



# Patois et digital, pour le passé et l'avenir



## Glossaire des patois de la Suisse romande

### Recherche textuelle dans le *Glossaire des patois de la Suisse romande*

Projet en Nouvelles technologies de l'information et de la communication (MA6)

Fabrice Camus

Mai 2021

Université de Genève - Faculté des lettres

## Avant-Propos

Dans le cadre du module NTIC, le projet réalisé et documenté ici s'est focalisé sur l'intégration des articles du Glossaire des patois de la Suisse romande dans son application web ainsi que le développement de la recherche plein-texte dans les articles. Ce projet s'inscrit dans une collaboration de plus grande ampleur, collaboration qui existe depuis plus de 20 ans. Le développement de leur application web est un projet qui a démarré en 2015.

## Partenaires du projet



**Haute école de gestion Arc, Neuchâtel  
HES-SO**

**Glossaire des patois de la Suisse  
romande  
Université de Neuchâtel**

# Table des matières

<b>Avant-Propos .....</b>	<b>2</b>
<b>Partenaires du projet.....</b>	<b>2</b>
<b>1 Introduction .....</b>	<b>4</b>
<b>2 Contexte du projet.....</b>	<b>4</b>
2.1 Présentation du Glossaire des patois de la Suisse romande.....	4
2.2 Articles.....	4
2.3 Particularités typographiques .....	5
2.4 Rédaction des articles.....	7
2.5 Partenariat avec la Heg-Arc .....	7
<b>3 Intégration des 40'000 articles dans l'application Web .....</b>	<b>8</b>
3.1 Rétrodigitalisation du Glossaire .....	8
3.2 Objectifs .....	8
3.3 Contraintes .....	9
3.4 Intégration dans le SI existant .....	9
<b>4 Recherche textuelle – Analyse des besoins .....</b>	<b>9</b>
4.1 Use case système .....	9
4.2 Maquettes .....	10
4.2.1 Recherche par mots-clés .....	10
4.2.2 Affichage d'un article .....	11
4.3 Synthèse des besoins .....	11
<b>5 Conception de la solution à développer .....</b>	<b>12</b>
5.1 Persistance des articles .....	12
5.2 Indexation, recherche textuelle et diacritiques non-standards.....	13
<b>6 Génération des articles au format xhtml.....</b>	<b>14</b>
6.1 Analyse des sources XML.....	14
6.2 Chaîne de traitement.....	16
6.2.1 Etape 1 : Nettoyage et correction du contenu .....	16
6.2.2 Etape 2 : Transformation xhtml .....	18
6.2.3 Etape 3 : Chargement des articles dans la base de données Oracle.....	19
<b>7 Développement de la recherche textuelle dans les articles .....</b>	<b>20</b>
7.1 Création d'un tri linguistique propre au GPSR .....	20
7.2 Indexation des articles .....	21
<b>8 Développement de l'interface Web .....</b>	<b>22</b>
8.1 Recherche par mots-clés.....	22
8.2 Recherche personnalisée .....	23
8.3 Interface web de développement d'application .....	24
<b>9 Conclusion.....</b>	<b>24</b>

## 1 Introduction

Ce document présente le projet sur lequel j'ai travaillé et qui est attendu dans le cadre du module MA6 – NTIC. Ce projet s'inscrit dans un mandat de plus grande ampleur réalisé dans le cadre de mon activité professionnelle à la Haute école de gestion Arc / Filière Informatique de gestion. Ce document ne se veut pas exhaustif dans les détails techniques, car cela ne serait pas nécessairement pertinent, et ne pourrait être exploité comme documentation de projet. Par contre, j'ai fait le choix de présenter la globalité du processus mis en place, en m'attardant sur certains points technologiques intéressants ou spécifiques. Pour terminer, il dépasse largement la taille demandée, mais cela s'explique par ma volonté de présenter le client et ses particularités en étant relativement précis pour bien comprendre le projet et les besoins.

## 2 Contexte du projet

### 2.1 Présentation du Glossaire des patois de la Suisse romande

"Le Glossaire des patois de la Suisse romande (GSPR) est, depuis 1899, un acteur essentiel dans la mise en valeur du patrimoine linguistique romand. Etabli à Neuchâtel, il est l'un des quatre Vocabulaires nationaux de la Confédération helvétique. Tout comme ses confrères alémanique, grison et tessinois, il a pour mission de documenter le plus complètement possible les patois de son domaine linguistique, d'en faire l'analyse lexicologique et de rendre celle-ci accessible au public et au monde scientifique sous la forme d'un dictionnaire dialectal de grande ampleur."<sup>1</sup>

La constitution du corpus du Glossaire a nécessité un quart de siècle et est formé d'environ 3 millions de fiches manuscrites. Ce fond documentaire est à la base de la rédaction des articles du Glossaire, qui compte plus de 7000 pages rédigées, représentant plus de 30'000 articles. La rédaction du Glossaire a démarré en 1924 avec la lettre A-. Début 2021, la rédaction des articles des lettres I- et J- a commencé... Le GPSR dispose encore d'autres sources manuscrites enrichissant leurs matériaux telles que des dictionnaires et lexiques sur les parlers locaux, le français régional, des centaines de textes patois parus dans les journaux, des documents anciens publiés par des historiens etc...

### 2.2 Articles

Chaque article traite d'un mot, le mot vedette ou en-tête. Tout l'article lui est consacré.

Selon le *Guide et complément*<sup>2</sup> du Glossaire des patois de la Suisse romande, la structure-type d'un article est la suivante :

- Première partie, les variantes phonétiques : « paragraphe consacré à l'exposé des formes que revêt un mot dans les patois, les langues anciennes et le français régional de la Suisse romande. »
- Seconde partie, le corps de l'article : « [...] présentation des sens et emplois attestés pour ce mot, généralement illustrés par des exemples, suivis de leur traduction (pour le patois) et de leur localisation. »
- Troisième partie, le commentaire scientifique : « paragraphe réservé au commentaire scientifique des données lexicographiques exposées dans le paragraphe initial et le corps de l'article. »

---

<sup>1</sup> Glossaire des patois de la Suisse romande [en ligne]. [Consulté le 01.05.2021]. Disponible à l'adresse : <http://www.unine.ch/gpsr>

<sup>2</sup> Guide et complément, F. Python, 2018, disponible au GPSR

Exemple avec l'article *gatòlyao*

1

**gatòlyao**, -aou, f. -**za**, -ja Vd 51, 61, 64, 9 MAT., Vd BR., DUM., -āu 31, -(l)yā F 1, 33, 3 L'HOMME, 45, 63, -āeu V 32, 36, -u N 31, 41, -zao Vd 20, 21, -é 15 a, -ólyók V 71 KE. (f. -*ḡouja*), -òulèu 47 GABBUD (-*oloou*, -ly- BJERROME), -ou 60, -oulyè 50, BA., -ök 75 JJ., -lööü 51 A. PRAZ, -uyóouk 75 mod. (f. -*ḡouja*), -alyè 23 (f. -*ḡeuz*), -yœu 35, -zè 18, -òæ Vd 14 (-òè JA.), -dæu V 43. **Fr.** *gatoilleux* F GR.; — **gatilyō**, f. -**za**, -z Vd 80, -(ə)you J 40, 42, 47, 60, 6 BIE., VA. (et -tè-), *gaklyè* N 22. **Anc.** patois *gateyou*, f. -se J 4 XVIII<sup>e</sup> s. RA.; — **gatayou**, f. -z J 6 GUÉ., VA., -quèyou 50, 54, 6 VA.; — **gòtelyou**, f. -zè N 30.

