Exploitation de données avec Allegro

Exemple des *Actes de Langages Stéréotypés*

Recueil des besoins

Recueil des besoins

Besoins récurrents pour l’exploitations des données textuelles semi-structurées produites au laboratoire :

- Mise à disposition en accès libre / accès contrôlé
- Recherches croisées sur l'intégralité des données, de leurs structuration et de leurs métadonnées
- Recherche plein texte / concordances en contexte

Recueil des besoins

Les besoins plus précis s'affinent souvent au fur et à mesure des développements, en fonction du projet.

Plusieurs itérations sont souvent nécessaires pour obtenir un résultat satisfaisant.

Un cycle de développement complet comprend en général

- Un entretien avec les porteurs du projet
- Une potentielle révision de la structure des données
- Un affinement de la mise en forme et du rendu

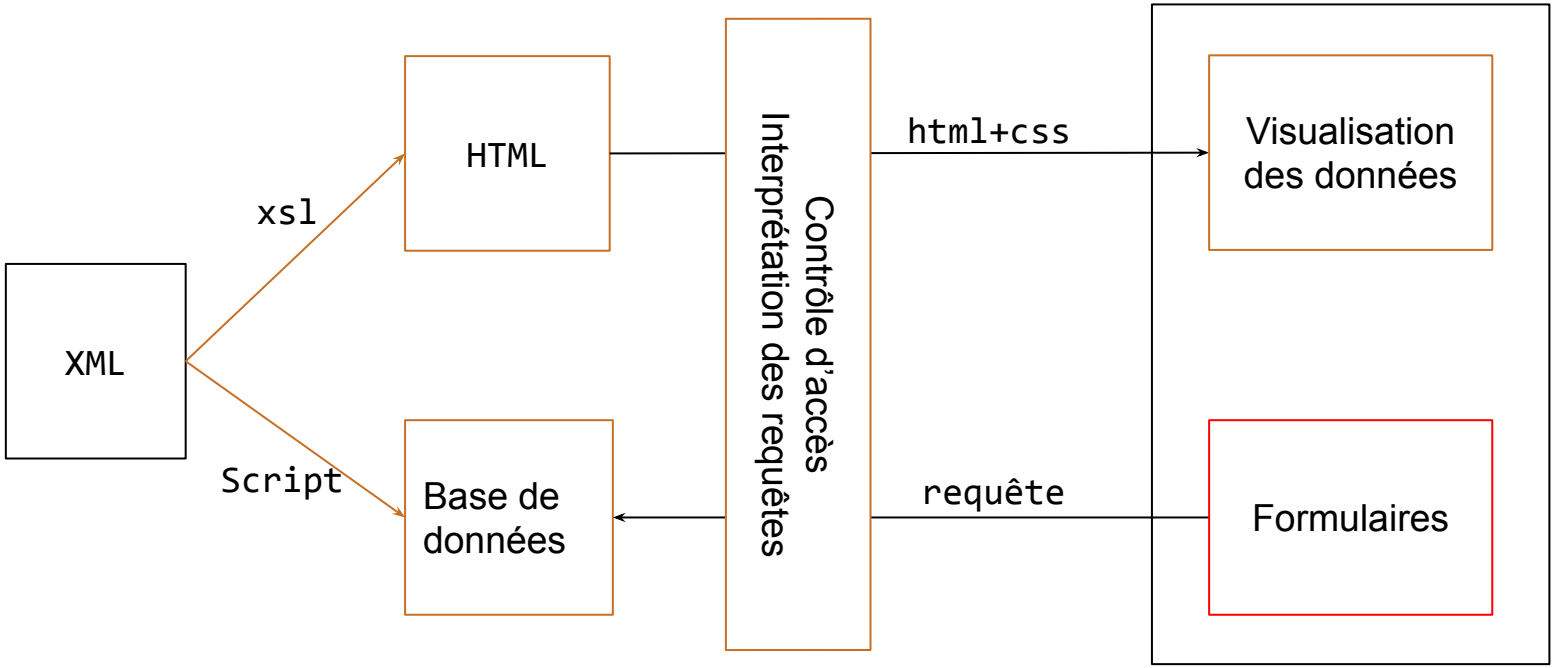
Implémentation classique

Implémentation classique

La solution la plus classique à ce genre de problème consiste généralement à :

- Indexer les données sous la forme d’une base de donnée
- Transformer les données vers un format compatible avec le web (généralement HTML)
- Concevoir les feuilles de styles correspondantes
- Implémenter le contrôle d’accès, l'interprétation des requêtes, etc…

