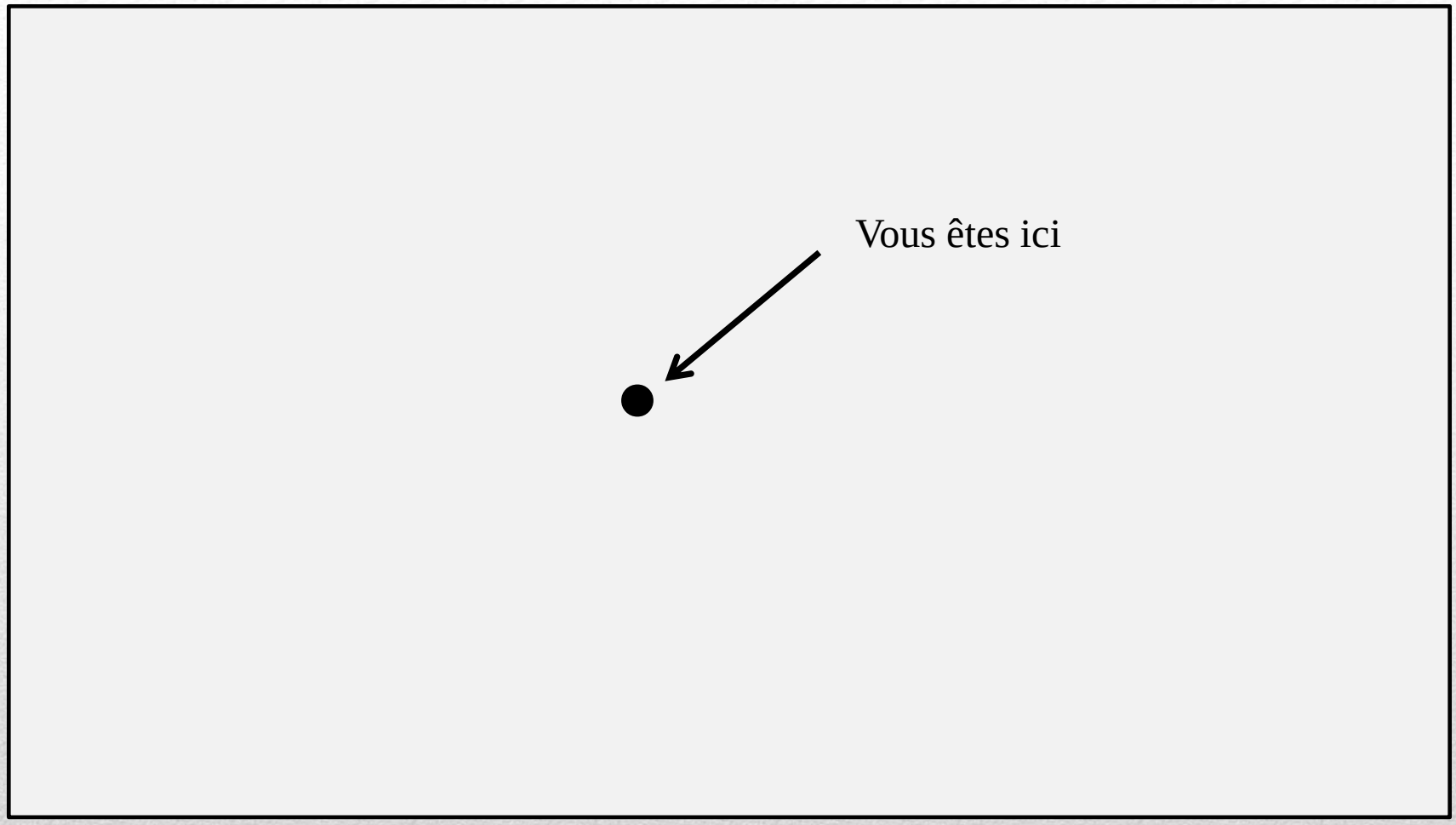

# Il n'y a pas si longtemps


# "Moi je veux Google"




# A la Google, c'est quoi ?


# Inadapté...




# Perdu ?

Xapian







# Les solutions...

Xapian

Lucene

Apache Solr

ElasticSearch

Sphinx

# ...Open Source...


# ...Qu'on a compris en moins de 10 minutes.


...Qu'on a compris en moins de 10 minutes.