2

|| Adj. et s. **1°** Chatouilleux, sensible au chatouillement, ou sans précisions (Vd-V et F spor., N C.-aux-F., Cern.-Pég., Brév., Ch.-de-F., J spor.; anc. J; fr. rég. F GR.). *Lé fémalā son din la rālya* [règle] *plyə gatòlyāuzə tyé lèz òmo*, les femmes sont en général plus chatouilleuses que les hommes (Vd Blon.). *L è tan gatòyā dējā lè pi!* il est si chatouilleux sous les pieds! (F Roche). *S k'èl ā gatouèyou! an n'i ōj touchi lé yindr lā toué di kō*, ce qu'il est chatouilleux! on n'ose pas lui toucher les glandes autour du cou (J Épauv.). || En parlant d'un animal (Vd Penth., V Nend., Sav., J Charm.). *Si tchva ā gatyou*,

ce cheval est *g.* (J Charm.). *Atsə gatoulḡouja*, vache qui supporte mal les attouchements lors de la traite (V Nend., var. J Charm.). *Chon gatoulou fou kòufiron, dri ky'oun èjé tḡtsé avoqéi oun fətou, chḡrton dou bḡgan*, ces grillons sont chatouilleux, dès qu'on les touche avec un brin d'herbe, ils sortent de leur trou (V Sav.). Autre ex. sous *férir* I, 6<sup>o</sup> 1. **2°** Susceptible, pointilleux, irritable (Vd Leys., Penth., Pailly, Aubers., V Isér., Nend., F Gruy. spor., Rom., Hte-Glâne L'HOMME, J Vermes, Pleigne), grincheux (V Isér.). *Kan lè brāntḡḡ pāson su on vāzin gatòlyao*, *lè fḡ ron-nyī pḡ avai la pé*, quand les branches surplombent le terrain d'un voisin susceptible, il faut les couper pour avoir la paix (Vd Penth.). *Avḡ sóvan kā* [suivant qui] *oun pu pā ḡḡālyè, son troua gatòlyèu*, avec certains on ne peut pas plaisanter, ils sont trop grincheux (V Isér.). **3°** Délicat, d'un problème (Vd Cuves, F Gru.). *Chin ly è gatòlyā*, c'est une affaire difficile à arranger, litt. cela est *g.* (F Gru.).

Dérivé en -ōsu de *gatòlyi*; FEW, II, 511a et b; DURAFFOUR, *Gloss.* n° 4915. Pour le fr. rég. F GR. -eur, cf. parallèles *chatouilleux* (sous *chatouilleux*), *pointilleux* (sous *pointilleux*). Mül.

3

## 2.3 Particularités typographiques

Les articles possèdent une richesse typographique très importante. Non seulement la typographie véhicule des indications linguistiques mais la transcription phonétique des patois est mise en évidence par un système de transcription très fin et bien plus élaboré que le système phonétique international.

**gatòlyao**, -aou, f. -**za**, -ja Vd 51, 61, 64, 9 MAT., Vd BR., DUM., -āu 31, -(l)yā F 1, 33, 3 L'HOMME, 45, 63, -āeu V 32, 36, -u N 31, 41, -zao Vd 20, 21, -é 15 a, -ólyók V 71 KE. (f. -*ḡouja*), -òulèu 47 GABBUD (-*oloou*, -ly- BJERROME), -ou 60, -oulyè 50, BA., -ök 75 JJ., -lööü 51 A. PRAZ, -uyóouk 75 mod. (f. -*ḡouja*), -alyè 23 (f. -*ḡeuz*), -yœu 35, -zè 18, -òæ Vd 14 (-òè

L'en-tête *gatòlyao* est un mot patois car il est écrit avec une police grasse italique.

La forme *gatólyók* a été recensée à Hérémence<sup>3</sup>.

A. PRAZ fait référence à Arsène Praz, patoisant de Nendaz et « source » de la variante phonétique *gatoulööü*.

<sup>3</sup> Le Glossaire possède sa propre nomenclature pour les localisations. Ici V 71 correspond à Hérémence



La typographie des en-têtes indique sa nature :

<b>FER</b>	pour les mots empruntés au français et pour les mots patois qui ont un équivalent français;
<b>FERRONNIER</b> □	pour les mots empruntés au français qui ne sont pas publiés en raison de leur peu d'intérêt;
<i>fèralyôn</i>	pour les mots patois qui n'ont pas d'équivalent français;
« <i>ferran</i> »	pour les mots patois qui reproduisent la graphie des sources;
<b>ferrature</b>	pour les mots du français régional, les mots anciens et les noms propres locaux;
[ <b>ferrée</b> ]	pour les mots patois ou anciens classés sous la forme qu'ils auraient en français régional.

Dans la seconde partie (le corps) de l'article, Une définition est illustrée par des citations patoises, accompagnée de leur traduction et leur localisation.

|| Adj. et s. **1°** Chatouilleux, sensible au chatouillement, ou sans précisions (Vd-V et F spor., N C.-aux-F., Cern.-Pég., Brév., Ch.-de-F., J spor.; anc. J; fr. rég. F GR.). *Lé fèmalə son din la rālya* [règle] *plyə gatòlyāuzə tyé léz òmo*, les femmes sont en général plus chatouilleuses que les hommes (Vd Blon.). *L è tan gatòyā dèjə lè pi!* il est si chatouilleux

La richesse phonétique des patois se traduit par un système de diacritique très élaboré et très complet. A titre d'exemple, les différentes diacritiques pour la voyelle e

*e e ē ē ě ě*  
*è è è è ě ě*  
*é é ě ě ě ě*

Cette richesse typographique a nécessité la création de polices de caractères spécifique au GPSR. Ces polices ont été réalisées par l'ANRT, *Atelier National de Recherche Typographique* de Nancy (Fr.).

