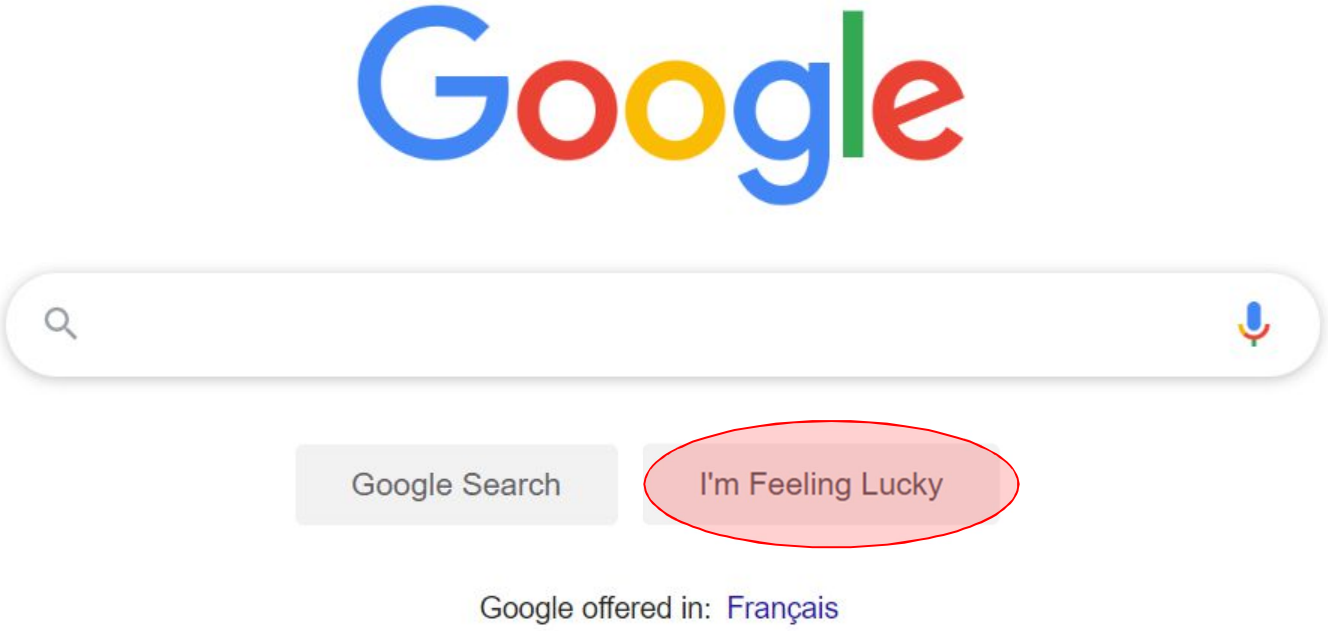
Relevant Facets

Lucian Precup
Radu Pop
@lucianprecup @a2lean
#haystackconf
Berlin EU 2019

// Poll

- How many of you are using facets with the search engines you implement ?
- Who is doing statistics on facet usage ?
- Who is using Solr ?
- Who is using Elasticsearch ?
- Other search technology ?
- Who speaks French ?

// Why this talk ?



// Facets ?

Used to define filters that refine the initial query

Used for disambiguation

Give a holistic view over the search results

Allow to find the needle in the haystack more quickly

Carrefour

tomate

Rayons
Promo, épicerie...

Maison & High Tech
Il y a du nouveau !

Services
Voyage, spectacle...

Trouver un magasin

Mes listes

Me connecter

Choisir Drive ou la livraison pour commencer vos courses

« tomate »

469 résultats

Trier par **Pertinence**

Affiner

Affiner par rayon

- Fruits et Légumes 77
- Viandes et Poissons 9
- Crémerie 1
- Charcuterie 1
- Traiteur 65

VOIR PLUS

Marque

Marques...

- CARREFOUR 89
- JARDIN BIO 18
- PANZANI 14
- SACLA 14
- MIX BUFFET 10

Prix imbattable !

Tomates rondes en grappe
le kilo
1.39 € / Kilogramme

FRANCE

1,39€

Tomate ronde en grappe cocktail
la barquette de 350 g
5.71 € / Kilogramme

FRANCE

2,00€

Tomate Cerise rouges rondes
la barquette de 1 kg
3.50 € / Pièce

MAROC

3,50€

Tomate Cerise rouges rondes
REFS SANS MARQUE

Promotion

Tomates cerises rondes
la barquette de 400g

Tomate ronde noire
le kilo

Tomate côtelée rouge
le kilo

jeudi 13 juin 2019

DIGGR

Categories

humanitarian priorities water sanitation

DASHBOARDS EN Adelean Digger

Did you mean : water hygiene and sanitation

Tips : With Diggr you can search from more than 35 differents MSF information platforms at the same time

About 153 results in 0.534 seconds

Display : 10 results per page Order by : Relevance Gallery Map

Hide search filters

ALL OOPS OOPSMED INSIDEOCB MSF WEBSITES TREASURY SHERLOG MAP CENTRE

Selected Filters :

MSF Websites

eng

LANGUAGES

- [x] eng (153)
- [ ] fra (5)
- [ ] ara (1)
- [ ] zho (1)

DATES OF TOPIC

PRODUCT

SOURCES

ORGANISATION

- [ ] Médecins Sans Frontières (122)
- [ ] World Health Organization (93)
- [ ] Drugs for Neglected

GO TO THE WATER HYGIENE AND SANITATION PAGE ON SHERLOG

GO TO THE WATER AND SANITATION PAGE ON OOPS MED

Setting priorities for humanitarian water, sanitation and hygiene research: a meeting report | DMello Guyett et al - 2018 - Setting priorities for humanitarian water, sanitation, and hygiene research.pdf

Last modified on 18/12/2018

http://fieldresearch.msf.org/msf/bitstream/10144/619251/1

Setting **priorities** for **humanitarian water**, **sanitation** and hygieneresearch: a meeting report MSF for **humanitarian water**, **sanitation** andhygiene research: a meeting report 2018, 12 (1) .....

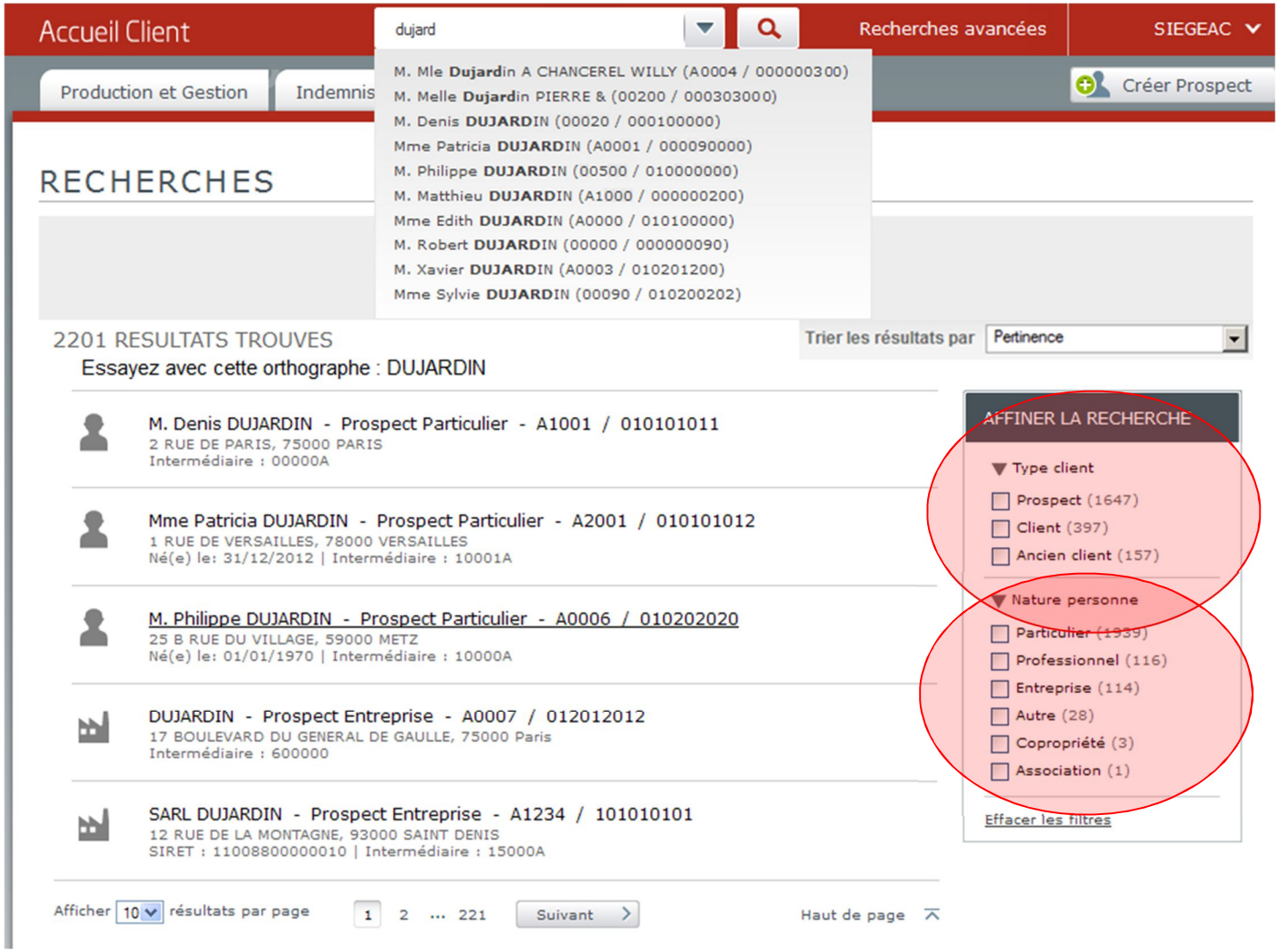
Microsoft Word - The Priorities UK New 2011.doc | the_priorities_uk_2011.pdf

