AlvisIR 2

Robert Bossy

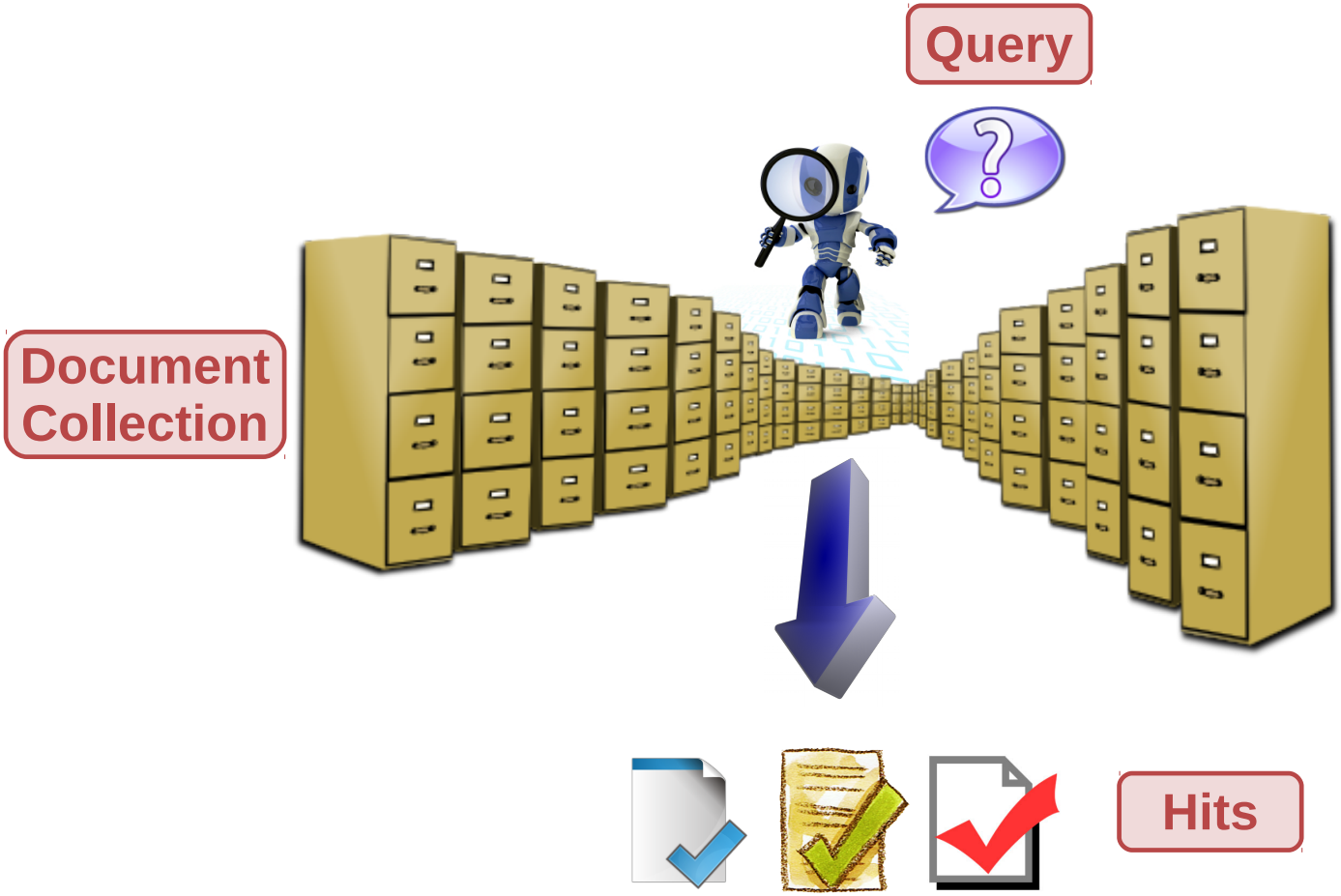
Mathématique Informatique et Génome – Bibliome
Institut National de la Recherche Agronomique

27 Novembre 2013 / Bibliome

Introduction

. . . to search engines

What is a search engine?