## 2.4 Rédaction des articles

La rédaction des articles a débuté en 1924. Elle s'est faite durant de nombreuses années avec des caractères au plomb. Les machines à écrire ont ensuite pris le relai, avant l'arrivée de l'ordinateur personnel. Aujourd'hui, la rédaction se fait en utilisant Microsoft Word. La saisie des signes spéciaux du Glossaire se fait par des macros spécifiques fournis dans le *Ruban* d'un modèle Word spécifique. Depuis peu, un clavier virtuel complète les outils de rédaction.

## 2.5 Partenariat avec la Heg-Arc

La HEG-Arc (Haute école de Gestion Arc à Neuchâtel) et le GPSR sont des partenaires de longues dates puisque la première collaboration a démarré en 1998.

Ce partenariat vieux de plus de 20 ans a permis d'élaborer ce que le GPSR appelle la « BDD ». Il s'agit d'un système d'information contenant les éléments-clés du Glossaire. Les objectifs de cette BDD sont :

- Capitaliser l'information, la connaissance acquise par le travail d'analyse et de rédaction
- Offrir aux rédacteurs un outil de recherche par les éléments-clés
- Produire des index intégrés en fin de tome



Cette BDD est accessible par internet par une application web qui a été développée par la Heg-Arc<sup>4</sup>.

Les éléments technologiques à préciser sont :

- La BDD est persistée dans le SGBD-R Oracle depuis 2015 (précédemment dans une base de données Microsoft Access).
- Une application Microsoft Access existe depuis 1998 et permet aux rédacteurs du Glossaire de saisir les éléments de cette BDD. Elle offre également des impressions importantes comme les index de fin de tome, des impressions de contrôle de saisie ou de nomenclature.
- L'application web a été développée à l'aide d'Apex<sup>5</sup> (Application Express). Il s'agit du RAD (*Rapid Application Development*) fourni par Oracle, gratuit et intégré dans toute distribution du SGBD-R. Ce RAD permet de développer des applications web d'entreprise, axées sur les données. L'interface utilisateur est entièrement générée dynamiquement. Des développements spécifiques peuvent être réalisés en javascript pour le côté client ou en PL/SQL pour les traitements côté serveur.

Les choix technologiques ont tous été fait tout au long de ces années en fonction des compétences fortes de l'équipe de projet.

<sup>4</sup> <https://portail-gpsr.unine.ch/>

<sup>5</sup> <https://apex.oracle.com/fr/>

### 3 Intégration des 40'000 articles dans l'application Web

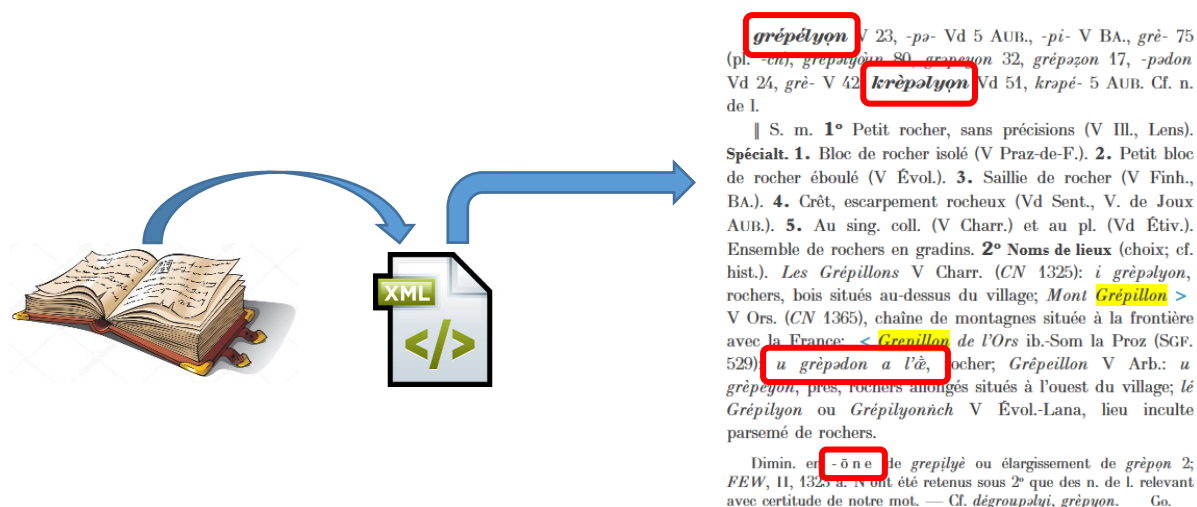
L'accessibilité tant à un large public qu'au monde scientifique est primordial pour mettre en valeur ce patrimoine linguistique de la Suisse romande. Le projet qui nous intéresse ici consiste donc à mettre en ligne l'intégralité du Glossaire déjà rédigé et d'y permettre des recherches "plein-texte".

### 3.1 Rétrodigitalisation du Glossaire

Etape préalable à la réalisation de ce projet, la numérisation des articles du Glossaire. Les articles actuels, regroupés dans 8 tomes (plus de 7000 pages rappelons-le) ne sont disponibles que dans un format papier.

Le GPSR a initié en 2014 le projet *Rétrodigitalisation du GPSR* : "[...] grâce au soutien financier de l'Académie suisse des sciences humaines et sociales (ASSH), ce projet a pour objectif de créer une version rétrodigitalisée du *Glossaire des patois de la Suisse romande* (GPSR) qui puisse être publiée sur Internet et ainsi être accessible à un vaste public."<sup>6</sup>

La rétrodigitalisation est assurée par l'Université de Trèves (All.) et se traduit par la mise à disposition du Glossaire au format XML. A partir de là, il nous incombe d'une part de reconstituer les articles pour permettre leur affichage dans un navigateur web et d'autre part d'indexer ces articles pour permettre des recherches de type "plein-texte".



### 3.2 Objectifs

Les deux objectifs sont donc :

- **Afficher** dans un format textuel les articles dans l'application web existante.
- **Permettre** des recherches plein-textes dans les articles : saisir des mots-clés, afficher les articles correspondants avec une mise en contexte et une surbrillance des mots-clés

<sup>6</sup> Rétrodigitalisation du GPSR [en ligne]. [Consulté le 10.01.2018]. Disponible à l'adresse : <https://www.unine.ch/islc/home/recherche/glossaire-des-patois-de-la-suiss/retrodigitalisation-du-gpsr.html>



### 3.3 Contraintes

Le GPSR utilisant une typographie et un système phonétique très riche, il est capital de respecter les exigences suivantes :

- Les articles au format textuel dans l'application web doivent reproduire la typographie originale. Aucune perte typographique ne peut être envisagée. Des divergences typographiques peuvent apparaître (plusieurs décennies de rédaction), dans ce cas il faudra harmoniser les pratiques dans les articles.
- La recherche textuelle doit pouvoir s'affranchir de la casse des caractères mais également des nombreux diacritiques.