Potentiellement répéter ces étapes à chaque itération



Implémentation classique

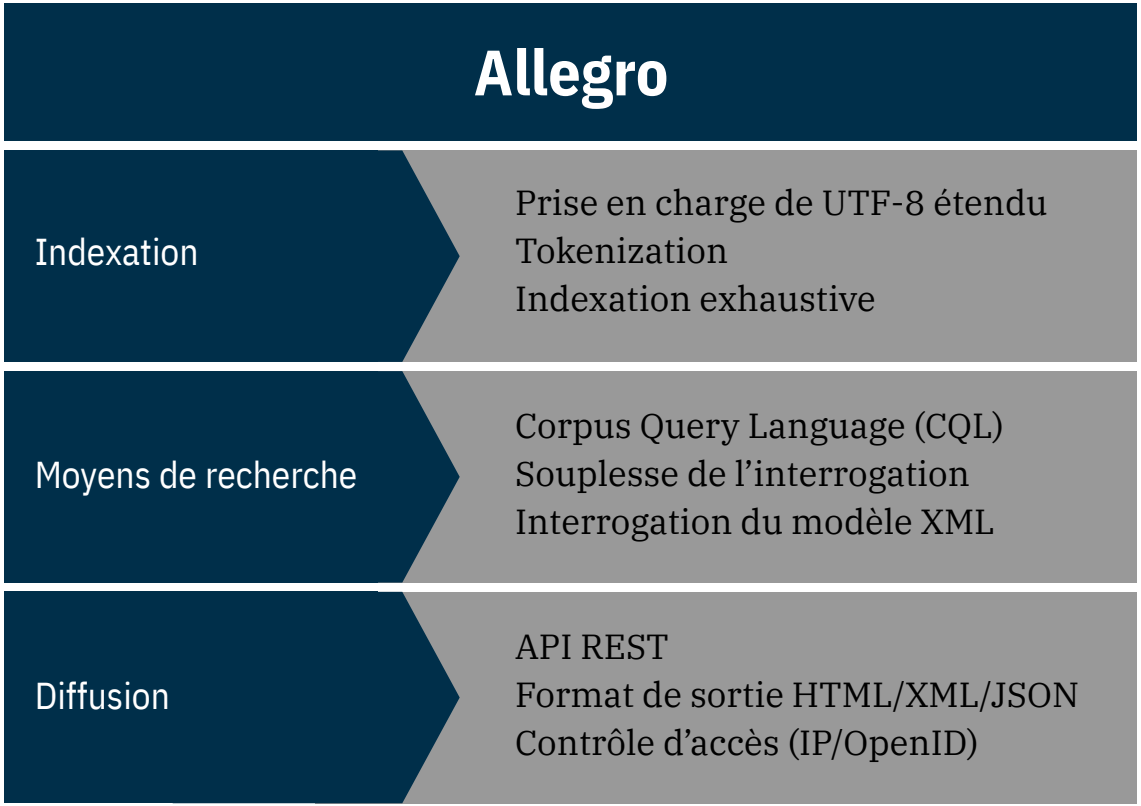
Implémentation avec Allegro

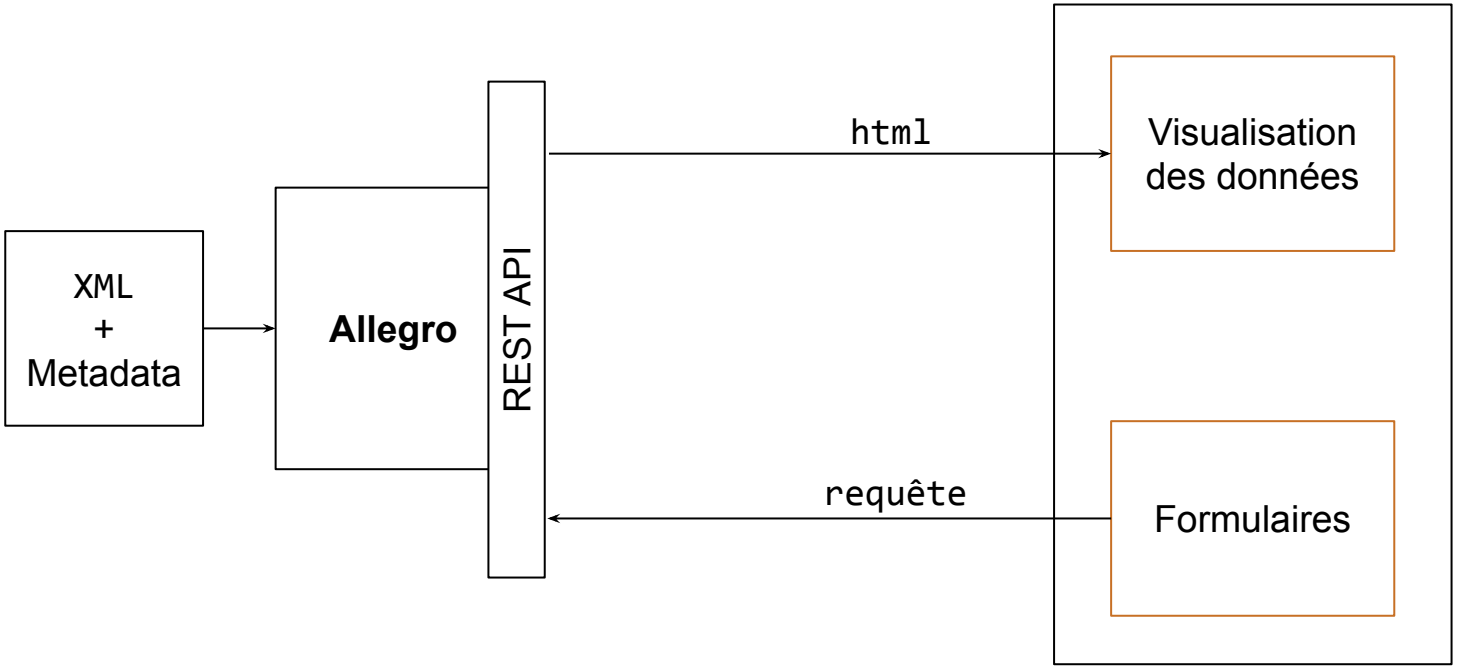
Implémentation avec Allegro

Allegro permet d’indexer, d’interroger et de diffuser des corpus de données semi-structurées, comme le XML.

L’architecture présentée ici est donc centrée sur l’exploitation de données de ce type.

Elle a vocation à simplifier la chaîne de traitement et de réduire les coûts de développement pour ce type de projets.


Allegro
Indexation
Prise en charge de UTF-8 étendu
Tokenization
Indexation exhaustive
Moyens de recherche
Corpus Query Language (CQL)
Souplesse de l'interrogation
Interrogation du modèle XML
Diffusion
API REST
Format de sortie HTML/XML/JSON
Contrôle d'accès (IP/OpenID)



Implémentation avec Allegro

Structuration des données

Exemple des *Actes de Langages Stéréotypés*

Structuration des données

Dans l’exemple des ALS les données proviennent d’un ensemble de fichier *word*, chaque fichier contenant un article.

Pour permettre de les exploiter, les fichiers sont modélisés en XML TEI (merci Jessika).

Petit dictionnaire permanent
des « actes de langages stéréotypés » (ALS)

Microstructure de *C'est vite dit*

Caroline Pernot
avec la collaboration des membres du GLFA
et des étudiants de l'Université de Lorraine – Metz
Coordination : Maurice Kauffer

PRÉSENTATION GÉNÉRALE