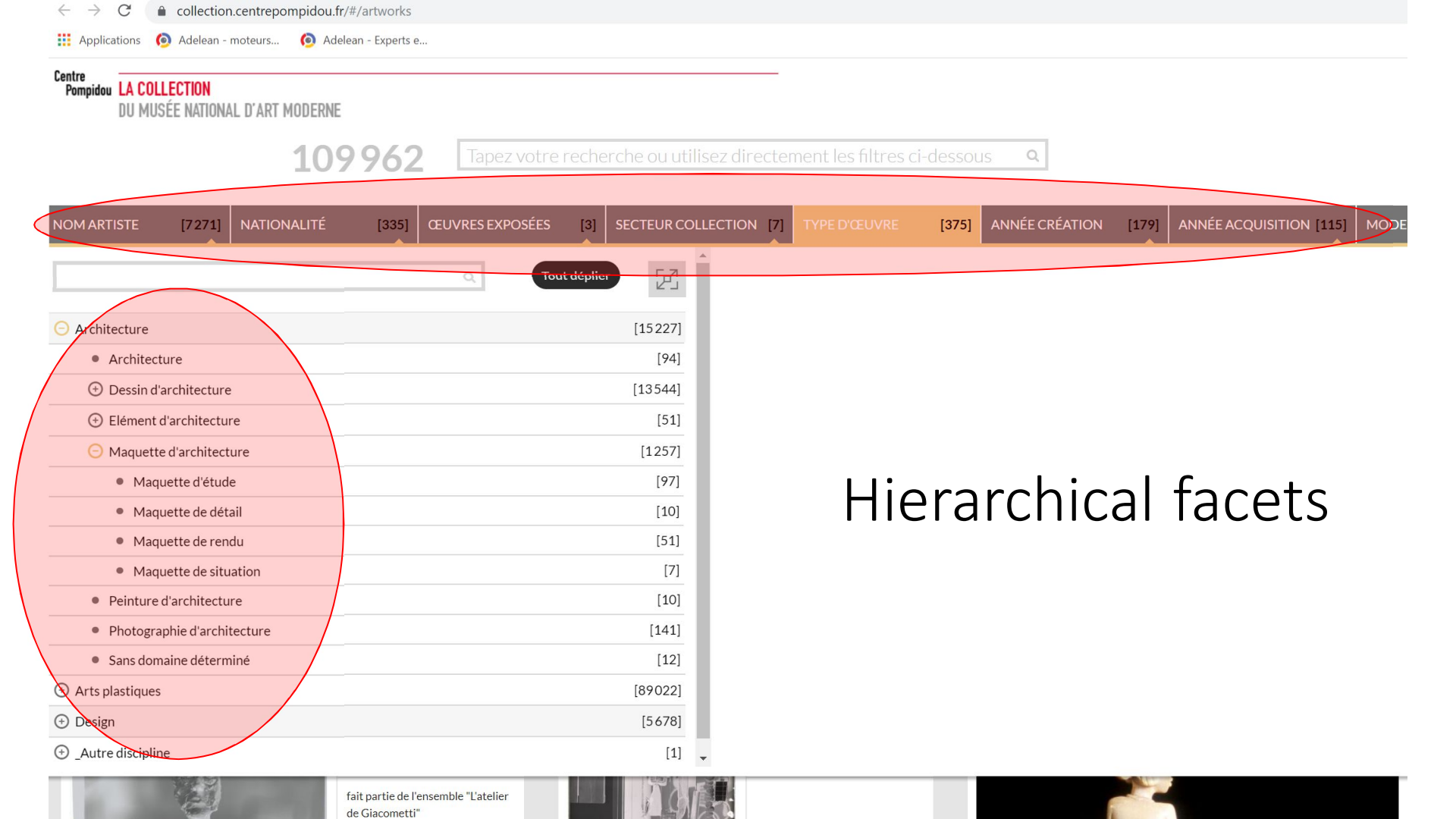
Last modified on 17/12/2018

https://evaluation.msf.org/sites/evaluation/files

Water, **hygiene and sanitation**.....9 4. Number, type and location of **water** points **Hygiene and sanitation**.....

Using lot quality assurance sampling to assess access to water, sanitation and hygiene services in a refugee camp setting in South Sudan: a feasibility study | Harding E et al - 2017 - Using lot quality assurance sampling


Accueil Client
dujard
Recherches avancées
SIEGEAC
M. Mle Dujardin A CHANCEREL WILLY (A0004 / 000000300)
M. Melle Dujardin PIERRE & (00200 / 000303000)
M. Denis DUJARDIN (00020 / 000100000)
Mme Patricia DUJARDIN (A0001 / 000090000)
M. Philippe DUJARDIN (00500 / 010000000)
M. Matthieu DUJARDIN (A1000 / 000000200)
Mme Edith DUJARDIN (A0000 / 010100000)
M. Robert DUJARDIN (00000 / 000000090)
M. Xavier DUJARDIN (A0003 / 010201200)
Mme Sylvie DUJARDIN (00090 / 010200202)
Production et Gestion
Indemnis
Créer Prospect
RECHERCHES
2201 RESULTATS TROUVES
Essayez avec cette orthographe : DUJARDIN
Trier les résultats par
Pertinence
M. Denis DUJARDIN - Prospect Particulier - A1001 / 010101011
2 RUE DE PARIS, 75000 PARIS
Intermédiaire : 00000A
Mme Patricia DUJARDIN - Prospect Particulier - A2001 / 010101012
1 RUE DE VERSAILLES, 78000 VERSAILLES
Né(e) le: 31/12/2012 | Intermédiaire : 10001A
M. Philippe DUJARDIN - Prospect Particulier - A0006 / 010202020
25 B RUE DU VILLAGE, 59000 METZ
Né(e) le: 01/01/1970 | Intermédiaire : 10000A
DUJARDIN - Prospect Entreprise - A0007 / 012012012
17 BOULEVARD DU GENERAL DE GAULLE, 75000 Paris
Intermédiaire : 600000
SARL DUJARDIN - Prospect Entreprise - A1234 / 101010101
12 RUE DE LA MONTAGNE, 93000 SAINT DENIS
SIRET : 11008800000010 | Intermédiaire : 15000A
AFFINER LA RECHERCHE
Type client
Prospect (1647)
Client (397)
Ancien client (157)
Nature personne
Particulier (1939)
Professionnel (116)
Entreprise (114)
Autre (28)
Copropriété (3)
Association (1)
Effacer les filtres
Afficher 10 résultats par page
1 2 ... 221
Suivant
Haut de page



Hierarchical facets

All Departments | bike

NextDay delivery | Delivering to 94066

1 - 40 of 1,000+ results

Refine by | Price | Top Brands | Store Availability

Sort by | Best Match

Departments

- Sports & Outdoors
  - Bikes
  - All Kids' Bikes
  - Cruiser Bikes
  - Kids' BMX Bikes
- See all Sports & Outdoo...
- See All Departments

Delivery & Pickup

- (•) Show all
- ( ) 2-day delivery
- ( ) Ship to home
- ( ) Free pickup + discount
- ( ) Free pickup
- ( ) Free pickup today

Bicycle Type

- [ ] BMX Bikes
- [ ] Comfort & Hybrid Bikes

Kids bike buying guide
See our guide for ride on toys for kids of all ages.

ROLLBACK

Featured Item

Bonus Helmet and Accessories with Mongoose Excursion Mountain Bike!

★★★★★

· Buy a Mongoose Excursion mountain bike and receive a helmet, lock, and cargo bag!