Functions of a search engine

- Find documents that match the user query (hits).
- Show the hits.
- Explain why hits matched.

Other functions

- Summary of hits.
- Help the user to compose queries.
- Give access to other services.

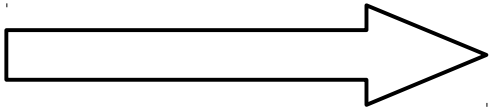
Document vector

Vibrio vulnificus
Vibrio vulnicus causes potentially fatal food poisoning Vibrio vulnificus is a a lactose-fermenting, halophilic, Gram-negative , opportunistic pathogenic bacterium from the same family as those that cause cholera. It normally lives in warm seawater and is part of a group of vibrios that are called "halophilic" because they are salt requiring organisms. This organism causes wound infections, gastroenteritis, or a syndrome known as primary septicemia. Found in warm coastal waters, this bacterium is related to the cholera pathogen and can cause a severe and potentially fatal illness. Infections tend to occur through eating raw or improperly cooked shellfish, particularly oysters. The ingestion of V. vulnificus by healthy individuals can result in gastroenteritis. The "primary septicemia" form of the disease can follow. Wound infections result either from contaminating an open wound with sea water harboring the organism, or by lacerating part of the body on coral, fish, etc., followed by contamination with the organism. Persons who are immunocompromised, especially those with chronic liver disease, are more at risk from Vibrio vulnificus. There is no evidence for person-to-person transmission.

a
an
and
are
as
at
bacterium
because
body
by
called
can
cause
causes
cholera
chronic
coastal
contaminating
contamination
cooked
coral
Person
person
...

Normalized document vector

Vibrio vulnificus
Vibrio vulnicus causes potentially fatal food poisoning Vibrio vulnificus is a a lactose-fermenting, halophilic, Gram-negative , opportunistic pathogenic bacterium from the same family as those that cause cholera. It normally lives in warm seawater and is part of a group of vibrios that are called "halophilic" because they are salt requiring organisms. This organism causes wound infections, gastroenteritis, or a syndrome known as primary septicemia. Found in warm coastal waters, this bacterium is related to the cholera pathogen and can cause a severe and potentially fatal illness. Infections tend to occur through eating raw or improperly cooked shellfish, particularly oysters. The ingestion of V. vulnificus by healthy individuals can result in gastroenteritis. The "primary septicemia" form of the disease can follow. Wound infections result either from contaminating an open wound with sea water harboring the organism, or by lacerating part of the body on coral, fish, etc., followed by contamination with the organism. Persons who are immunocompromised, especially those with chronic liver disease, are more at risk from Vibrio vulnificus. There is no evidence for person-to-person transmission.



a
an
and
are
as
at
bacterium
because
body
by
called
can
cause
causes
cholera
chronic
coastal
contaminating
contamination
cooked
coral
Person
person
...

Enriched document vector

Vibrio vulnificus
Vibrio vulnicus causes potentially fatal food poisoning Vibrio vulnificus is a a lactose-fermenting, halophilic, Gram-negative , opportunistic pathogenic bacterium from the same family as those that cause cholera. It normally lives in warm seawater and is part of a group of vibrios that are called "halophilic" because they are salt requiring organisms. This organism causes wound infections, gastroenteritis, or a syndrome known as primary septicemia. Found in warm coastal waters, this bacterium is related to the cholera pathogen and can cause a severe and potentially fatal illness. Infections tend to occur through eating raw or improperly cooked shellfish, particularly oysters. The ingestion of V. vulnificus by healthy individuals can result in gastroenteritis. The "primary septicemia" form of the disease can follow. Wound infections result either from contaminating an open wound with sea water harboring the organism, or by lacerating part of the body on coral, fish, etc., followed by contamination with the organism. Persons who are immunocompromised, especially those with chronic liver disease, are more at risk from Vibrio vulnificus. There is no evidence for person-to-person transmission.

bacterium
because
body
call
cause
cholera
chronic
coastal
contamin
cook
coral
person
Bacteria:<Vibrio vulnificus>
Habitat:<food>
Habitat:<seawater>
Habitat:<wound>
...