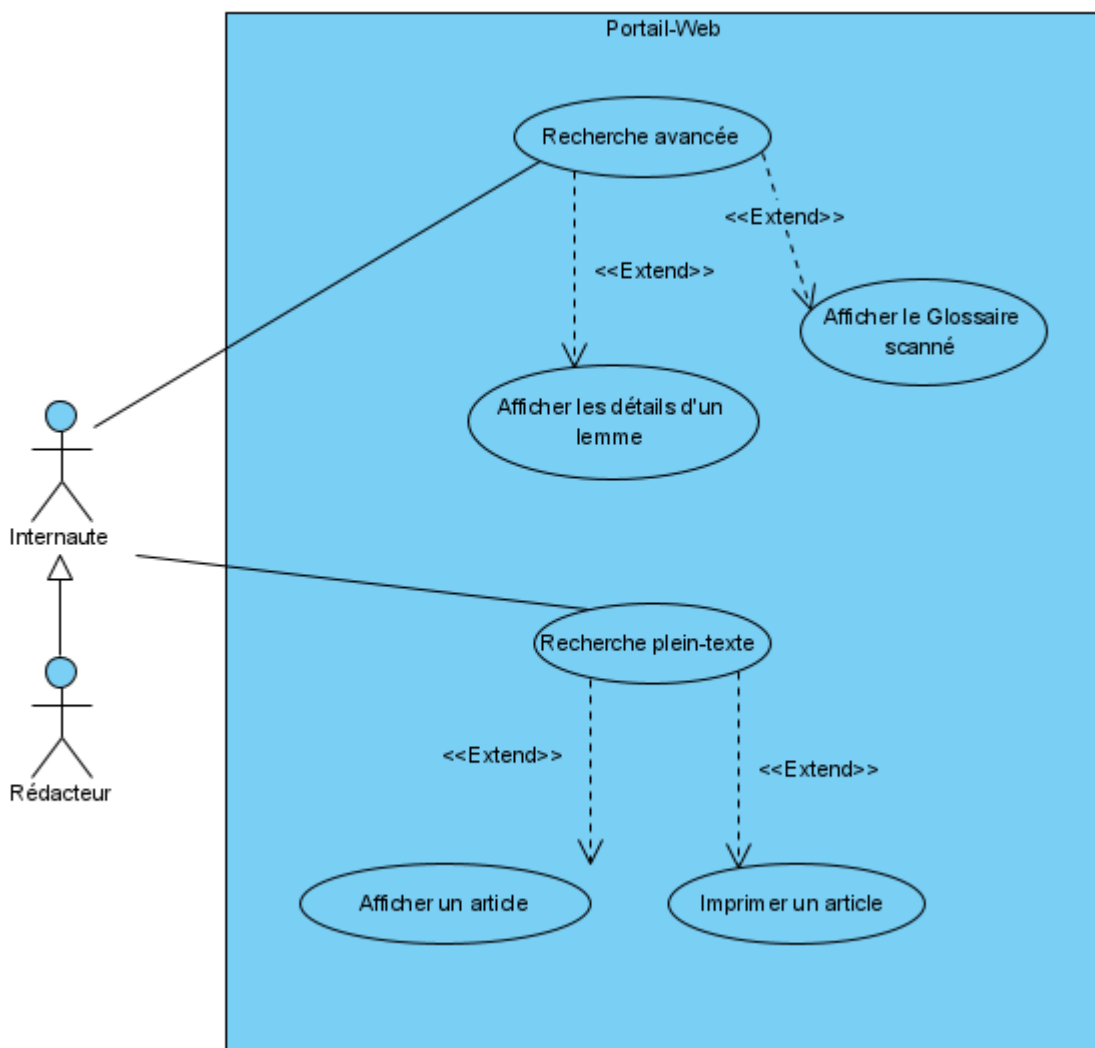
### 3.4 Intégration dans le SI existant

Les articles et la fonctionnalité de recherches doivent être intégrés dans l'application web actuelle.

## 4 Recherche textuelle – Analyse des besoins

### 4.1 Use case système

Le portail web actuel offre le service de recherche avancée, permettant d'interroger la BDD. Le nouveau service à offrir sera la recherche plein-texte qui se basera sur les articles, eux-mêmes devant être intégrés dans le SI existant.



## 4.2 Maquettes

Etape primordiale dans tout projet, le maquettage permet non seulement de mieux comprendre les besoins du client, mais permet de lui faire des propositions et d'anticiper les problèmes.