$148.00

Free delivery
Free pickup

Men's Bikes

Women's Bikes

Kids Bikes

Rollback

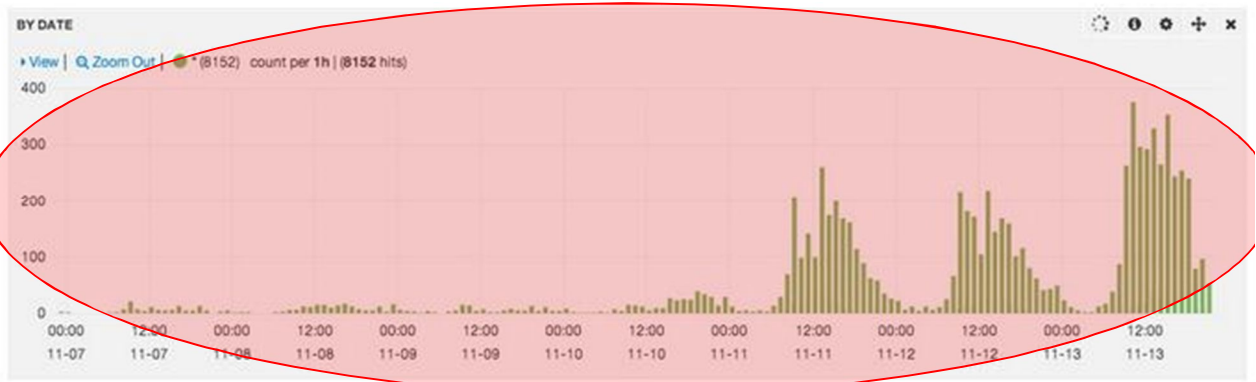
All Bike

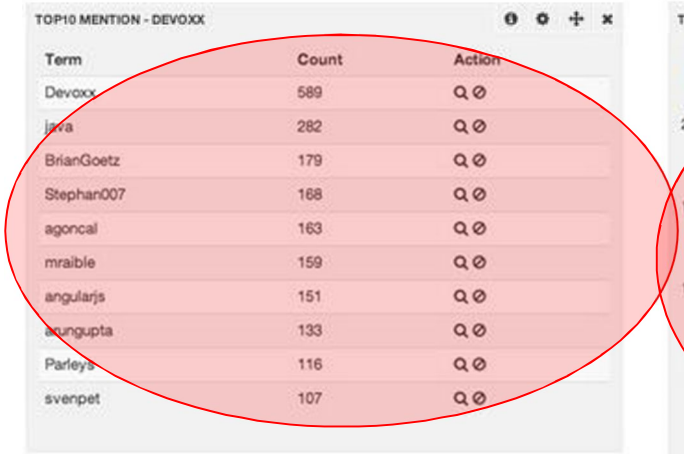
BEST SELLER

BEST SELLER

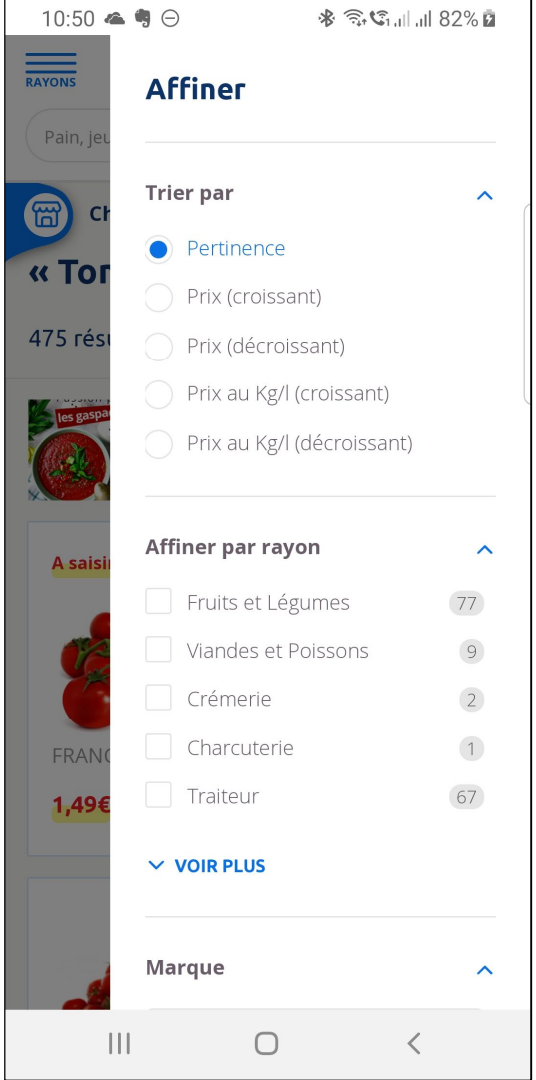
BEST SELLER

Other “exotic” facets

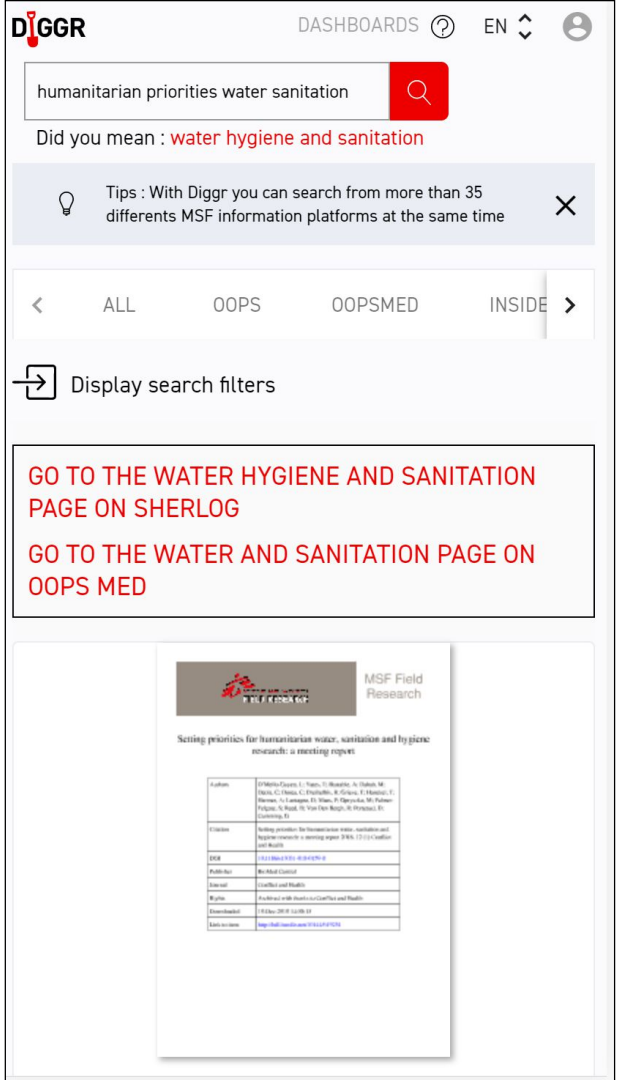


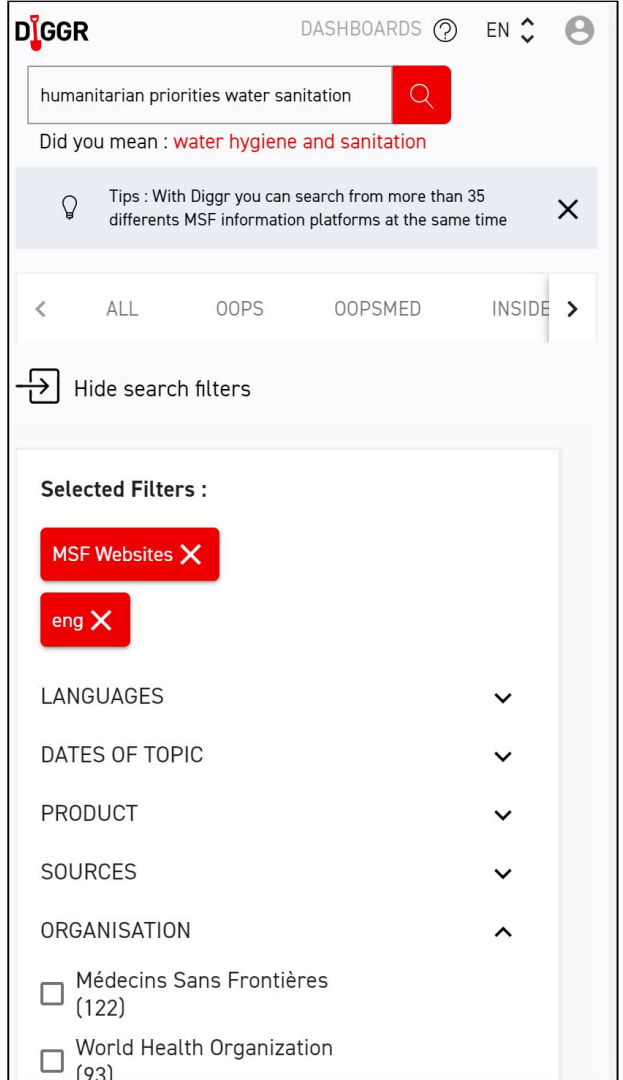


Facets on a mobile device

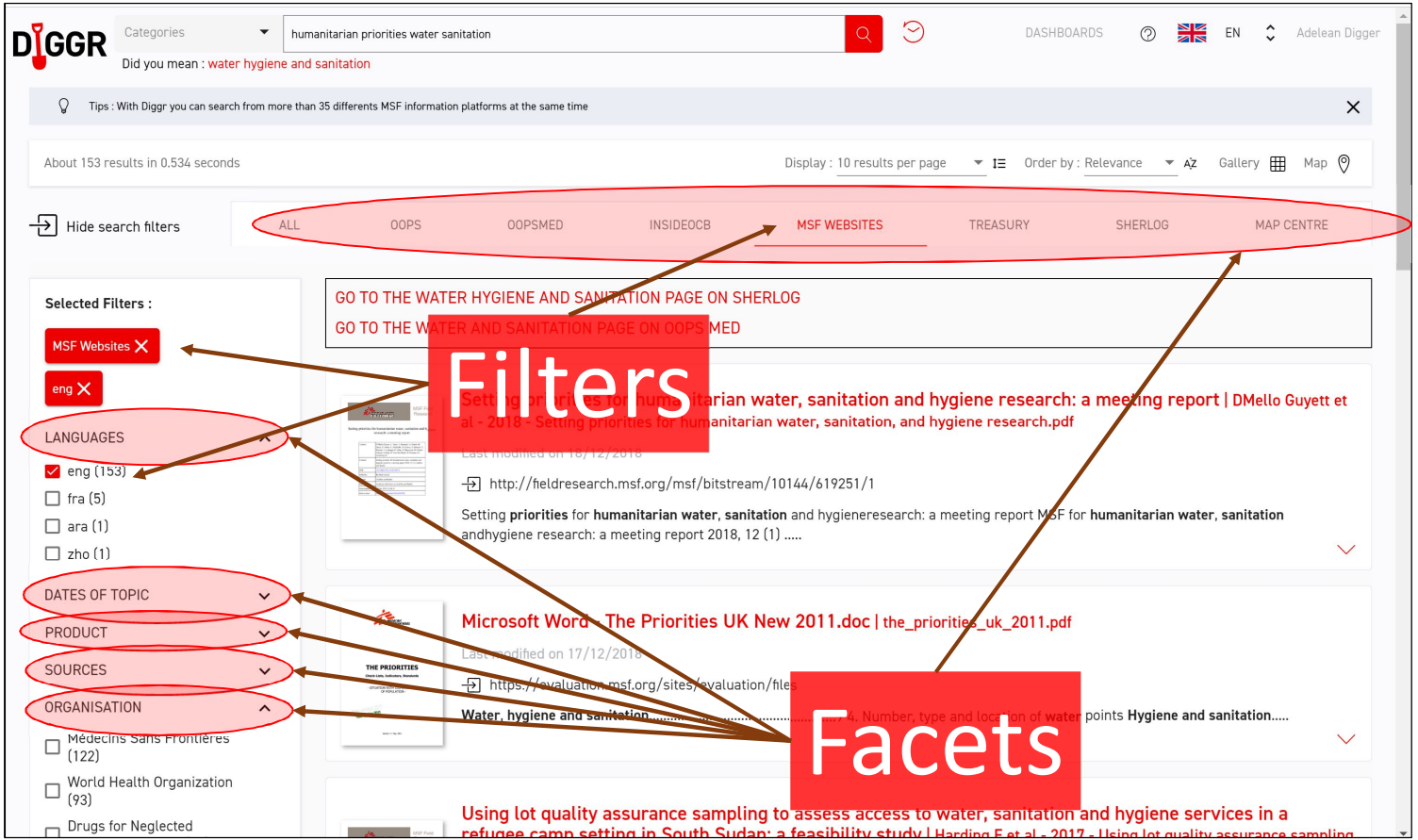


Facets on a mobile device





// Facets and Filters

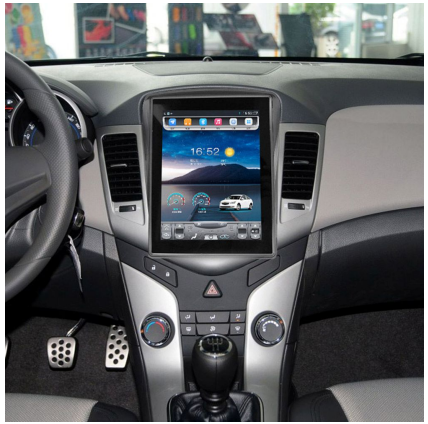


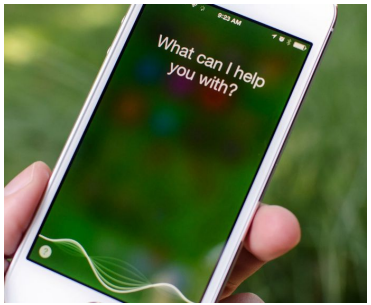
// Why are facets important ?

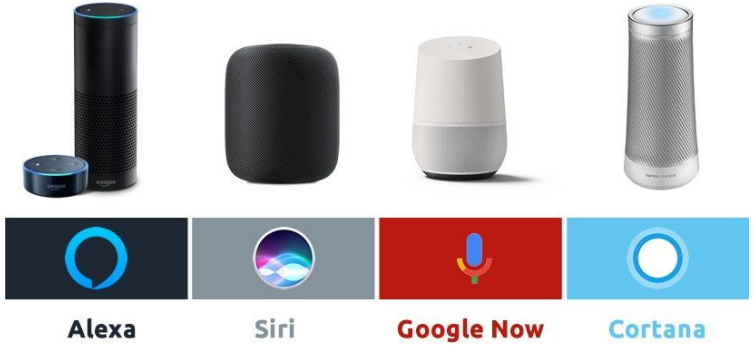
VOICE + SCREEN
(multimodal)

- More and more data and less and less space to display it
- New ways of searching: voice, assistants, chat bots

VOICE ONLY







// How are facets implemented ?

Facets are a standard feature of modern search engines.

Apache Lucene has great support for everything around facets

- Solr : field value faceting, range faceting, pivot faceting, interval faceting, block join faceting, …
- Elasticsearch : aggregations, sub-aggregations, top hits aggregation, histogram aggregation, range aggregations, geo aggregations, …