Inverted index

Example

adherent	BTID-60171 BTID-60091 BTID-60170 BTID-60276 BTID-20330 BTID-10620 BTID-60619 BTID-60461
extraction	BTID-60338 BTID-20312 BTID-60637 BTID-50051 BTID-60561 BTID-60018 BTID-50013 BTID-60410
container	BTID-60366 BTID-60593
medicinal	BTID-60037 BTID-60356 BTID-60593
filter	BTID-60576 BTID-60143 BTID-60587 BTID-60262 BTID-60379 BTID-10227
trophozoites	BTID-60333 BTID-60536 BTID-60211 BTID-60066 BTID-60262 BTID-60263 BTID-60485 BTID-60049

- An *inverted index* is the collection of documents that contain each *term* (word).
- Allows to find documents that match each query term.
- *Posting*: term – document pair.

Composite queries

"x `AND` y"	$D_x \cap D_y$
"x `OR` y"	$D_x \cup D_y$
"x `NOT` y"	$D_x - D_y$
"x `~5` y"	???

Posting extension

Computation of relevance

$$tfidf = \frac{tf}{df}$$

adherent	BTID-60171 BTID-60091 BTID-60170 BTID-60276 BTID-20330|*2* BTID-10620 BTID-60619|*2* BTID-60461
extraction	BTID-60338 BTID-20312 BTID-60637 BTID-50051|*2* BTID-60561 BTID-60018 BTID-50013 BTID-60410
container	BTID-60366 BTID-60593
medicinal	BTID-60037 BTID-60356 BTID-60593
filter	BTID-60576|*2* BTID-60143|*2* BTID-60587|*2* BTID-60262 BTID-60379 BTID-10227
trophozoites	BTID-60333 BTID-60536|*2* BTID-60211|*14* BTID-60066|*6* BTID-60262|*4* BTID-60263|*11*

And more...

"near" queries	Token position of each occurrence.
Highlights	Character offset of each occurrence

Performance optimization

Prefix tree (trie) of terms

- $O(L_{term})$
- Prefix queries ("Bacill*").
- Construction can be distributed.

Software optimization

- Persistance of data structures: minimization of size and traversal.
- Special focus on concurrent access.

Solutions

. . . of AlvisIR2

Technological option: Lucene

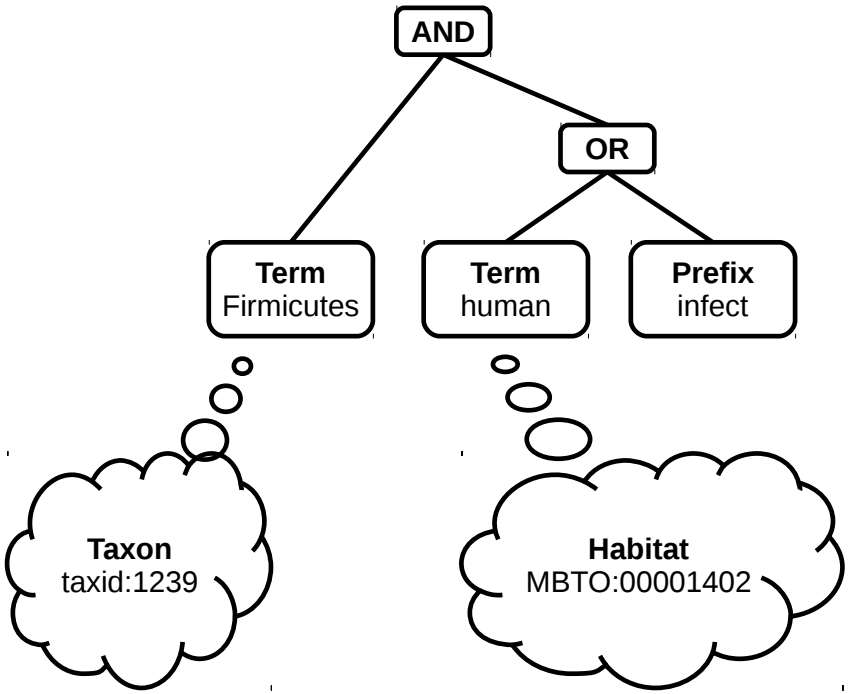
- Apache license: free (speech) and permissive.
- Active developper community.
- Users: industry, academic (especially SemWeb and NLP).