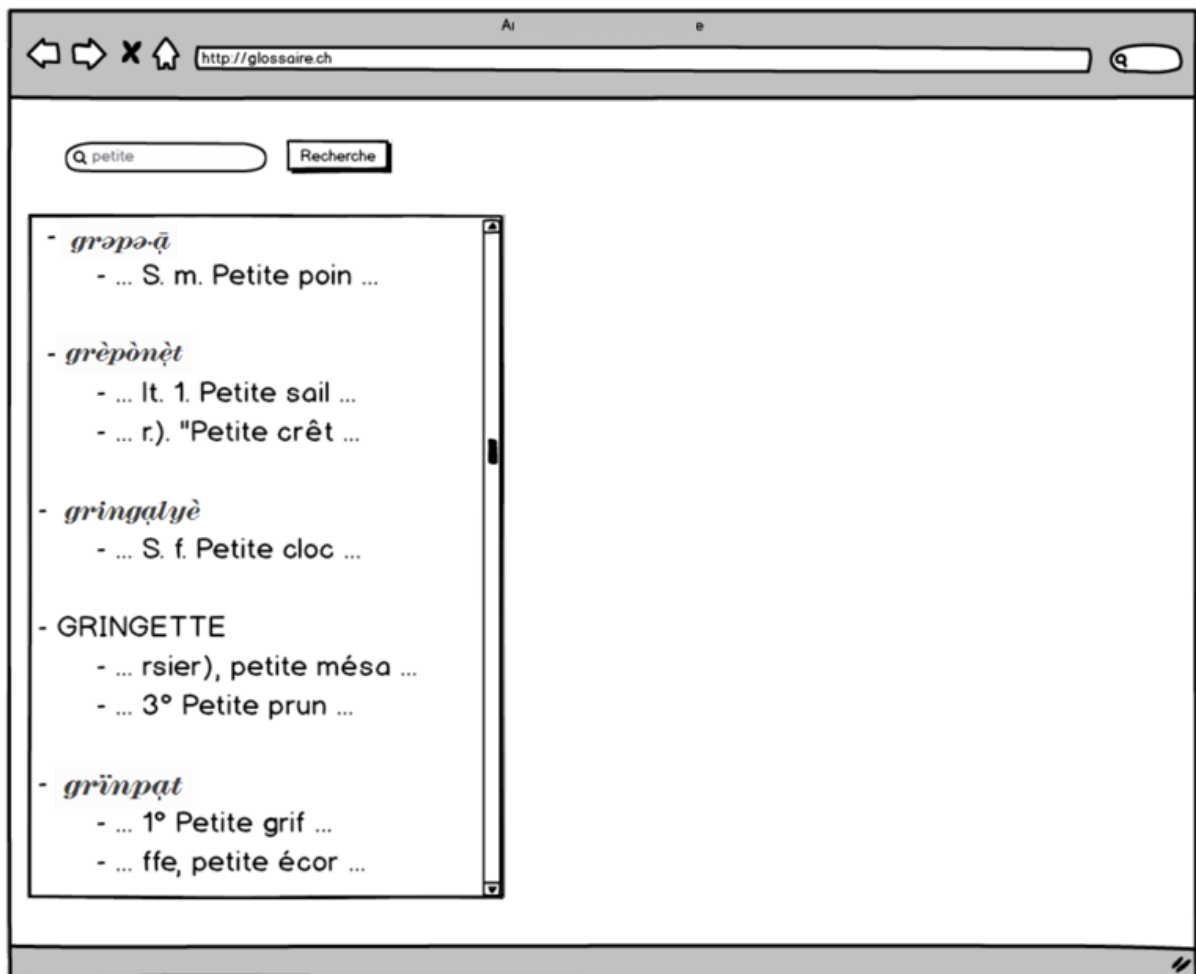
Des maquettes ont été présentées aux rédacteurs afin de bien expliciter notre compréhension de ses besoins et de formuler des solutions à ceux-ci.

### 4.2.1 Recherche par mots-clés

L'utilisateur va pouvoir saisir un ou plusieurs mots-clés.

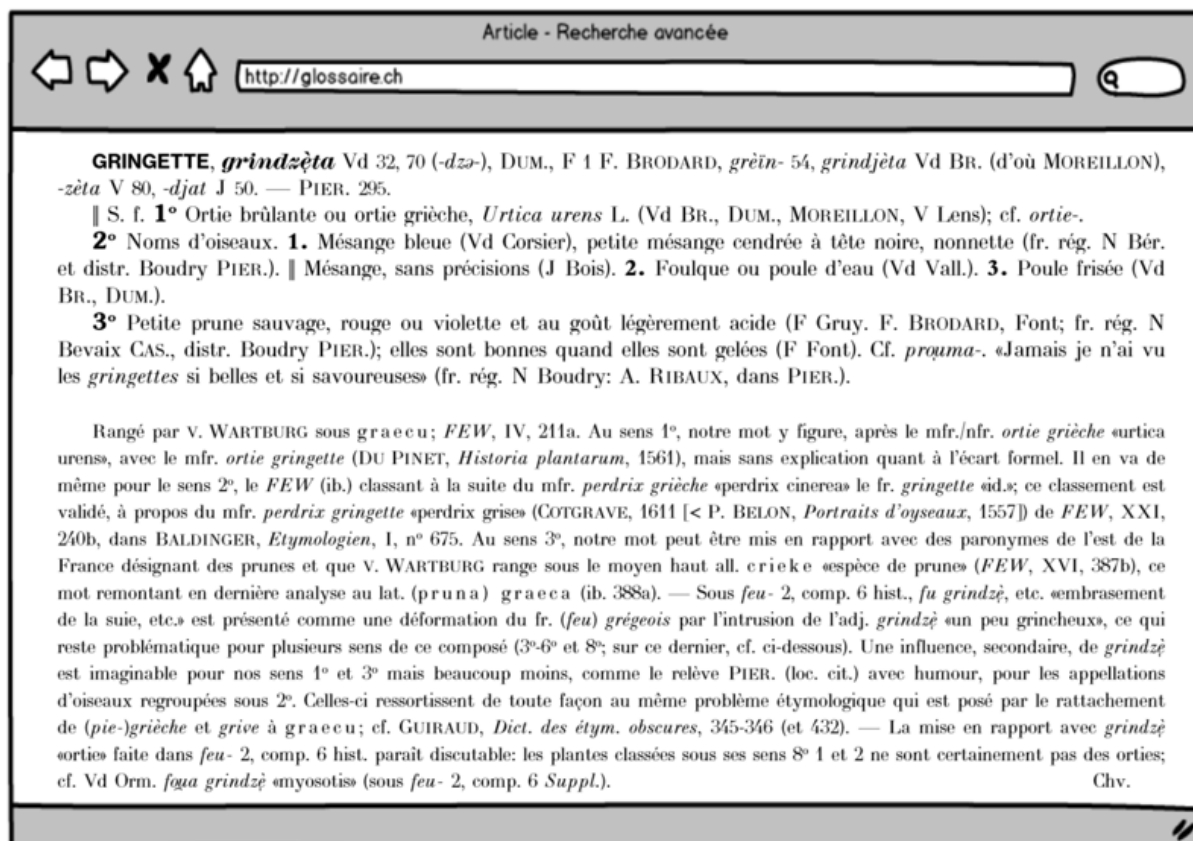
Le résultat de la recherche sera la liste des articles qui contiennent ce(s) mot(s)-clé(s)

Chaque article sera affiché par son lemme, le mot vedette, mais aussi par une mise en contexte des mots-clés. La mise en contexte correspond au contexte dans lequel les mots-clés existent, c'est-à-dire quelques mots avant et après.