FORME ET SYNTAXE

Variantes : *vite dit !* (familier, langue parlée) ; *c'est plus vite dit que fait.*

Figement : Incomplet. Substitutions possibles au pronom : N *est vite dit.* Reprise ou explicitation possibles du pronom. Le verbe est toujours au présent, l'imparfait signalant un discours indirect libre (*c'était vite dit,* rare) ou un emploi descriptif.

Configurations syntaxiques : *c'est vite dit !* ; *c'est vite dit de* + GInf ; *c'est vite dit,* GInf ; *c'est vite dit que* + subord. ; *[ça] c'est vite dit [ça]* + reprise anaphorique d'un élément contextuel.

SENS / FONCTIONS

Type d'acte : DÉSACCORD

Fonctions :

- *C'est vite dit* sert à exprimer un désaccord, fondé sur un manque de réflexion que le locuteur présume chez son interlocuteur.
- Le désaccord porte sur la pertinence d'un énoncé, sur sa véracité ou sur un contenu évalué par le locuteur comme étant difficilement réalisable.

Concurrents : *[ça] c'est toi qui le dis ! ; ça n'est pas si vrai que ça ; ce n'est pas si simple [que ça] ; ce n'est rien de le dire ; c'est facile à dire [mais moins facile à faire] ; c'est / ce n'est pas dit ! ; c'est un peu fort ; comme tu y vas ; le dire, c'est une chose [le faire en est une autre] ; il faut le dire vite ; on dit ça ! ; quand même ; que tu crois / dis ! ; sais-tu seulement de quoi tu parles ? ; tout le monde peut le dire.*

USAGE

Registre : Standard.

Partenaires privilégiés : Aucun.