Technical advantages

- API: easy intégration.
- Document vector customization: AlvisNLP/ML provides it.
- Arbitrary extension of postings.

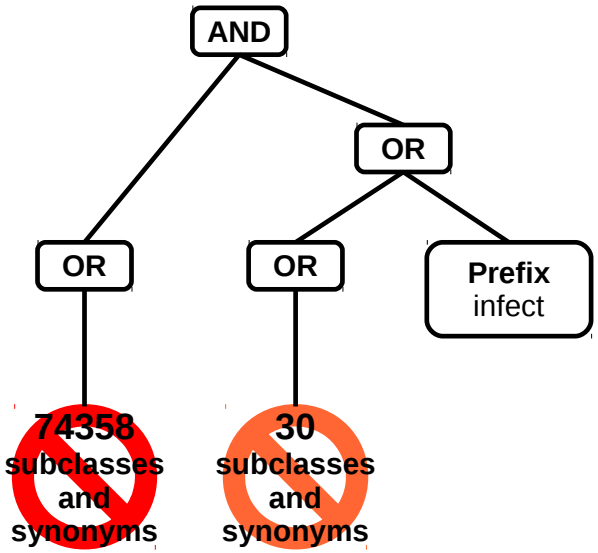
Query expansion

Firmicutes (human OR infect*)



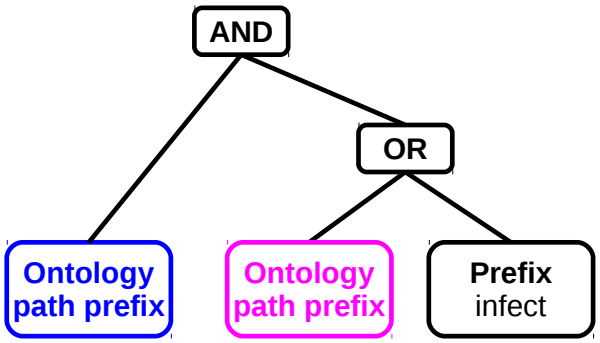
Query expansion (full expansion)

Firmicutes (human OR infect*)



Query expansion (explained canonical expansion)

Firmicutes (human OR infect*)



Path : /1/131567/2/1239
Canonical : Firmicutes
Synonyms : low G+C gram-positive bacteria, Firmacutes, ...
Sub-concepts : ...

Path : /MBTO:00000872/.../MBTO:00001514/MBTO:00001402
Canonical : human
Synonyms : person, people, ...
Sub-concepts : children, ...

Query expansion

The function that expands queries has two outputs:

1. The expanded query.
2. An *explanation* for each expanded term.

Responsibilities of the explanation

- Create the actual query (transparent).
- Provide expansion details.
- Highlight snippets.

Indexing conventions

Campylobacter jejuni contaminates the water

Indexing conventions (tokens)

	Campylobacter	jejuni	contaminates	the	water

Terms	campylobacter	jejuni	contaminate	the	water
ID	001	002	003		004
Position	1	2	3		5
Offset	0-13	14-20	21-33		38-43

Indexing conventions (semantic units)

	Campylobacter	jejuni	contaminates	the	water

Terms	campylobacter	jejuni	contaminate	the	water
ID	001	002	003		004
Position	1	2	3		5
Offset	0-13	14-20	21-33		38-43

Terms	{bacteria}/2/.../197/				{habitat}/.../MBTO:00000707/
ID	005				006
Position	1				5
Offset	0-20				38-43

Indexing conventions (relations)