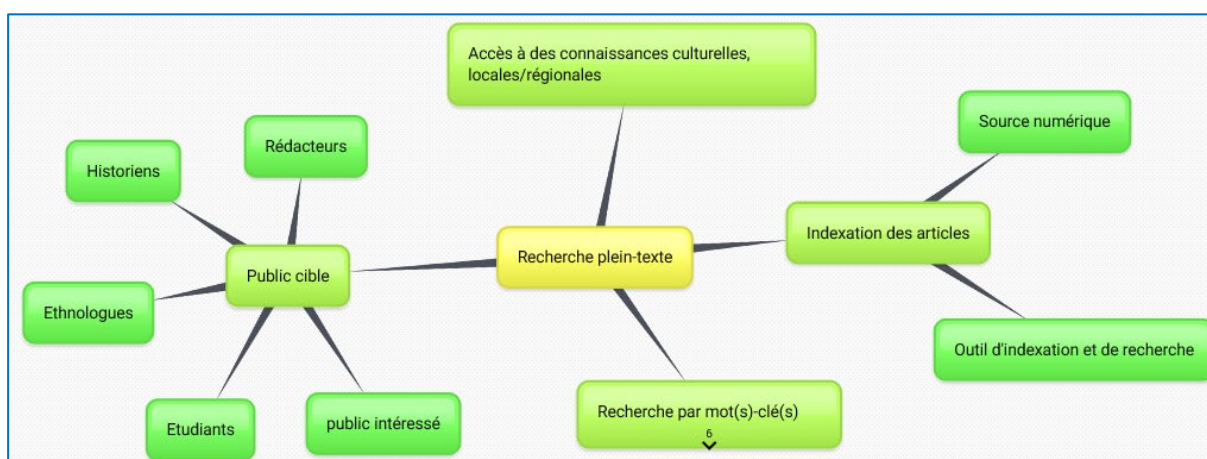
### 4.2.2 Affichage d'un article

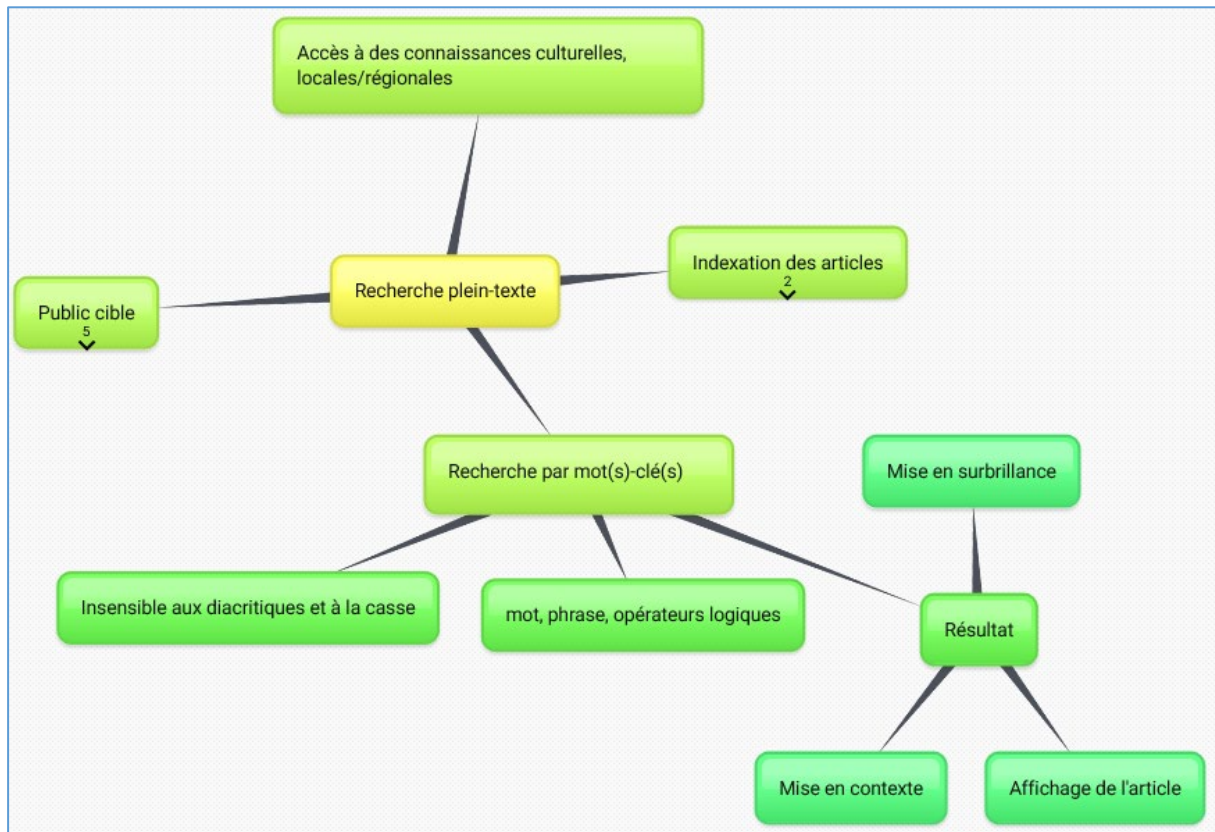
Lorsque l'utilisateur clique sur un lemme, l'article en question s'affiche dans une nouvelle fenêtre/onglet du navigateur.



### 4.3 Synthèse des besoins

Les besoins ont été synthétisés ainsi :





## 5 Conception de la solution à développer

L'analyse des besoins a permis de mettre en évidence que les articles devront être intégrés unitairement dans le SI existant et affichés dans une interface web. Il est temps de s'intéresser à la conception de la solution à développer.

### 5.1 Persistance des articles

Le choix stratégique le plus important et impactant toutes les fonctionnalités et donc tout le développement est le choix de la persistance des articles :

- Où seront stockés les articles ? Dans la base de données même ou dans le système de fichier du serveur ?
- Dans quel format ?

S'intéressant à l'emplacement physique des articles, deux alternatives sont possibles : dans une colonne d'une table dans le SGBD-R ou à l'extérieur du SGBD-R, sur le système de fichier du serveur. Nous ne nous attarderons pas sur l'évaluation de ces possibilités. Nous avons retenu comme choix d'intégrer les articles dans le SGBD-R, ceci pour bénéficier des mécanismes de journalisation déjà mis en place dans la BDD. D'un point de vue fonctionnelle, il n'y a aucune différence qui a été relevée.

Le second critère, le plus délicat, doit déterminer dans quel format seront stockés les articles. Ce choix va être guidé par 3 considérations :

- Format des articles retrodigitalisés
- Stratégie d'affichage des articles dans l'interface web, en fonction du point suivant
- Fonctionnalités d'indexation et de recherche textuelle

Les articles ont été rétrodigitalisés par l'Université de Trèves et nous ont été fournis dans un format XML (nous avons le choix entre RTF et XML). La stratégie d'affichage des articles révèle deux alternatives possibles :

- 1) Stockage des articles dans un format XML et affichage html par une transformation XSL dynamiquement, lors du rendu de l'article dans la page web.
- 2) Stockage de l'article au format xhtml dans la base de données. La restitution sera naturellement sans transformation.

	Avantages	Inconvénients
Stockage XML Rendu html	Séparation entre les données et la technologie de présentation. Intéressant à long terme, en cas d'évolution technologique	Nécessite une transformation automatique lors de l'affichage de chaque article, ce qui peut dégrader les temps de réponses
Stockage xhtml	Affichage de l'article tel qu'il est récupéré dans la base de données, sans traitement	Dépendance vis-à-vis du langage de présentation (html).