ÉQUIVALENTS

- Habituels *: das ist leicht gesagt ; das ist schnell gesagt*
- Habituels *: das ist leicht gesagt ; das ist schnell gesagt*
- Occasionnels : *das ist leicht / schnell gesagt / dahingesagt / hingesagt ; das ist schnell behauptet ; das sagt sich [so] leicht / [so] leicht dahin / [so] leicht hin / [so] schnell ; es ist simpel zu sagen ; leicht gesagt ;* [GV ou GV] *ist [vielleicht] ein bisschen übertrieben ; wie rasch das gesagt ist!*

PLAN :

DESCRIPTION DETAILLEE DES FONCTIONS ET EMPLOIS

I. *DESACCORD FONDÉ SUR LA PERTINENCE*

1. Le désaccord porte sur la justesse d'une description

– J'ai vu Nicole, ce matin. […] Elle avait affaire à une espèce de désaxé. […]
– Oui... enfin, un désaxé, c'est vite dit. On ne l'a pas retrouvée à l'hôpital, que je sache. (PDB 346-322)

„Ich habe Nicole heute morgen gesehen. [Sie hatte] es mit einem Psychopathen zu tun."
„Ja... aber weißt du, ein Psychopath, das ist schnell gesagt. Soweit ich weiß, hat man ihn nicht in einem Krankenhaus aufgefunden."

Il n'a jamais jugé utile de savoir à quel prix et sur quels renoncements s'était finalement stabilisé notre couple. En tout cas, il n'a jamais voulu en parler. Il n'a pesé que son renoncement à lui, très réel. Mais renoncement... sacrifice... c'est vite dit. Il est malhonnête de condamner rétrospectivement son passé sous prétexte qu'il n'a pas été éternel ou

Er hat es nie für sinnvoll gehalten, in Erfahrung zu bringen, unter welchen Opfern und mit wieviel Verzicht sich am Ende unsere Partnerschaft stabilisiert hat. Auf jeden Fall wollte er nie darüber sprechen. Er hat nur seinen eigenen, ganz realen Verzicht abgewogen. Aber Verzicht, Opfer, das ist leicht gesagt. Es ist unehrlich, rückblickend seine Vergangenheit zu

Article word fourni par l'équipe des rédacteurs

Petit dictionnaire permanent
des « actes de langages stéréotypés » (ALS)

Microstructure de *C'est vite dit*

Caroline Pernot
avec la collaboration des membres du GLFA
et des étudiants de l'Université de Lorraine – Metz
Coordination : Maurice Kauffer

PRÉSENTATION GÉNÉRALE

FORME ET SYNTAXE
Variantes : *vite dit !* (familier, langue parlée) ; *c'est plus vite dit que fait.*
Figement : Incomplet. Substitutions possibles au pronom : N *est vite dit.* Reprise ou explicitation possibles du pronom. Le verbe est toujours au présent, l'imparfait signalant un discours indirect libre (*c'était vite dit,* rare) ou un emploi descriptif.
Configurations syntaxiques : *c'est vite dit !* ; *c'est vite dit de* + GInf ; *c'est vite dit,* GInf ; *c'est vite dit que* + subord. ; *[ça] c'est vite dit [ça]* + reprise anaphorique d'un élément contextuel.

SENS / FONCTIONS
Type d'acte : DÉSACCORD
Fonctions :
- *C'est vite dit* sert à exprimer un désaccord, fondé sur un manque de réflexion que le locuteur présume chez son interlocuteur.
- Le désaccord porte sur la pertinence d'un énoncé, sur sa véracité ou sur un contenu évalué par le locuteur comme étant difficilement réalisable.

Concurrents : *[ça] c'est toi qui le dis ! ; ça n'est pas si vrai que ça ; ce n'est pas si simple [que ça] ; ce n'est rien de le dire ; c'est facile à dire [mais moins facile à faire] ; c'est / ce n'est pas dit ! ; c'est un peu fort ; comme tu y vas ; le dire, c'est une chose [le faire en est une autre] ; il faut le dire vite ; on dit ça ! ; quand même ; que tu crois / dis ! ; sais-tu seulement de quoi tu parles ? ; tout le monde peut le dire.*

USAGE
Registre : Standard.
Partenaires privilégiés : Aucun.

ÉQUIVALENTS
- Habituels *: das ist leicht gesagt ; das ist schnell gesagt*

```
<form type="variante">
    <orth>vite dit !</orth>
    <usg type="registre">familier</usg>
    <usg type="registre">langue parlée</usg>
</form>
<form type="variante">
    <orth>c'est plus vite dit que fait</orth>
</form>
<note type="figementMorphoSyntaxique">Incomplet. Substitutions possibles au
    pronom : N <hi rend="italique">est vite dit</hi>. Reprise ou explicitation
    possibles du pronom. Le verbe est toujours au présent, l'imparfait signalant
    un discours indirect libre (<hi rend="italique">c'était vite dit</hi>, rare)
    ou un emploi descriptif.</note>
<note type="configurationSyntaxique"><hi rend="italique">c'est vite dit !</hi> ;
        <hi rend="italique">c'est vite dit de</hi> + GInf ; <hi rend="italique"
        >c'est vite dit</hi>, GInf ; <hi rend="italique">c'est vite dit que</hi>
    + subord. ; <hi rend="italique">[ça] c'est vite dit [ça]</hi> + reprise
    anaphorique d'un élément contextuel.</note>
<usg type="acte">DESACCORD</usg>
```