	Campylobacter	jejuni	contaminates	the	water

Terms	campylobacter	jejuni	contaminate	the	water
ID	001	002	003		004
Position	1	2	3		5
Offset	0-13	14-20	21-33		38-43

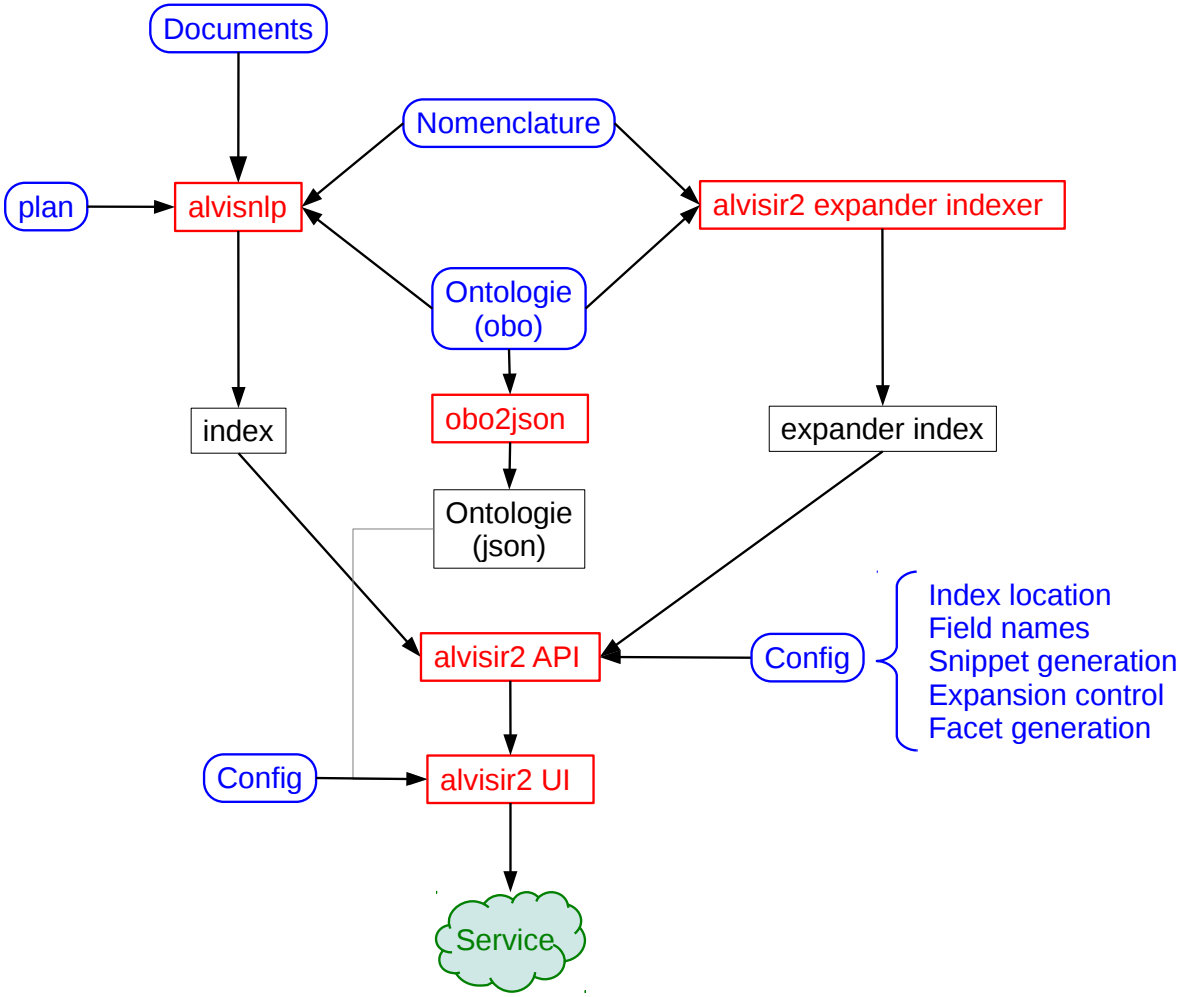
Terms	{bacteria}/2/.../197/	{habitat}/.../MBTO:00000707/
ID	005	006
Position	1	5
Offset	0-20	38-43