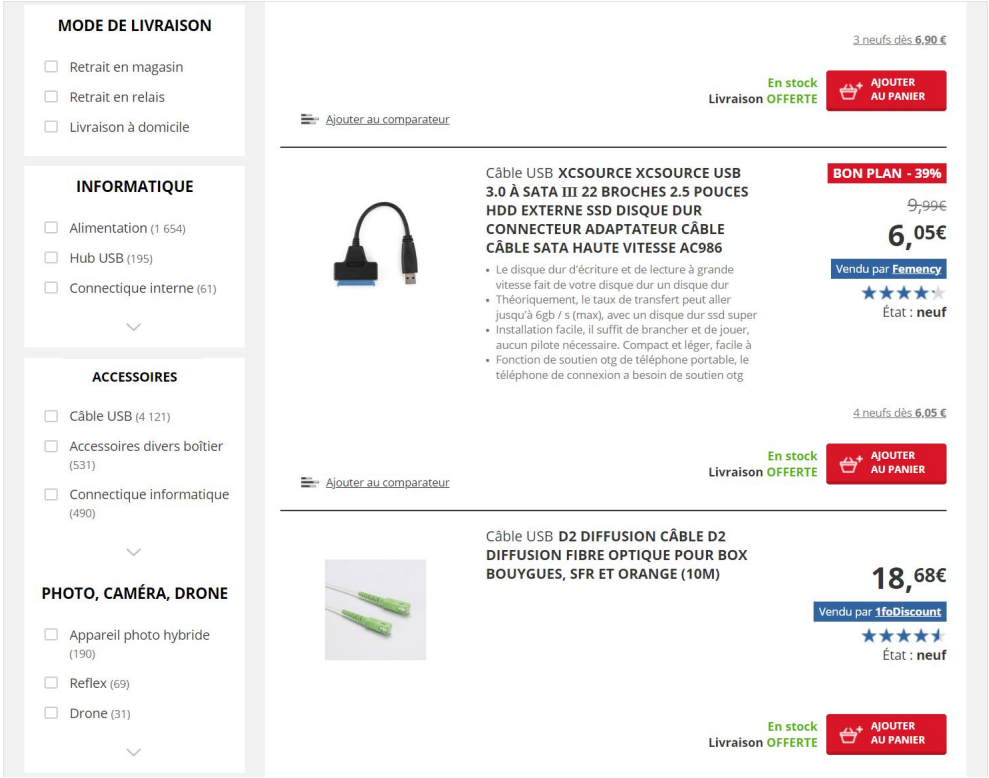
The User Experience with facets and the way they are "displayed" can be very diverse

// Structure of the talk

- Examples of facet implementations
- Challenges with facets and possible solutions
- Challenges with search in general and how facets can help
- Technical implementation examples are with Elasticsearch
- We are addressing less the "graphical" display of facets and more the technical issues with their relevancy

// Challenge #1: marketplaces

- Issue: the heterogeneity of results and the number of candidate facets



Heterogeneity of results: Solution 1

Facets based on top N results:

- Fetch the top N results (first page + a few of the next ones)
- Retain only the facets applicable to these top N results

Implementation details:

- First query: query term
- Fetch the first N document ids (let's say max 1024)
- Second query : terms filter on document ids and aggregations

Heterogeneity of results: Solution 2

- Modeling with a single facet-name / facet-value field tuple and the nested type
- Need to treat differently strings, numbers and booleans