Structuration et modélisation du contenu de l'article

- Habituels : *das ist leicht gesagt ; das ist schnell gesagt*
- Occasionnels : *das ist leicht / schnell gesagt / dahingesagt / hingesagt ; das ist schnell behauptet ; das sagt sich [so] leicht / [so] leicht dahin / [so] leicht hin / [so] schnell ; es ist simpel zu sagen ; leicht gesagt ;* [GV ou GV] *ist [vielleicht] ein bisschen übertrieben ; wie rasch das gesagt ist!*

PLAN :

I. *DESACCORD FONDÉ SUR LA PERTINENCE*
1. Le désaccord porte sur la justesse d'une description
2. Le désaccord porte sur une situation donnée
3. Le désaccord porte sur la validité d'un argument

II. *DESACCORD FONDÉ SUR LA VÉRACITE D'UN ÉNONCÉ*

III. *DESACCORD LIÉ AU CARACTÈRE IRRÉALISTE D'UN PROJET*
1. En réaction à une proposition
2. En réaction à un énoncé prospectif
3. En réaction à une injonction

DESCRIPTION DETAILLEE DES FONCTIONS ET EMPLOIS

I. *DESACCORD FONDÉ SUR LA PERTINENCE*

1. Le désaccord porte sur la justesse d'une description

– J'ai vu Nicole, ce matin. […] Elle avait affaire à une espèce de désaxé. […]
– Oui... enfin, un désaxé, c'est vite dit. On ne l'a pas retrouvée à l'hôpital, que je sache. (PDB 346-322)

„Ich habe Nicole heute morgen gesehen. [Sie hatte] es mit einem Psychopathen zu tun."
„Ja... aber weißt du, ein Psychopath, das ist schnell gesagt. Soweit ich weiß, hat man ihn nicht in einem Krankenhaus aufgefunden."

Il n'a jamais jugé utile de savoir à quel prix et sur quels renoncements s'était finalement stabilisé notre couple. En tout cas, il n'a jamais voulu en parler. Il n'a pesé que son renoncement à lui, très réel. Mais renoncement... sacrifice... c'est vite dit. Il est malhonnête de condamner rétrospectivement son passé sous prétexte qu'il n'a pas été éternel ou

Er hat es nie für sinnvoll gehalten, in Erfahrung zu bringen, unter welchen Opfern und mit wieviel Verzicht sich am Ende unsere Partnerschaft stabilisiert hat. Auf jeden Fall wollte er nie darüber sprechen. Er hat nur seinen eigenen, ganz realen Verzicht abgewogen. Aber Verzicht, Opfer, das ist leicht gesagt. Es ist unehrlich, rückblickend seine Vergangenheit zu


```
<div type="fonctionEmploi">
    <div n="I">
        <head>DESACCORD FONDE SUR LA PERTINENCE</head>
        <div n="1">
            <head>Le désaccord porte sur la justesse d'une description</head>
            <cit type="exemple">
                <cit type="original" xml:lang="fr">
                    <quote>– J'ai vu Nicole, ce matin. […] Elle avait affaire à une
                        espèce de désaxé. […]<lb/> – Oui... enfin, un désaxé, c'est vite
                        dit. On ne l'a pas retrouvée à l'hôpital, que je sache. </quote>
                    <ref target="#PDB">(PDB 346-322)</ref>
                </cit>
                <cit type="traduction" xml:lang="de">
                    <quote>„Ich habe Nicole heute morgen gesehen. [Sie hatte] es mit
                        einem Psychopathen zu tun."<lb/> „Ja... aber weißt du, ein
                        Psychopath, das ist schnell gesagt. Soweit ich weiß, hat man ihn
                        nicht in einem Krankenhaus aufgefunden." </quote>
                </cit>
            </cit>
```

Structuration et modélisation du contenu de l'article

Gestion des métadonnées

Pour compléter cette modélisation, on peut produire un fichier de métadonnées qui sera exploité par Allegro.

Cela nous permet de caractériser les fichiers de notre corpus en fonction de critères spécifiques.

C’est aussi une façon d’agréger des informations qui ne sont pas forcément présents dans les fichiers XML.