Terms	{loc}{bacteria}/2/.../197/~{habitat}/.../MBTO:00000707/
ID	007
Position	1
Offset	0-43
Args	bacteria:005,habitat:006

HOWTO

DIY

Architecture



Prérequis

Logiciels

- AlvisNLP
- bibliome-utils
- AlvisIR2-core
- glassfish/AlvisIR2.war

Ressources

- Plan d'annotation.
- REN basée sur des ressources OBO ou CSV.
- Corpus.
- Une petite idée des requêtes

Pas à pas

1. Ajouter `AlvisIRIndexer` à la fin du plan.
2. Annoter (et donc indexer) avec AlvisNLP.
3. Générer l'index pour l'expansion.
4. Générer `ontologie.json`.
5. Configurer le moteur de recherches.

Si ça change

- Plan d'annotation → 2.
- Ressources → 2, 3, 4.
- Types d'EN → 1, 2, 3, 4, 5.

AlvisIRIndexer module

```
<index class="AlvisIRIndexer">
    <indexDir>path/to/index</indexDir>
    <tokenPositionGap>128</tokenPositionGap>
    <fieldNames>
        title,abstract,author,pmid,year,journal,mesh
    </fieldNames>
    <relations>
        <loc>taxon,habitat</loc>
    </relations>
    <propertyKeys/>
    <documents>
        <fields>
            <instances>
                sections:title | sections:abstract
            </instances>
            <annotations>
                <instances>layer:bacteria</instances>
                <text>"{taxon}" ^ @path ^ "/"</text>
            </annotations>
            <annotations>
                <instances>layer:habitats</instances>
                <text>"{habitat}" ^ @concept-path ^ "/"</text>
            </annotations>
```

AlvisIRIndexer module: indexDir

```
<index class="AlvisIRIndexer">
    <indexDir>path/to/index</indexDir>
    <tokenPositionGap>128</tokenPositionGap>
    <fieldNames>
        title,abstract,author,pmid,year,journal,mesh
    </fieldNames>
    <relations>
        <loc>taxon,habitat</loc>
    </relations>
    <propertyKeys/>
    <documents>
        <fields>
            <instances>
                sections:title | sections:abstract
            </instances>
            <annotations>
                <instances>layer:bacteria</instances>
                <text>"{taxon}" ^ @path ^ "/"</text>
            </annotations>
            <annotations>
                <instances>layer:habitats</instances>
                <text>"{habitat}" ^ @concept-path ^ "/"</text>
            </annotations>
```

Supprime le répertoire avant de réindexer !

AlvisIRIndexer module: tokenPositionGap

```
<index class="AlvisIRIndexer">
    <indexDir>path/to/index</indexDir>
    <tokenPositionGap>128</tokenPositionGap>
    <fieldNames>
        title,abstract,auth  pmid,year,journal,mesh
    </fieldNames>
    <relations>
        <loc>taxon,habita
    </relations>
    <propertyKeys/>
    <documents>
        <fields>
            <instances>
                sections:title | sections:abstract
            </instances>
            <annotations>
                <instances>layer:bacteria</instances>
                <text>"{taxon}" ^ @path ^ "/"</text>
            </annotations>
            <annotations>
                <instances>layer:habitats</instances>
                <text>"{habitat}" ^ @concept-path ^ "/"</text>
            </annotations>
```

C'est un paramètre magique.
Doubler si AlvisNLP refuse d'indexer.