Sachant que les articles qui seront stockés dans la base de données seront obtenus par une chaîne de traitement automatisée, nous avons fait le choix de stocker les articles au format xhtml dans la base de données. Ainsi les performances seront meilleures et en cas de changement de langage de présentation (ce qui ne risque pas de se faire avant de nombreuses années...), nous reprendrons notre processus de génération des articles et adapterons le format de présentation.

## 5.2 Indexation, recherche textuelle et diacritiques non-standards

L'indexation des articles sera réalisée par Oracle Text. Oracle Text est une fonctionnalité native du SGBD-R Oracle permettant d'indexer des documents et d'effectuer des recherches plein-texte.

Toutes les fonctionnalités à mettre en œuvre dans ce projet sont prises en charge par Oracle Text. D'autres alternatives ont été envisagées, telles que Solr/Lucene, mais n'ayant aucune compétence technique dans ces outils, contrairement à Oracle Text, nous les avons écartés, malgré leur qualité et efficacité reconnue.

Dans les contraintes mentionnées au chapitre 3.3, il a été mentionné que la recherche plein-texte doit s'effectuer en n'étant pas sensible aux diacritiques, ni aux majuscules. Cela signifie que le processus d'indexation d'Oracle Text doit être capable d'indexer les mots des articles en désaccentuant toutes les lettres accentuées, ce qu'il sait très bien faire de manière native.

Mais qu'en est-il des signes spéciaux du Glossaire ? Pour être très précis, un certain nombre de signes spéciaux du Glossaire ne sont pas prévus par la norme Unicode et ils ont donc dû être placés, lors de l'élaboration de leurs polices de caractères, dans une des zones privées que prévoit Unicode.

Par exemple la voyelle *ǎ* doit être indexée avec la lettre de base *a*.

Une étude préalable d'Oracle Text a dû être effectuée afin de s'assurer qu'il soit bien capable d'indexer des documents contenant des caractères spéciaux.

Ceci est fort heureusement possible, grâce au logiciel Oracle Locale Builder qui permet de spécifier et paramétrer des *tris linguistiques* (nous y reviendrons dans la partie liée au développement).

## 6 Génération des articles au format xhtml

### 6.1 Analyse des sources XML

Les tomes ont été retrodigitalisés et donnent lieu chacun à un fichier xml, un par tome.

L'analyse des fichiers XML de Trèves a permis de cerner avec précision le balisage mis en place dans les articles. Le balisage est principalement d'ordre typographique, mais certaines balises révèlent une indication structurelle ou linguistique. Il est à noter qu'aucun schéma XSD n'a été fourni par Trèves...

Extrait d'un article :

```
008"/><art vol="7_2" ln="1140. 008"><P> <lem3><b><i>fyol</i></b>
009"/><s4><b>1</b><sup>o</sup> Épicéa, <i>Picea Abies</i> (L.) Kar
010"/></s4><s4><b>2</b><sup>o</sup> Petit buisson (N C.-aux-F.). <
011"/><i>Fiolet</i> N Fleur. (<i>CN</i> 1183), pâturage; <i>le Fio
012"/>Sulpice, bois; <i>le Fiolet</i> J Genevez: <i>ā fyólè,</i> c
013"/><i>les Fiolets</i> J St-Br. (<i>CN</i> 1105): <i>é fyólè,</i>
014"/>forêt. □ <i>Clos Fiola</i> B Mall.: <i>i tyó fyòla,</i> bâti
```

Ci-dessous je liste les balises XML utilisées dans le balisage des articles avec une brève description :

- Balisage des articles eux-mêmes (début-fin)
  - Balise <art ln="..." lem="..." type="..." >
- Balisage des différents types de lemmes (français, patois, anc. fra. rég, renvois, ...)
  - Balises <lem1> à <lem7>
- Balisage des différents styles typographiques (majuscule, grasitalique, gras, exposant, petites capitales, retrait de paragraphe, etc...)
  - <AN> → majuscule, utilisé avec <lem1>
  - <b>, <i>, <sup> → gras, italique, exposant
  - <c> → petite capitale
  - <spr> → étendu
  - <Am1>, <Am2> → polices plus petites (2 niveaux)
- Balisage des différentes parties d'un article (variantes phonétiques, corps, dérivés, commentaires scientifiques, encycl.)
  - <varpho> → variante phonétiques
  - <deriv> → dérivés
  - <hist> → commentaire scientifique
  - <encycl> → partie « encyclopédique » dans le commentaire scientifique
  - <p> → paragraphe
  - <semx>, <sem1a>, <sem1e> → les sens
- Balisage des différentes numérations de sens
  - <s1> à <s6>
- Composés
  - <lemc>, <lemsub> → lemmes liés à des composés
  - <subart> → article qui est un composé dans un article principal
- Balisage des sommaires et des niveaux de sens (y.c. mise en forme)
  - <sommaire>
  - <Z>, <E>, <negEZ>, <Ep1>, <Ep2>
  - <nivS1> à <nivS6>, <niv1> à <niv4>

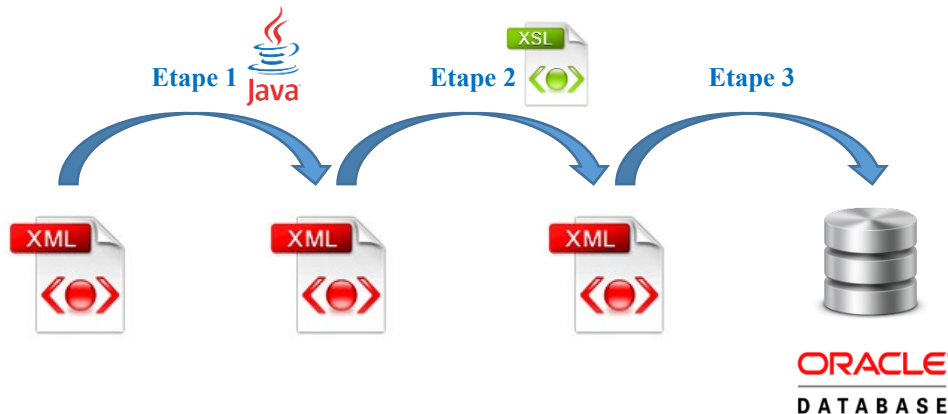


- Balisage de l'auteur
  - <R>
- Balisage des particularités (poèmes, tableau, etc...)
  - <Biaoge>, <BiaogeHist> (dans les 1ers volumes uniquement)
    - <tr>, <td>, <rTab>
  - <shi> (dans les 1ers volumes)
  - <xiamian>
- Balisage des figures et des légendes
  - <figure>
  - <image>

## 6.2 Chaîne de traitement

Une chaîne de traitement automatisée a été mise en place et se déroule en plusieurs étapes.