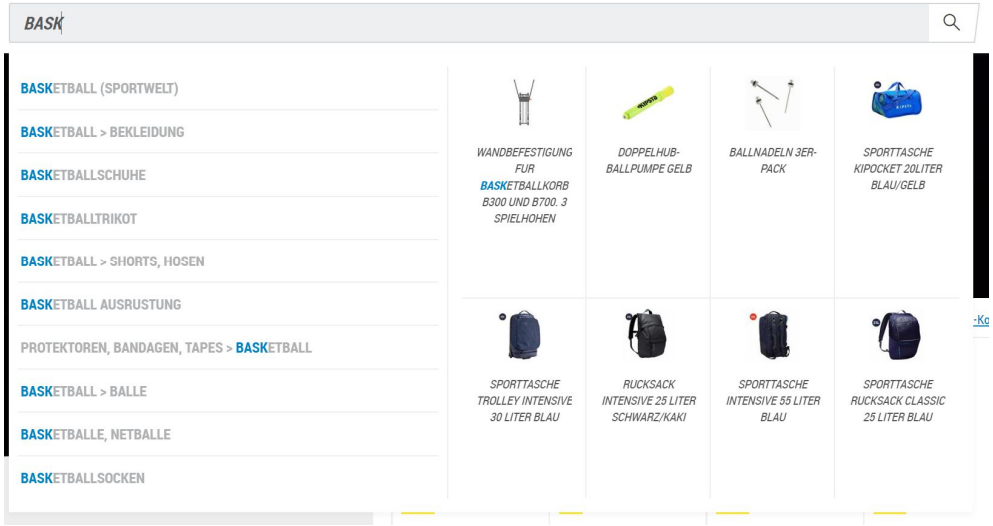
```
"string_facets": {
  "type": "nested",
  "properties": {
    "facet-name": {
      "type": "keyword"
    },
    "facet-value": {
      "type": "keyword"
    }
  }
},
"numeric_facet" : {
  "type": "nested",
  "properties": {
    "facet-name": {
      "type": "keyword"
    },
    "facet-value": {
      "type": "double"
    }
  }
}
```

```
"string_facets": [
  {
    "facet-name": "memory",
    "facet-value": "64 MB"
  },
  {
    "facet-name": "OS_type",
    "facet-value": "android"
  }
],
"numeric_facet": [
  {
    "facet-name": "price",
    "facet-value": 599
  },
  {
    "facet-name": "screen_size",
    "facet-value": 5,7
  }
]
```

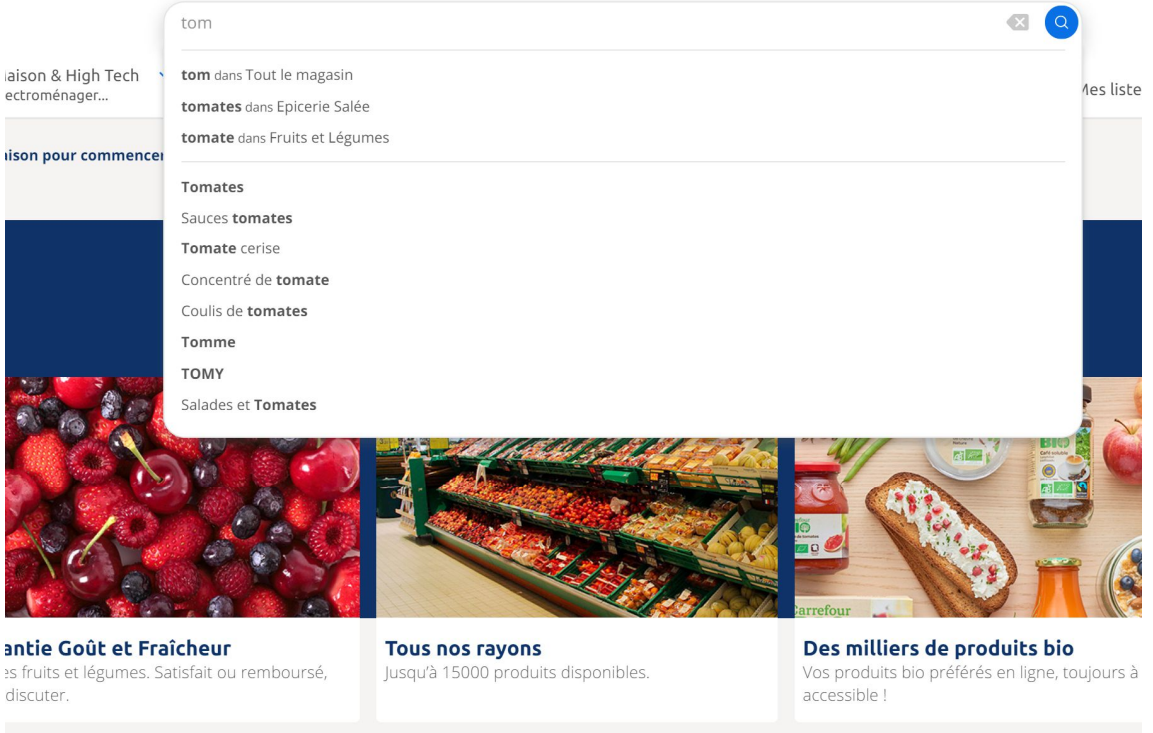
Heterogeneity of results: Solution 2 – the query

```
{
    "aggs": {
        "agg_string_facet": {
            "nested": {
                "path": "string_facets"
            },
            "aggs": {
                "facet_name": {
                    "terms": {
                        "field": "string_facets.facet-name"
                    },
                    "aggregations": {
                        "facet_value": {
                            "terms": {
                                "field": "string_facets.facet-value"
                            }
                        }
                    }
                }
            }
        },
        "agg_numeric_facet": {
            "nested": {
                "path": "numeric_facet"
            },
            "aggs": {
                "facet_name": {
                    "terms": {
                        "field": "numeric_facet.facet-name"
                    },
                    "aggs": {
                        "facet_value": {
                            "stats": {
                                "field": "numeric_facet.facet-value"
                            }
                        }
                    }
                }
            }
        }
    }
}
```

// Challenge #2: auto-completion



// Challenge #2: auto-completion



Auto-completion: solution

Products index

```
{
    "title": "Red tomatoes",
    "category": "Tomates",
    "department": "Fruits et Légumes",
    "price": "3"
}

{
    "title": "Yellow tomatoes",
    "category": "Tomates",
    "department": "Bio et Ecologie",
    "price": "30"
}
```

Suggestions index

```
{
    "text": "Tomates",
    "context": {
        "department": ["Fruits et Légumes"]
    }
}

{
    "term": "Tomates",
    "context": {
        "department": ["Bio et Ecologie"]
    }
}
```

Use the Update API here
and also increase the number of occurrences

Auto-completion: solution

The "Suggestions" index

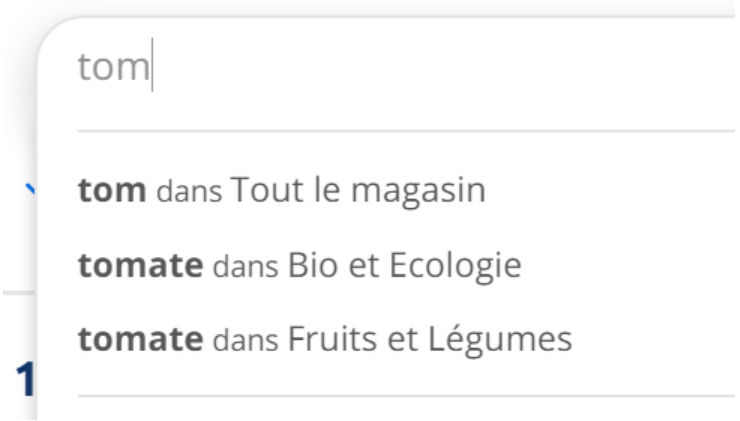
```
{
    "text": "Tomates",
    "context": {
        "product_categories_name": [
            "Fruits et Légumes",
            "Bio et Ecologie"
        ]
    },
    "occurrences": 81,
    "popularity": 16
}
```

The query

```
{
  "query": {
    "multi_match": {
      "query": "tom",
      "fields": ["text.smartComp"]
    }
  },
  "sort": [
    {
      "occurrences": {
        "order": "desc"
      }
    }
  ]
}
```