Exploitation

Exploitation

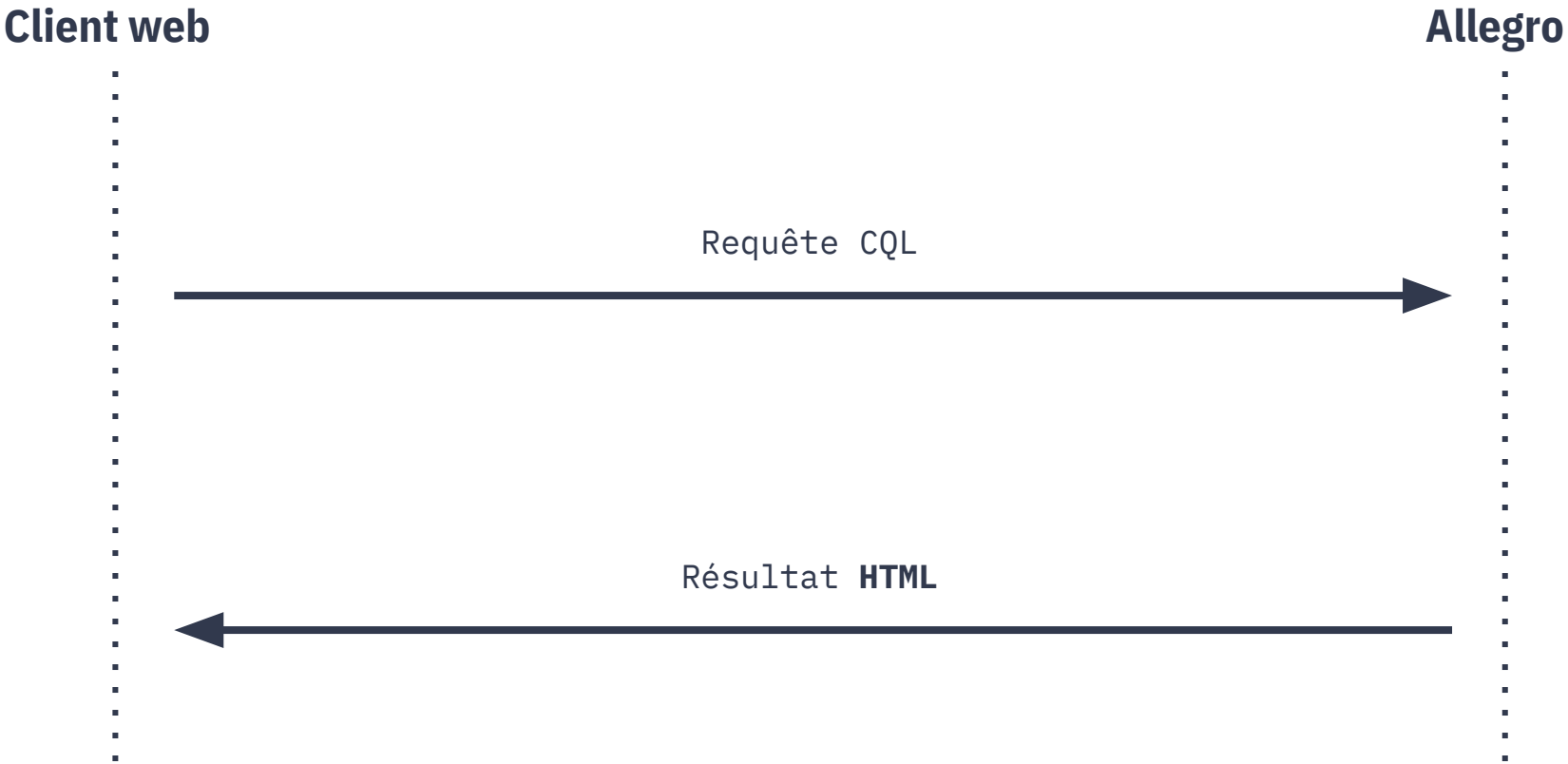
Interrogation

Une fois les données transcrites en XML elles sont directement exploitables par Allegro.