# C'est du JAVA !?


# RESTez ! ☺


- Open Source (licence Apache 2)
- Orienté Document *
- Pas de schéma établi (schema less)
- Scalabilité (oui, c'est un Anglicisme)
- JSON
- Rapide
- Langage de requête puissant

# En gros...

* Un peu comme mongoDB


# Une base de données orientée recherche

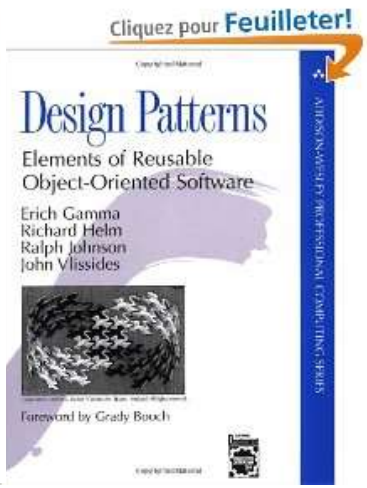
# Un peu de vocabulaire

- **Index** – Lieu de stockage pour un ensemble fonctionnel (exemple : boutique contient des CD, Livres, DVD, Aspirateurs, …)
- **Type** – Une famille de documents
- **Document** – Un élément typé et identifié

Type CD




Type Livre


Type DVD



# L'exemple d'une boutique

# Avant de mettre en pratique

REST… tout le monde connait ?


http://

**REST***

GET - Lecture
POST - Création
PUT - Modification
DELETE - Suppression

# Rappels sur la grammaire

* Le truc qui marche bien avec un *file_get_contents*

http://elastic_search:port/**index**/**type**/**identifiant**

http://elastic_search:port/**boutique**/**livre**/**978-0201633610**

{

Titre : Design Patterns

Auteurs : [Eric Gamma, Richard Helm, Ralph Johnson, John Vlissides]

}

# Vocabulaire en pratique

http://elastic_search:port/**index**/**type**/**identifiant**

http://elastic_search:port/**boutique**/**dvd**/**B001E08UMU**

```
{
  Titre : Matrix 2 Reloaded
  Minutes : 138
}
```

# Vocabulaire en pratique

# Remplir la base Elastic Search

```
curl -XPOST http://elastic_search:port/boutique/cd -d '{
   "titre" : "19 Nocturnes",
   "artiste" : "Arthur Rubinstein",
   "compositeur" : "Fryderyk Chopin",
   "genre" : "classique"
} '
```

```
{
  "ok" : true,
  "_index" : "boutique",
  "_type" : "cd",
  "_id" : "B000031WBV",
  "_version" : 1
}
```

REST

# Création

```
curl -XPUT http://elastic_search:port/boutique/cd/B000031WBV -d '{
   "titre" : "19 Nocturnes",
   "artiste" : "Arthur Rubinstein",
   "compositeur" : "Fryderyk Chopin",
   "genre" : "classique"
} '
```

```
{
   "ok" : true,
   "_index" : "boutique",
   "_type" : "cd",
   "_id" : "B000031WBV",
   "_version" : 2
}
```

REST

# Création / Mise à jour

# Une première recherche ?

Brute… mais recherche tout de même

```
curl -XGET http://elastic_search:port/boutique/_search?q=Chopin
```


```
{
   "took" : 153,
   "timed_out" : false,
   "_shards":{"total":5,"successful":5,"failed":0},
   "hits":
   {
      "total" : 1,
      "max_score" : 0.11506981,
      "hits":[{
               "_index" : "boutique",
               "_type" : "cd",
               "_id" : "4J7iyPPvSXW1Rjufu0oJbg",
               "_score" : 0.11506981,
               "_source" : {
                  titre : "19 Nocturnes",
                  artiste : "Arthur Rubinstein",
                  compositeur : "Fryderyk Chopin",
                  genre : "classique"
               }
            }]
   }
}
```