Auto-completion: solution

The result



Auto-completion: solution

The shortcut

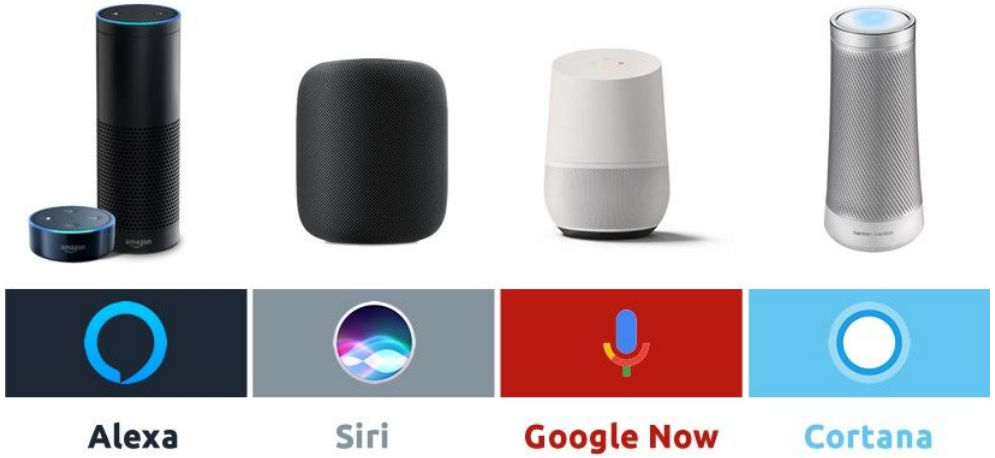
```
GET _a2/autocomplete
{
  "text": "tom"
}
```



```
{
    "engineTimeInMilis": 1,
    "totalHits": 1478,
    "product": [...],
    "suggestions": [
        {
            "text": [
                "tom"
            ],
            "highlight": [
                "<em>tom</em>"
            ],
            "type": "input"
        },
        {
            "text": [
                "Tomates"
            ],
            "highlight": [
                "<em>Tomates</em>"
            ]
        },
        ...
    ],
    "filteredsuggestions": [
        {
            "text": "Tomates",
            "highlight": "<em>Tomates</em>",
            "facetFilter": {
                "product.categories.name": "Bio et Ecologie"
            }
        },
        {
            "text": "Tomates ",
            "highlight": "<em>Tomates</em>",
            "facetFilter": {
                "product.categories.name": "Fruits et Légumes"
            }
        },
        ...
    ]
}
```

// Challenge #3: assistants

- Often the first responses of an assistant are suggestions for additional filters that refine the query.



Simulator

A2

Talk to A2

Alright. Here's the test version of ei two.

Welcome! What are you looking for today?

Chocolate

I found 920 products for you, what is your favorite brand?

Suggested input

Carrefour | Carrefour Bio | Lindt | Milka | Côte d'Or | cancel

Input

Try typing or saying "Talk to A2".

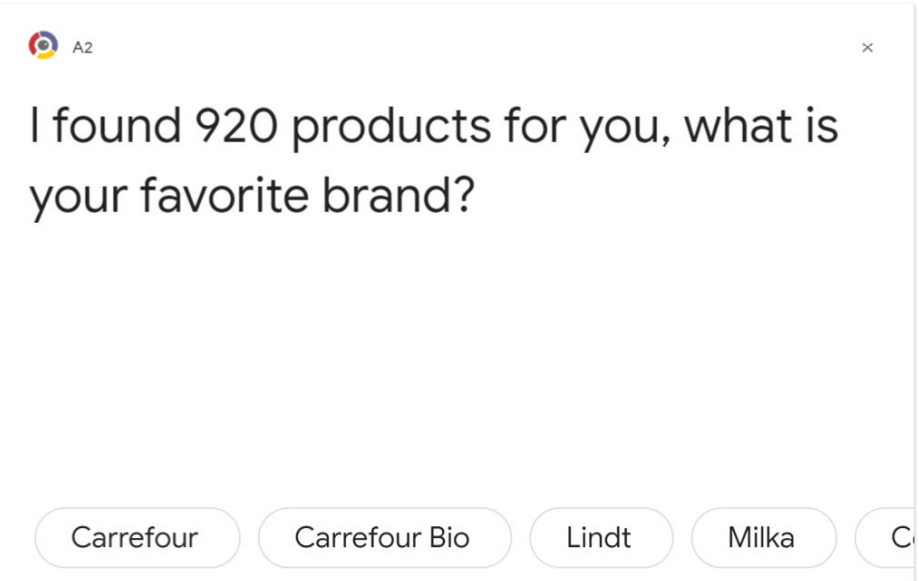
Change version

Surface: Smart Displ...

Language: English (United States)

Location: Schönwalder Str. 21, 13347 Berlin, Allem...

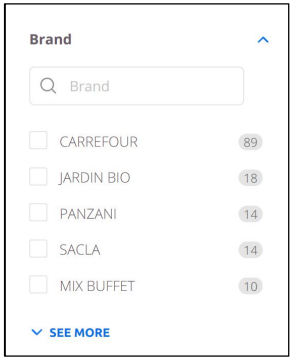
DISPLAY REQUEST RESPONSE AUDIO DEBUG ERRORS

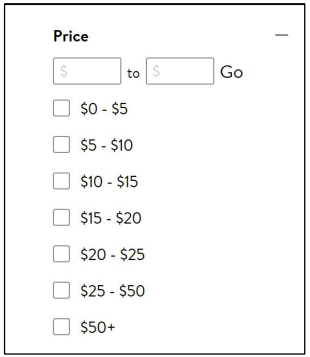


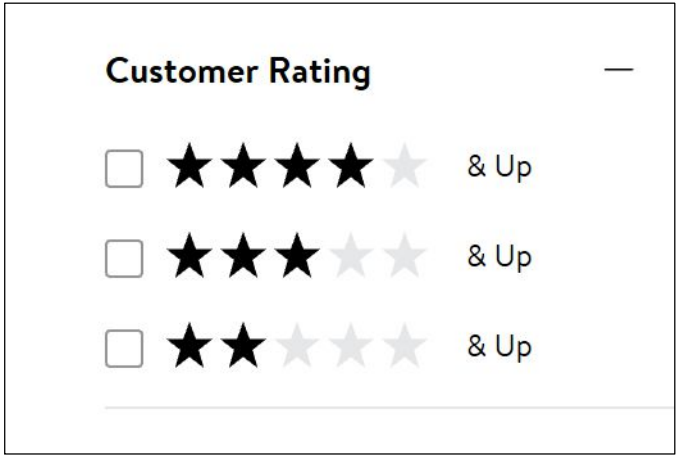
How to narrow ?

- Often the first responses of an assistant are suggestions for additional filters that refine the query

- “Quick win” solution :
  - Filters
- Issue :
  - Which facets to choose?
- Prerequisite:
  - Your search engine should already have relevant filters

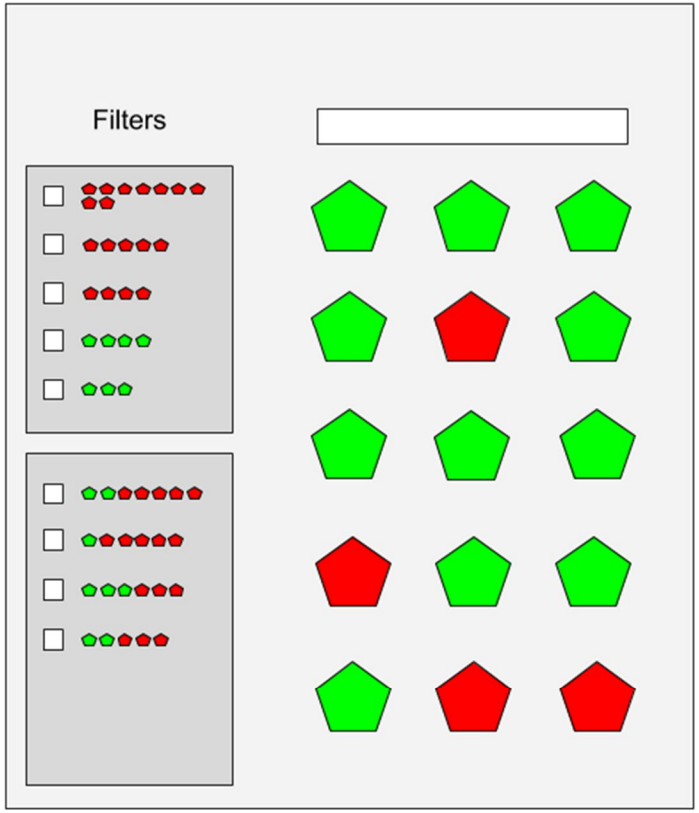


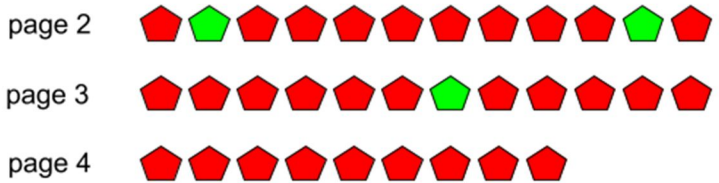




// Challenge #4: relevant facet values

- Issue: how to make facet values relevant in the context of many "less relevant" results ?





Carrefour

sucre

Rayons
Promo, épicerie...

Maison & High Tech
Electroménager...

Services
Voyage, spectacle...

Trouver un magasin

Mes listes

Me connecter

Marque

Marques...

- CARREFOUR 106
- TROPICANA 42
- GERBLE 36
- KARELEA 27
- JOKER 23

VOIR PLUS

Préférences alimentaires

- Sans sucres ajoutés 325
- Sans sucres 100
- Végétarien 54
- De saison 5
- Halal 2

1,28€

0,67€

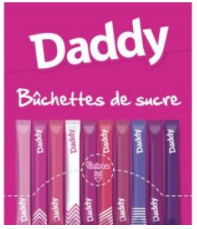
1,13€

1,25€

Sucre vanillé VAHINE
les 10 sachets de 7,5 g
26.40 € / Kilogramme

Sucre en poudre DADDY
la boite de 1 kg
1.34 € / Kilogramme

Sucre en poudre DADDY
la boite de 100 bûchettes - 500 g
4.42 € / Kilogramme



Sucre roux GRAEFFE
le paquet de 1 kg
2.14 € / Kilogramme

Carrefour

sucre

Rayons
Promo, épicerie...