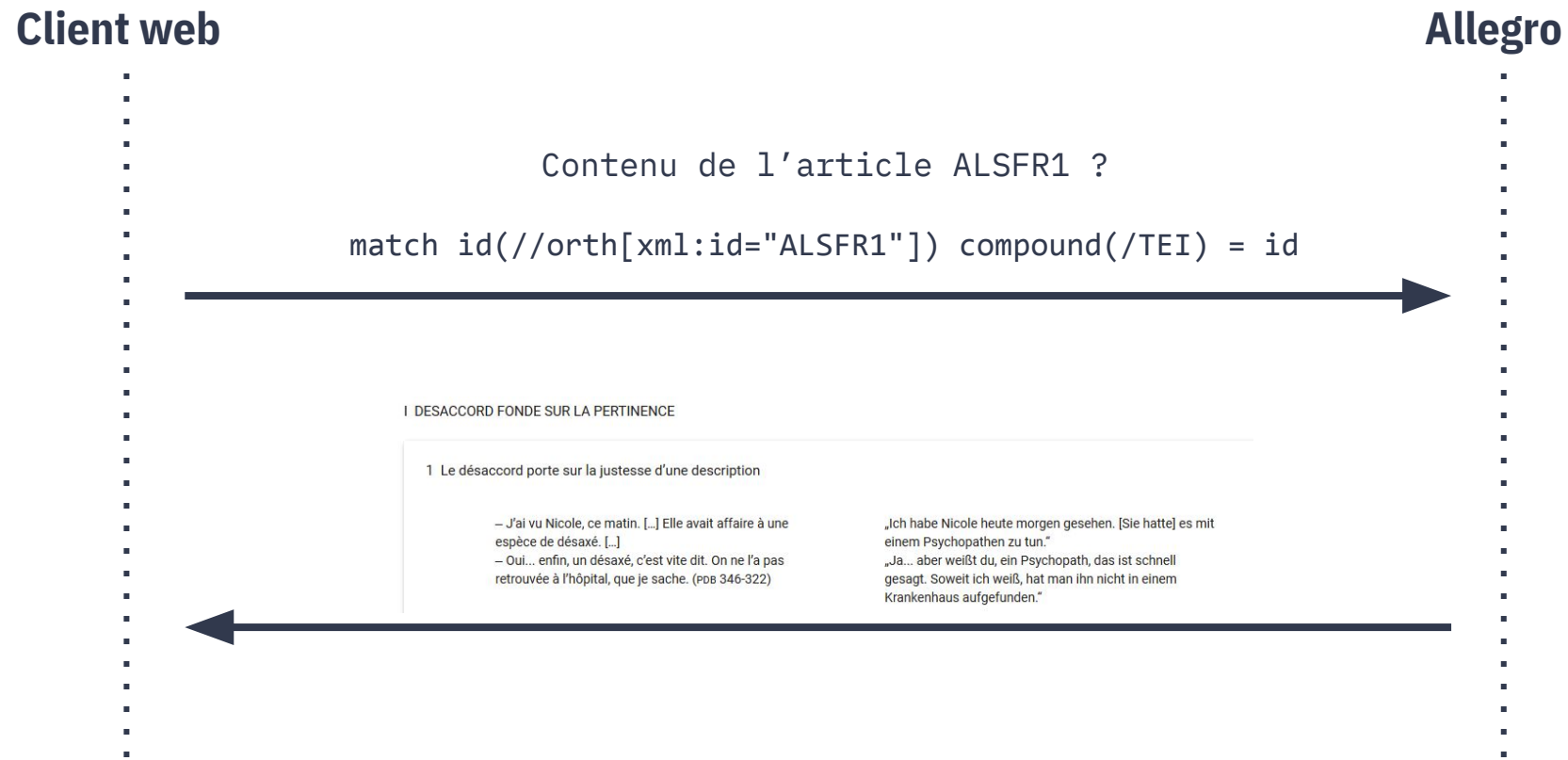
Il n’y a pas d’autres représentation intermédiaire de la donnée. On interroge directement le XML.

Il n’y a pas d’autres briques logicielles à configurer ou à maintenir.

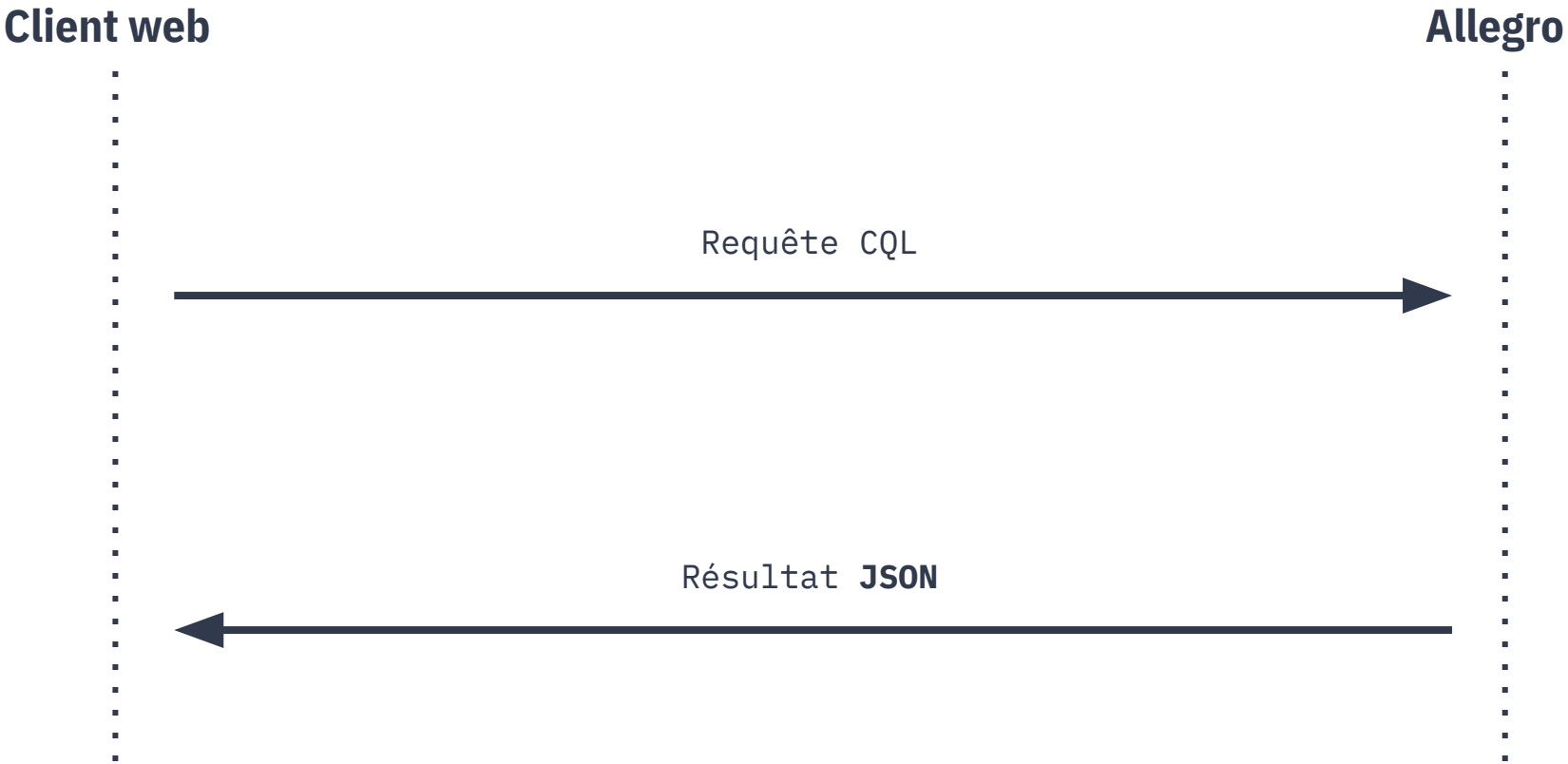
Dans notre exemple, on utilise un framework javascript pour écrire l’interface.