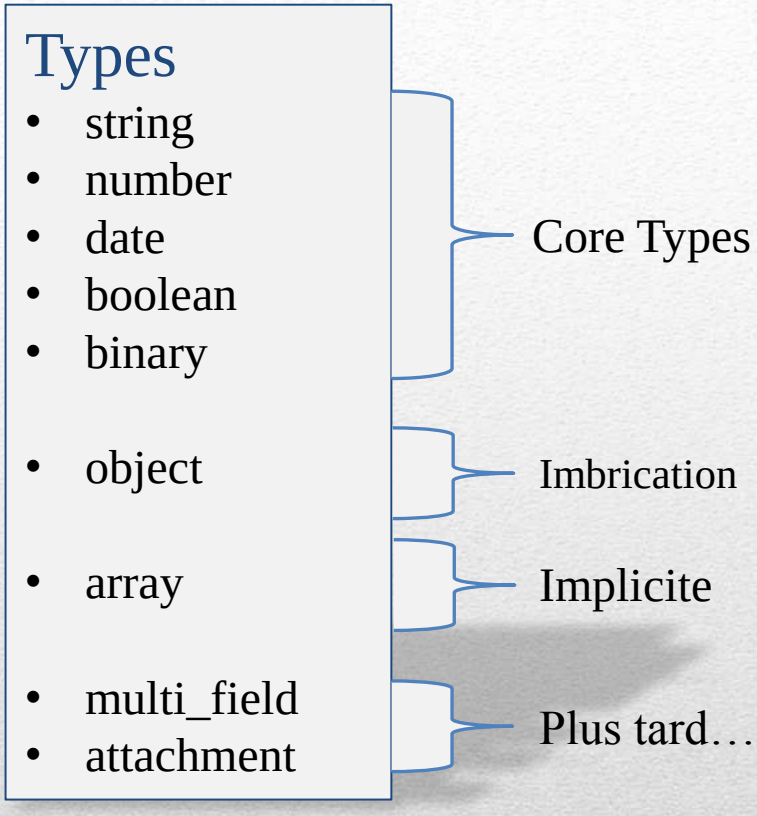

# Quels documents concernent Chopin ?

Dans ES, les actions sont préfixées de underscore "_"