Maison & High Tech
Electroménager...

Services
Voyage, spectacle...

Trouver un magasin

Mes listes

Me connecter

Marque

Marques...

- CARREFOUR 8
- BEGHIN SAY 15
- DADDY 9
- SAINT LOUIS 9
- CANDEREL 5

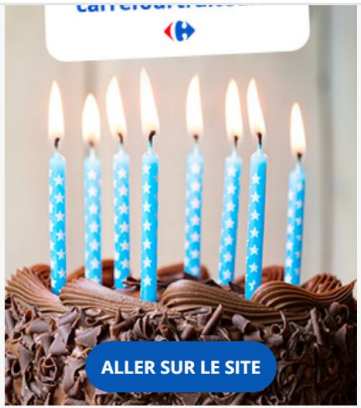
VOIR PLUS

Préférences alimentaires

- Végétarien 23

Labels qualité

- Bio 14
- Commerce équitable 5



0,74€

0,69€

0,99€

Sucre en poudre DADDY
la poche de 900 g
1.34 € / Kilogramme

Sucre Cristal ERSTEIN
le paquet de 1 kg
1.28 € / Kilogramme

Sucre roux morceaux LA PERRUCHE
la boite de 168 morceaux - 1 kg
1.78 € / Kilogramme

Sucre demi-morceaux DADDY
le paquet de 750 g
1.51 € / Kilogramme

Relevant facet values: the solution

Solutions: work on your search precision

Analytics and data science have clues: for instance, when clients type “tomato”, is there a category which regroup most of the clicks ?

All you must do is prefilter some facets (or even all the results) with this category : 80% of the result set will disappear and your filters will look good !

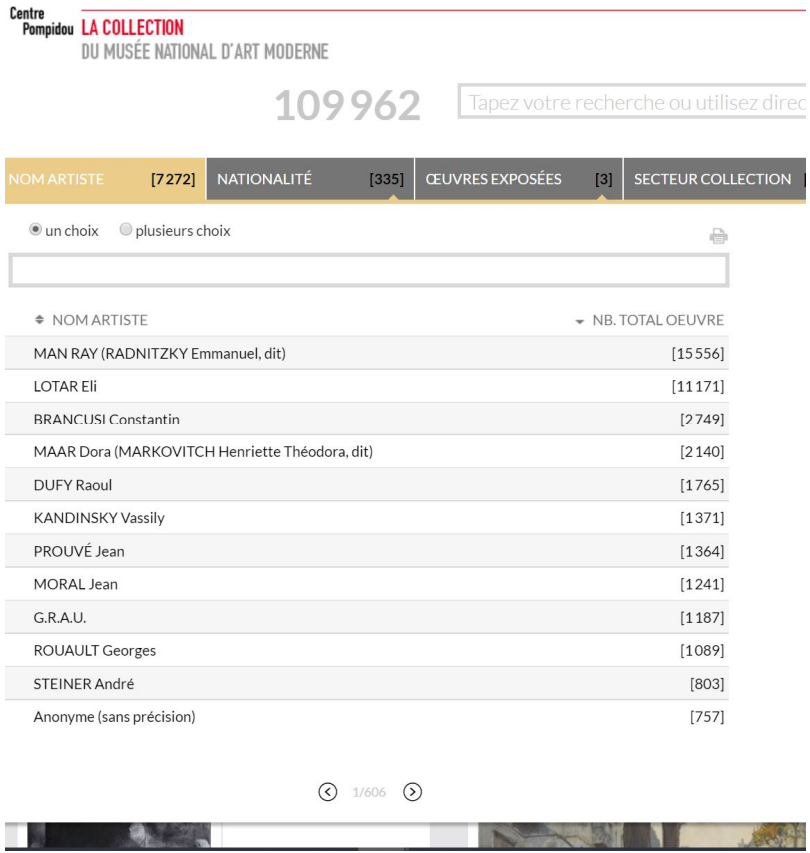
Examples of prefiltering at Carrefour:

- 11% of results for “tomatos” are in the “Fresh vegetables” category but they represent 86% of products added to basket
- 24% of results for “rice” are in the “Pasta and Rice” category and represent 90% of purchases
- 8% of results for “sugar” are in the “Sugar and sweeteners” category and represent 90% of purchases

// Challenge #5: search in facet values

- Issue: How to bring up facet values beyond the first top N values ?

- Solutions:
  - Pagination
  - Search in Search



CATÉGORIE

< Image - Son

Téléviseur

- TV 8K
- TV 4K UHD
- TV OLED
- TV QLED
- Smart TV
- TV LED
- TV reconditionné

MARQUE

Rechercher

- Essentielb
- Grundig
- Hisense
- LG
- Listo

TAILLE DE L'ÉCRAN (DIAGON...

- 80 pouces et plus (> 203 cm)
- 75 pouces (189 cm)
- 65 pouces (163 cm)
- 55 pouces (140 cm)
- 49 pouces (123 cm)
- 40 pouces (100 cm)
- 32 pouces (80 cm)
- 28 pouces (70 cm)

Comparer

TV LED **Essentielb** 65UHD-1291-SMART

Jusqu'à 10€ de remise pour l'achat d'une Chromecast*

Automnez-moi !

- 165 cm (65") - 4K UHD
- LxHxP (sans pied) 145.5 x 87.6 x 8.5 cm
- Son 2 x 10 Watts - HDMI x3 - USB x2
- **Le + :** Wifi

~~699,99€~~ - 14%

599,99€

AJOUTER AU PANIER

✓ Retrait magasin **dès mardi 29 octobre** à **Velizy 2** et à Les Clayes sous Bois "Alpha" (1) - offert
Autres magasins

✓ Livraison chez vous **dès mardi 29 octobre** - offerte (2)
Détails

Comparer

TV LED **Sony** KD49XG8096 Android TV

4.5 (7 avis)

Automnez-moi !

- 123 cm (49") - 4K UHD
- TV connectée Android TV
- LxHxP (sans pied) 110.1 x 64.5 x 5.7 cm
- Son 2 x 10 Watts - HDMI x4 - USB x3
- **Le + :** Android TV, google assistant, design aux bords fins, passage de câb... intégré au pied

~~949,00€~~ - 26%

699,00€

AJOUTER AU PANIER

✓ Retrait magasin **dès demain** à **Velizy 2** et à Les Clayes sous Bois "Alpha" (1) - offert
Autres magasins

✓ Livraison chez vous **dès mardi 29 octobre** - offerte (2)
Détails

TV OLED **Sony** KD55AG8 Android TV

5 (5 avis)

~~2490,00€~~ - 20%

1990,00€

AJOUTER AU PANIER

Search in facet values: implementation with Elasticsearch

```
GET vm-dev_370_2019-07-19_17-18-artworks/_search
{
  "size": 0,
  "aggregations": {
    "source.authors.nationality": {
      "filter": {
        "bool": {
      },
      "aggregations": {
        "source.authors.nationality.terms": {
          "terms": {
          "aggregations": {
            "details": {
              "top_hits": {
              }
            }
          },
          "source.authors.nationality.total": {
            "cardinality": {
              "field": "source.authors.nationality.keyword",
              "precision_threshold": 40000
            }
          }
        }
      }
    }
  }
}
```

Search in facet values: details of the filter aggregation

```
"filter": {
  "bool": {
    "should": [
      {
        "multi_match": {
          "query": "fra",
          "fields": [
            "source.authors.nationality.comp^1.0"
          ],
          "type": "phrase",
          "operator": "AND",
          "slop": 2,
          "prefix_length": 0,
          "max_expansions": 50,
          "lenient": false,
          "zero_terms_query": "NONE",
          "boost": 1
        }
      }
    ],
    "disable_coord": false,
    "adjust_pure_negative": true,
    "minimum_should_match": "1",
    "boost": 1
  }
},
```

Search in facet values: details of the terms sub-aggregation

```
4    "aggregations": {
5      "source.authors.nationality": {
6        "filter": {
32       "aggregations": {
33         "source.authors.nationality.terms": {
34           "terms": {
35             "field": "source.authors.nationality.keyword",
36             "size": 800,
37             "min_doc_count": 1,
38             "shard_min_doc_count": 0,
39             "show_term_doc_count_error": false,
40             "order": [
41               {
42                 "_count": "desc"
43               },
44               {
45                 "_term": "asc"
46               }
47             ]
48           },
49           "aggregations": {
50             "details": {
51               "top_hits": {
102            }
103          }
104        },
```