AlvisIRIndexer module: fieldNames

```
<index class="AlvisIRIndexer">
    <indexDir>path/to/index</indexDir>
    <tokenPositionGap>128</tokenPositionGap>
    <fieldNames>
        title,abstract,author,pmid,year,journal,mesh
    </fieldNames>
    <relations>
        <loc>taxon,h    tat</loc>
    </relations>
    <propert
```

Tous les champs à requêter ou à afficher doivent être déclarés

```
                                    sections:abstract

            <annotations>
                <instances>layer:bacteria</instances>
                <text>"{taxon}" ^ @path ^ "/"</text>
            </annotations>
            <annotations>
                <instances>layer:habitats</instances>
                <text>"{habitat}" ^ @concept-path ^ "/"</text>
            </annotations>
```

AlvisIRIndexer module: relations

```
<index class="AlvisIRIndexer">
    <indexDir>path/to/index</indexDir>
    <tokenPositionGap>128</tokenPositionGap>
    <fieldNames>
        title,abstract,author,pmid,year,journal,mesh
    </fieldNames>
    <relations>
        <loc>taxon,habitat</loc>
    </relations>
    <propertyKeys/>
    <documents>
        <fields>
            <instances>
                <text>"{taxon}" ^ @path ^ "/"</text>
            </annotations>
            <annotations>
                <instances>layer:habitats</instances>
                <text>"{habitat}" ^ @concept-path ^ "/"</text>
            </annotations>
```

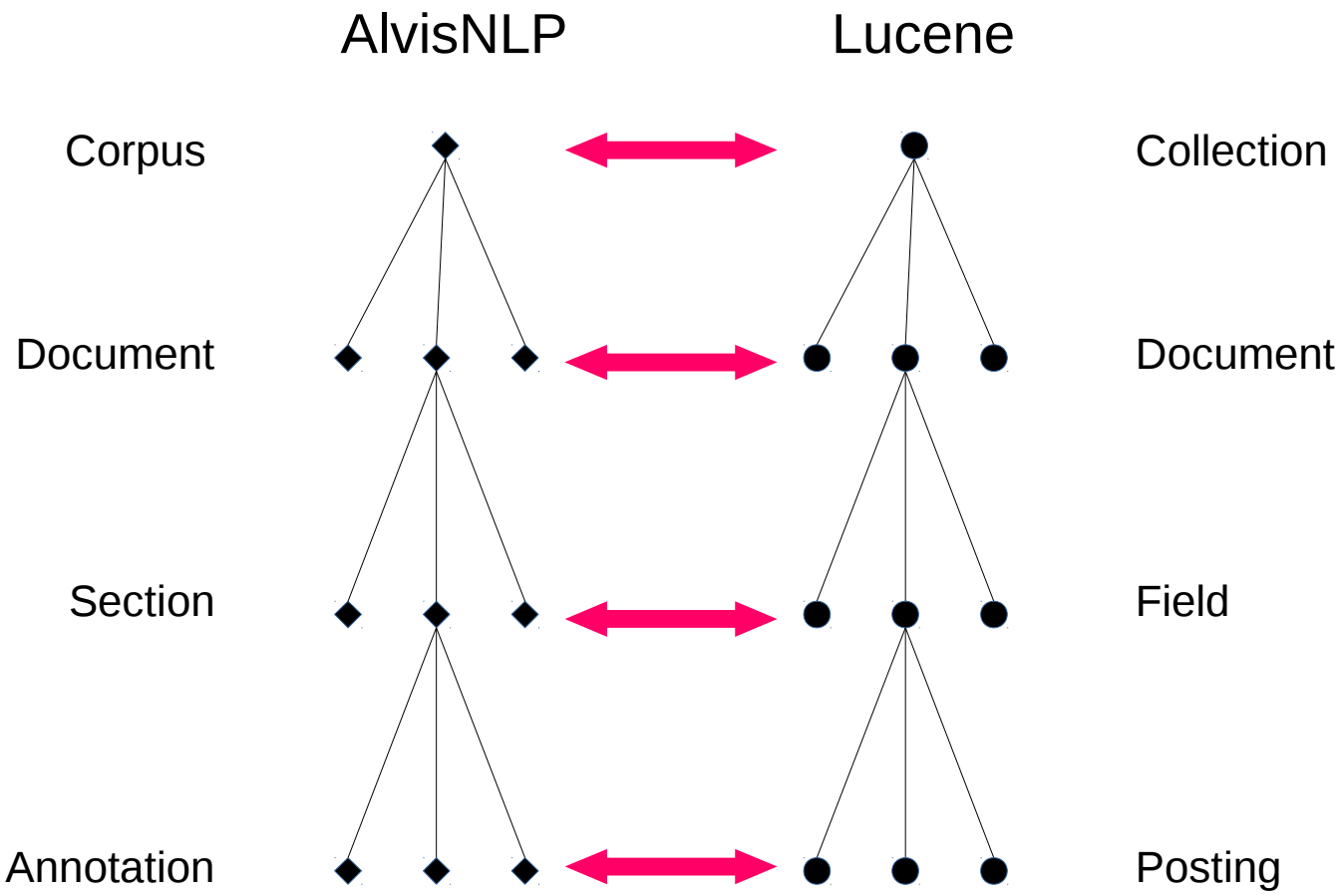
Déclaration des noms de relations et de leurs arguments.
Pas de schéma pour AlvisIR.