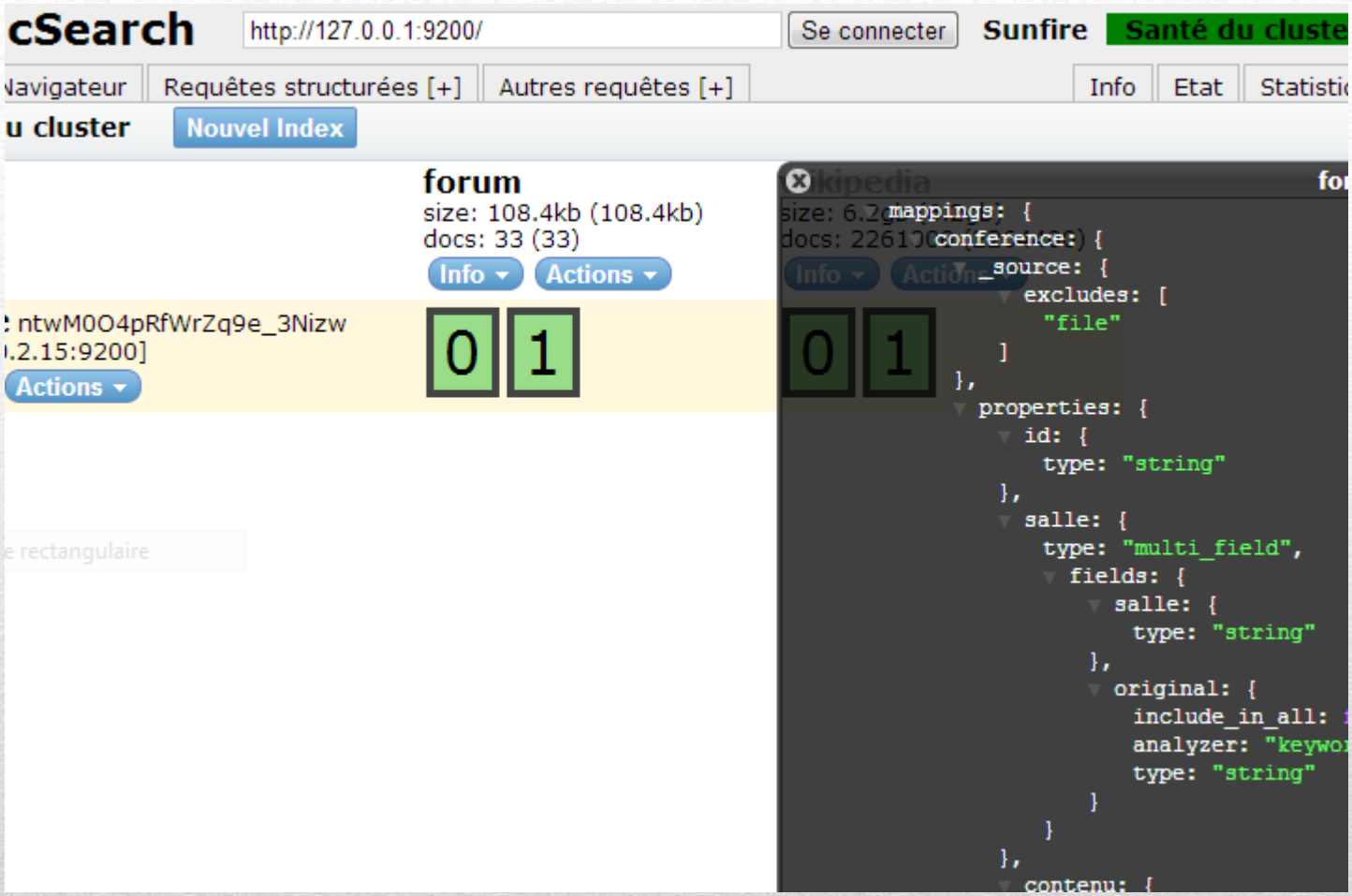
# La magie dans les coulisses

Qui ne tient toutefois pas du miracle

```
{
   "cd" : {
      "properties" : {
         "titre" : {
            "type" : "string"
         },
         "sortie" : {
            "type" : "date",
            "format" : "date_time_no_millis"
         },
         "disque" : {
            "properties" : {
               "piste" : string
            }
         }
      }
   }
}
```




# Décrire les types avec Json Schema




# elastichead

# Mais ce n'est toujours pas suffisant...

Indexation & Recherche

# La chaine d'analyse

Quand "WTF Dude ?" devient "Mais qu'est-ce que cette chose étrange, très cher ami ?"

# CharFilter

Pré-processeur sur la chaine de charactères

[ ‘ph’=>’f’]

[ ‘qu’=>’k’]

# CharFilter - Mapping

Texte
indexé

Texte indexé

# CharFilter - HTML Strip

# Tokenizer

Découpage de la chaine de caractère en token

Texte en token-standard

[Texte]

[en]

[token]

[standard]

# Tokenizer - Standard

Texte en token-whitespace

[Texte]

[en]

[token-whitespace]

# Tokenizer - Whitespace

Texte ngram

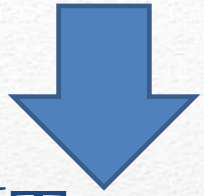
[Te][Tex][Text][Texte]

[ng][ngr][ngra][ngram]

# Tokenizer - Edge NGram

# Tokenizer - NGram

Texte ngram



[Te][Tex][Text][Texte][ex][ext][exte]

[xt][xte] [te]

[ng][ngr][ngra][ngram]
[gr][gra][gram]

[ra][ram] [am]

Texte en keyword

[Texte en keyword]

# Tokenizer - keyword

/root/branch/leaf

[/root]

[/root/branch]

[/root/branch/leaf]

# Tokenizer - path hierarchy

# Token Filter

Filtrage, Enrichissement et Modification des tokens

[Texte][à][indexer]

[texte][à][indexer]

# Token Filter - Lowercase

[texte][à][indexer]

[texte][a][indexer]

# Token Filter - AsciiFolding

[texte][a][indexer]

[texte]

# Token Filter – Length (ex. 3-5)

[indexe][indexer][indexation]

[index]

# Token Filter - Stemmer

[l’avion]

[avion]

# Token Filter - Ellision

[le][chien][est][noir]