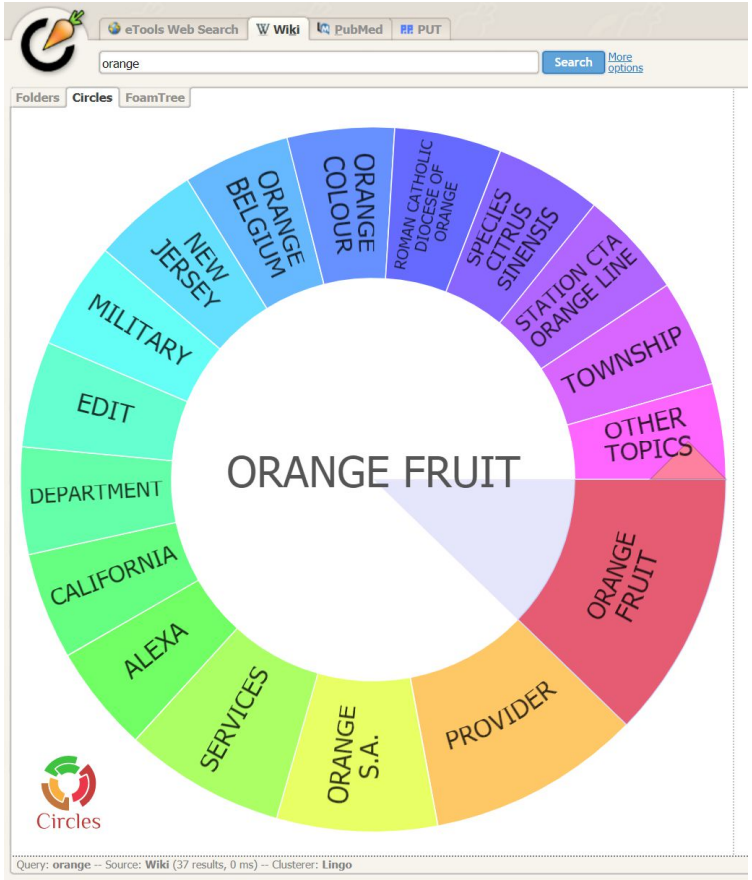
Search in facet values: details of the top_hits sub-aggregation and highlighting

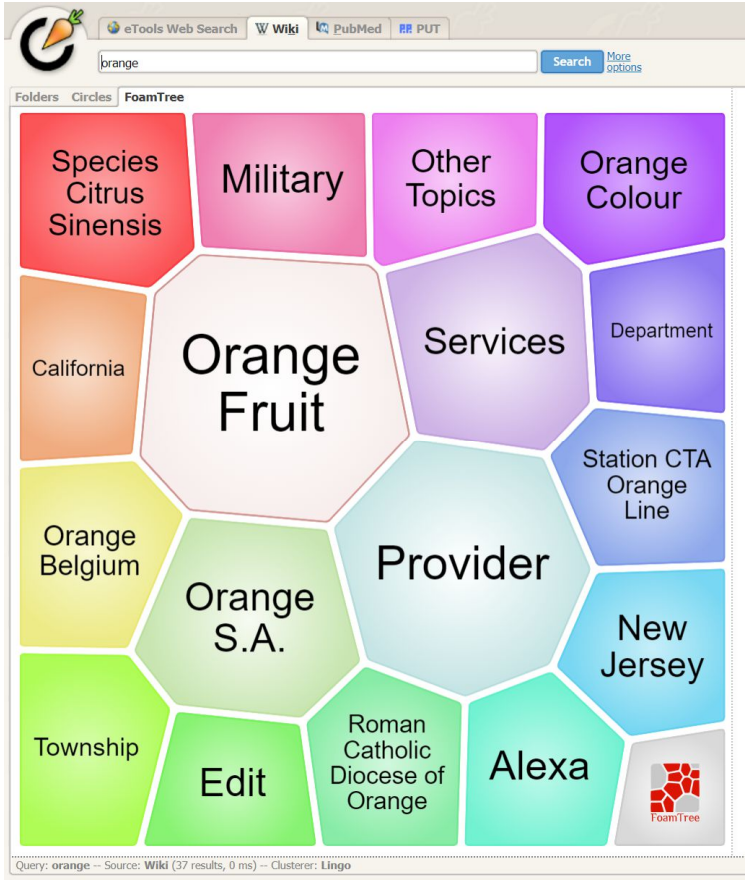
```
            "top_hits": {
              "from": 0,
              "size": 1,
              "version": false,
              "explain": false,
              "_source": {
                "includes": [
                  "source.authors.nationality"
                ],
                "excludes": []
              },
              "highlight": {
                "pre_tags": [
                "post_tags": [
                "highlight_query": {
                  "bool": {
                    "should": [
                      {
                        "multi_match": {
                          "query": "fra",
                          "fields": [
                            "source.authors.nationality.comp^1.0"
                          ],
                          "type": "phrase",
                          "operator": "AND",
                          "slop": 2,
                          "prefix_length": 0,
                          "max_expansions": 50,
                          "lenient": false,
                          "zero_terms_query": "NONE",
                          "boost": 1
                        }
                      }
                    ],
                    "disable_coord": false,
                    "adjust_pure_negative": true,
                    "minimum_should_match": "1",
                    "boost": 1
                  }
                },
                "fields": {
                  "source.authors.nationality.comp": {
                    "number_of_fragments": 0
                  }
                }
              }
            }
```

// Challenge #6: unstructured data

- Issue: the lack of structure makes difficult to suggest additional query refinements
- Solutions:
  - Clustering (like http://project.carrot2.org/)
  - Entity extraction (like https://www.basistech.com/text-analytics/rosette/entity-extractor/ or https://twitter.com/dep4b/status/1121141764503609345)

Display “facets" with clustering





http://project.carrot2.org/

Enrich the data with entity extraction

Haystack is the conference for improving search relevance. If you're like us, you work to understand the shiny new tools or dense academic papers out there that promise the moon. Then you puzzle how to apply those insights to your search problem, in your search stack. But the path isn't always easy, and the promised gains don't always materialize.

Haystack is the conference for organizations where search, matching, and relevance really matters to the bottom line. For search managers, developers, relevance engineers & data scientists finding ways to innovate, see past the silver bullets, and share what actually has worked well for their unique problems. Please come share and learn!

https://haystackconf.com/

Conference: Haystack

Domain: search

Facets on unstructured text after entity extraction

DIGGR | Categories | msf | DASHBOARDS | EN | Adelean Digger

About 24.368 results in 0.244 seconds | Display : 10 results per page | Order by : Relevance | Gallery | Map

Hide search filters | ALL | OOPS | OOPSMED | MSF WEBSITES | UNIDATA | MAP CENTRE | TREASURY | SHERLOG | NESTOR | MYHELP | ONEDRIVE

Selected Filters :

MSF Websites ✕

- DATES OF TOPIC
- PRODUCT
- SOURCES
- LANGUAGES
- ORGANISATION
- PERSON
- LOCATION
  - ☐ United States of America (12415)
  - ☐ Africa (12402)
  - ☐ France (12257)
  - ☐ Switzerland (11628)
  - ☐ Democratic Republic of the Congo (11487)
  - ☐ United Kingdom (11472)

Microsoft Word - 090810 chagas policy doc FINAL | MSF DNDi Chagas Policy Doc 090810.pdf

Last modified on 17/12/2018

https://www.dndi.org/wp-content/uploads/2009/10

population: 8-15 million people · 14,000 Deaths / Year · Largest parasitic killer in the Americas **Medecins Sans Frontieres/Doctors Without Borders (MSF)** and Drugs for Neglected Dise.....

MIM MSF-DNDi SMC symposium at MIM_leaflet | MSF-DNDi _SMCsymposium_MIM.pdf

Last modified on 17/12/2018

https://www.dndi.org/wp-content/uploads/2013/07

Médecins Sans Frontières/ **Doctors Without Borders (MSF)** & the Drugs for Neglected 6.30 pm Place: Hall 4B Photo: Patients queue in Moïssala District's Bouna health zone, where **MSF** Prof Chris Drakeley, LSHTM, UK SMC: The exper.....

Microsoft Word - 102019 joint msf dndi briefing paper format_v3.doc | msf dndi briefing paper_Feb2010.pdf

Last modified on 17/12/2018

https://www.dndi.org/wp-content/uploads/2010/02

Doctors Without Borders/Médecins Sans Frontières (**MSF**) and the Drugs for Neglected Diseases initiative

// Conclusions and takeaways

More data, less space → Facets are more and more important

In order to be useful → Facets should be relevant

Modern search engines have great support for facets

// Conclusions and takeaways

Marketplaces	When too many possible facets → the relevant ones should be driven by the most relevant results
Auto-completion	Use facet values as suggestions and disambiguation techniques
Assistants	When too many results → chose the facet and filter suggestions that disambiguate most as the first answer
Relevant facet values	When there is a risk of noise in the results → avoid bringing it to facet values
Search in facet values	When too many facet values → bring up those beyond the top N with search (not with JavaScript
Unstructured data	Use clustering and entity extraction to be able to define facets

Thank You !

- Lucian Precup
- Radu Pop
- @lucianprecup
- @a2lean
- #haystackconf
- @o19s
- Berlin EU 2019

adelean
EXTRACT TRANSFORM SEARCH

HAYSTACK