AlvisIRIndexer module: propertyKeys

```
<index class="AlvisIRIndexer">
    <indexDir>path/to/index</indexDir>
    <tokenPositionGap>128</tokenPositionGap>
    <fieldNames>
        title,abstract,author,pmid,year,journal,mesh
    </fieldNames>
    <relations>
        <loc>taxon,habitat</loc>
    </relations>
    <propertyKeys/>
    <documents>
        <fields>
            <instances>
                sections:title | sections:abstract
            </instances>
            <annotations>
                <instances>layer:bacteria</instances>
                <text>"{taxon}" ^ @path ^ "/"</text>
            </annotations>
            <annotations>
                <instances>layer:habitats</instances>
                <text>"{habitat}" ^ @concept-path ^ "/"</text>
            </annotations>
```

Obligatoire.
dsl

AlvisNLP vs Lucene data models



AlvisIRIndexer module: documents

```
<index class="AlvisIRIndexer">
    <indexDir>path/to/index</indexDir>
    <tokenPositionGap>128</tokenPositionGap>
    <fieldNames>
        title,abstract,author,pmid,year,journal,mesh
    </fieldNames>
    <relations>
        <loc>taxon,habitat</loc>
    </relations>
    <propertyKeys/>
    <documents>
        <fields>
            <instances>
                sections:title | sections:abstract
            </instances>
            <annotations>
                <instances>layer:bacteria</instances>
                <text>"{taxon}" ^ @path ^ "/"</text>
            </annotations>
            <annotations>
                <instances>layer:habitats</instances>
                <text>"{habitat}" ^ @concept-path ^ "/"</text>
            </annotations>
```

AlvisIRIndexer module: fields

```
<index class="AlvisIRIndexer">
    <indexDir>path/to/index</indexDir>
    <tokenPos
    <fie
    </fi
    <relations
        <loc>taxon,habitat
    </relations>
    <propertyKeys/>
    <documents>
        <fields>
            <instances>
                sections:title | sections:abstract
            </instances>
            <annotations>
                <instances>layer:bacteria</instances>
                <text>"{taxon}" ^ @path ^ "/"</text>
            </annotations>
            <annotations>
                <instances>layer:habitats</instances>
                <text>"{habitat}" ^ @concept-path ^ "/"</text>
            </annotations>
```

Default : all sections.
Annotations différentes selon les champs.

AlvisIRIndexer module: annotations

```
<index class="AlvisIRIndexer">
    <indexDir>path/to/index</indexDir>
    <tokenPos
    <fie
    </fi
    <relations
        <loc>taxon,habitat
    </relations>
    <propertyKeys/>
    <documents>
        <fields>
            <instances>
                sections:title    sections:abstract
            </instances>
            <annotations>
                <instances>layer:bacteria</instances>
                <text>"{taxon}" ^ @path ^ "/"</text>
            </annotations>
            <annotations>
                <instances>layer:habitats</instances>
                <text>"{habitat}" ^ @concept-path ^ "/"</text>
            </annotations>
```

Indiquer :

- les postings,
- le texte indexé.