[chien][noir]

# Token Filter - Stop

[Documentation]

[Documentation][Forum][Manuel][RTFM]

# Token Filter - Synonym

Hunspell

Kstem

Snowball

Phonetic

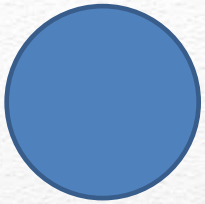
…

# Token Filter

Mon super texte, avec accents, créées

Char Filter

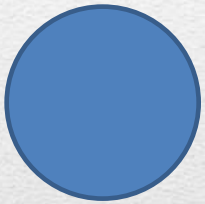
HTMLStrip



Mon super texte, avec accents , créées

Tokenizer

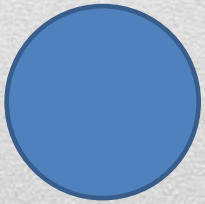
Letter Tokenizer
Whitespace



- mon
- super
- texte
- avec
- accents
- créées

Token Filter

Stemmer
Ascii folding
Snowball



- super
- text
- accen
- cree

# Dans l'ordre !

# La recherche

Profiter du travail réalisé

# Le langage de requête

## _search

```
curl -XPOST
curl -XGET

'http://elastic_search:port/boutique/_search'  -d
```

## Query DSL

```
{"query":
   {"match":
      {"titre" : "matrix  reloaded"}
   }
}
```

```
{
   "took" : 153, …
   {
      "total":2,  "max_score":0.91506981,
      "hits":[{
          "_index" : "boutique",
          "_type" : "dvd",
          "_id":"5Gh3y…",
          "_score":0.91506981,
          "_source" : {
             titre:"Matrix Reloaded", …
          }
      },  …]
   }
}
```

# La réponse

```
{ match:
      { ‘titre’: ‘Matrix reloaded’ }
}
```

[matrix] [reloaded]

Titre : *Matrix Reloaded*
Description : *Neo apprend à mieux contrôler ses dons naturels …*

# Les tokens, toujours les tokens…

```
{ match:
      { 'titre.ngram': 'Matrix' }
}
```

[ma] [mat] [matr] [matri] [matrix]

Titre : *Matrix Reloaded*
Description : *Neo* ...

Titre : *Maman j'ai raté la conf*
Description : ...

# Un problème mis en lumière

```
{ match:
      { ‘titre.ngram’: {'query' : ‘Matri’,
        ‘analyzer’    : ’keyword’ }}
}
```

[matri]

Titre : *Matrix Reloaded*
Description : *Neo* …

Titre : *Matrix a raté la conf*
Description : …

# Forcer l'analyzer

```
{"query":
   "bool":{
         "should": [
               {"match":
                  {"titre": "neo"}
                },
               {"match":
                  {"description": "neo"}
               },
          ]
    }
}
```

# Requête multi propriétés

```
[
   { match:   { ‘titre’: ‘neo’} },
   { match:   { ‘description’: ‘neo’} },
]
```

Titre : *Matrix Reloaded*
Description : *Neo* …

# Ce qui donne

```
[
   { match:   { 'titre': 'Matrix'} },
   { match:   { 'description': 'Matrix'} },
]
```

Titre : *Cloud atlas*
Description : *Par les réalisateurs de Matrix, Matrix Reloaded …*

Titre : *Matrix*
Description : *Neo …*

# Le plus important ?

# La pertinence

# Boost

```
[
   { match: { 'titre^4': 'Matrix'} },
   { match: { 'description': 'Matrix'} },
]
```

Titre : *Matrix*
Description : *Neo* ...
*_score* : *0.945*

Titre : *Cloud atlas*
Description : *Par les réalisateurs de Matrix, Matrix Reloaded* ...
*_score* : *0.911*

# Et si seulement c'était tout...

Il existe tellement d'options que nous ne pouvons pas tout citer

- Filter
- Facets
- Suggests
- Highlight
- River
- Plugins
- Fuzzy
- Like this
- Sharding
- …

# Des mots clefs à retenir

# Cerise sur le globiboulga

Un plugin existe pour gérer les documents bureautiques classiques

# Conclusion & DEMO ! :-)

Réalisée en une demi journée

# Questions ?