Interrogation de la donnée



Implémentation des formulaires de recherche



Interrogation de la donnée


Client web
Allegro
Liste des valeurs de registre existants ?
unique(//registre)
Registre
parlé
langue parlée
standard
familier

Exploitation

Mise en forme

Le contenu HTML retourné par Allegro est issu d’une transformation bi-univoque du XML

Nous pouvons donc écrire les règles de mise en forme directement à partir du modèle XML

Cela facilite la mise en forme typographiques de n’importe quel contenu balisé

Transformation XML/HTML

Extrait du XML indexé

```xml
<head>Le désaccord porte sur la justesse d’une description</head>
<cit type="exemple">
    <cit type="original" xml:lang="fr">
        <quote>- J’ai vu Nicole, ce matin. […] Elle avait affaire à une
            espèce de désaxé. […]<lb/> - Oui... enfin, un désaxé, c’est vite
            dit. On ne l’a pas retrouvée à l’hôpital, que je sache. </quote>
        <ref target="#PDB">(PDB 346-322)</ref>
    </cit>
    <cit type="traduction" xml:lang="de">
        <quote>„Ich habe Nicole heute morgen gesehen. [Sie hatte] es mit
            einem Psychopathen zu tun.“<lb/> „Ja... aber weißt du, ein
            Psychopath, das ist schnell gesagt. Soweit ich weiß, hat man ihn
            nicht in einem Krankenhaus aufgefunden.“ </quote>
    </cit>
</cit>
```

Version HTML retournée

```html
<span class="n-head">Le désaccord porte sur la justesse d’une description</span>
<span class="n-cit" data-type="exemple">
    <span class="n-cit" data-xml-ns-lang="fr" data-type="original">
        <span class="n-quote">- J’ai vu Nicole, ce matin. […] Elle avait affaire
            à une espèce de désaxé. […] <span class="n-lb"></span> - Oui... enfin, un désaxé, c’est vite
            dit. On ne l’a pas retrouvée à l’hôpital, que je sache. </span>
        <span class="n-ref" data-target="#PDB">(PDB 346-322)</span>
    </span>
    <span class="n-cit" data-xml-ns-lang="de" data-type="traduction">
        <span class="n-quote">„Ich habe Nicole heute morgen gesehen. [Sie hatte] es mit
            einem Psychopathen zu tun.“ <span class="n-lb"></span> „Ja... aber weißt du, ein
            Psychopath, das ist schnell gesagt. Soweit ich weiß, hat man ihn
            nicht in einem Krankenhaus aufgefunden.“
        </span>
    </span>
</span>
```

Quelques exemples

Base d’énoncés sapientiaux Aliento

https://base.aliento.eu

Dictionnaire de l’Académie

https://academie.atilf.fr

Dictionnaire des Actes de Langages Stéréotypés

https://als.atilf.fr

FEW rétroconvertit

https://few-webapp.atilf.fr

Frantext

https://www.frantext.fr