Le but est de pouvoir régénérer l'intégralité des articles du Glossaire de manière automatisée. Ceci est primordial dans l'évolution du contenu des articles. En effet, si le Glossaire souhaite pouvoir apporter de nouvelles fonctionnalités dans la recherche d'articles ou améliorer le contenu des articles (en ajoutant des indications supplémentaires, des liens hypertextes sur des contenus externes...), il faudra modifier la structure ou le contenu de tous les articles. Parallèlement à cela, il a été décidé que les articles ne seront jamais modifiés par les rédacteurs, afin de garantir l'objectif cité précédemment. En cas d'erreur dans un article, je corrigerai cette erreur en l'intégrant dans la chaîne de traitement et régènerai l'article en question.



### 6.2.1 Etape 1 : Nettoyage et correction du contenu

Les traitements réalisés sont :

- Analyse du contenu par des requêtes Xpath : comptage des éléments et identification d'éventuelles erreurs de balisages.
- Corrections des erreurs de rétrodigitalisation (erreur de retranscription, de structuration d'un article...)

- Gestion des coupures de mot entièrement automatisée : le processus de rétrodigitalisation s'est fait de telle manière que chaque fin de ligne de chaque article a été balisé en tant que tel. Les mots coupés en fin de ligne sont donc également coupés dans le fichier source. Il a fallu reconstituer ces mots en identifiant les mots à reconstruire et ceux pour lesquels le tiret devait être conservés. Pour cela, j'ai utilisé un fichier texte contenant une liste de plus 336500 mots utilisés dans la langue française que j'ai trouvé sur internet<sup>7</sup>. Pour les mots patois qui sont dans ce cas, j'ai transmis au Glossaire une liste de mots patois dont il faut indiquer s'il faut conserver ou non le tiret. J'ai ensuite complété le fichier de mots français avec les mots patois

*flamèyè*, -ë, -é, -i, -më, -mé- V 50, *flanm-* 35, *ham-* 21, *hlanm-* 51 A. PRAZ (-é-é), *hlanm-* 46-47, *shanm-* Vd 16 (-i), V 36, *shanm-* Vd 15 a ISABEL (-i), *shanm-* V 43, 5 BER. (-é-é).

|| V. intr. 1<sup>o</sup> Flamber, flamboyer, faire de **flam-**mes, de belles flammes (partout). *Lə fəua şaməyè kan t'oua sòşè*, le feu flambe quand le vent souffle (Vd Ollon ISABEL). *Lə fəua haméyè-t-ə rè toa* [rien trop]? le feu ne fait-il pas de trop grandes flammes?

Le mot « flammes » étant coupé dans la source XML, il faut le reconstruire, pour pouvoir l'afficher correctement mais également l'indexer pour la recherche plein-texte

On voit ci-dessous la source XML :

```
<eol/><semx><P>|| V. intr. <s4><b>1</b><sup></sup> Flamber, flamboyer, faire des flam-  
<eol/>mes, de belles flammes (partout). <i>Lə fəua şaməyè kan</i>
```

Chippis, *flama* 86-86 a MEYER, *flanma*, -anm- 22 MÜLLER, 22 a J.J., 2 Éviennaz ZI., 35, 42, 45 var., 4 Bruson, Médières, Sarreyer, Verbier, 50, 60, 62, 6 Miserier, Salins, 7, 83 var., 84, 85, 8 Granges St-Léonard, -anm B 33 et 3 Grandval ZI., *flama*, -a, -a- Vd 80, 8 Bullet, V 10, 11, 14, 1 Bouveret, 20, 22 corr.,

Par contre, ici « St-Léonard » ne doit pas être reconstruit sans le tiret...

- Homogénéisation de la typographie entre les différents volumes
- Suppression des retours à la ligne
- Intégration de tirets insécables, espaces insécables dans la numérotation des sens, etc...
- Identification des références faites au FEW (dictionnaire étymologique français) afin d'inclure un lien hypertexte sur le dictionnaire en ligne

Tous ses traitements sont réalisés par programmation (find-replace, regex, etc...), à l'aide d'un programme Java.

Illustration avec divers traitements

```
//Finalisation traitement coupure de mots
lignes = findReplace(lignes, "<eol/>", "-");
lignes = findReplace(lignes, "<eol/>", " ");

//Renvoi sur le FEW
//corrections
lignes = findReplace(lignes, " FEW,</i>", " </i><i>FEW,</i>");
lignes = findReplace(lignes, "<i>FEW,</i></Aml> <Aml>", "<i>FEW,</i> ");
lignes = findReplace(lignes, "<i>FEW</i></Aml> <Aml>", "<i>FEW</i> ");

//balisage
lignes = executeRegex(lignes, "<i>FEW,</i> \\((? ([ locitetn0-9 \\.)*)?([IVXC DLM]+(?:/[1-9])?) (, ) ([0-9]+) ( ?[ab] )?: n. [0-9]+) ?>");

//nettoyage (regex pas optimisé, il reste des balises lecteurFEW en trop. Ex : XVII, 234 a et b ou 234 a et b, XVI
lignes = executeRegex(lignes, "<lecteurFEW tome=\\([IVXC DLM]+(?:/[1-9])?) ?>\" page=\\\">([\\(\\), a-z]*)?</lecteurFEW>\", \"$2\");
```

<sup>7</sup> <http://www.pallier.org/liste-de-mots-francais.html>

### 6.2.2 Etape 2 : Transformation xhtml

La transformation en xhtml est effectuée par une feuille de style xsl. Le résultat reste un fichier xml, composés de balises <article>. Le contenu à proprement parlé des articles est stocké dans un CDATA dans le fichier XML.

Extraits du fichier xsl :

Création de la balise <article> et création d'un identifiant unique afin d'offrir un permalien dans l'application web

```
<?xml version="1.0"?>
<xsl:stylesheet version="2.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:output method="xml" indent="no" encoding="UTF-8"/>
  <!--génération finale du xml/xhtml -->

  <!--Changer la valeur avant de lancer la transfo ! -->
  <xsl:variable name="tome" select="8"/>

  <xsl:template match="/xml">
    <articles>
      <xsl:apply-templates/>
    </articles>
  </xsl:template>
  <xsl:template match="//art">
    <xsl:variable name="artPage" select="@ln"/>
    <article>
      <xsl:attribute name="idPage">
        <xsl:choose>
          <xsl:when test="contains(@ln, '.')"><xsl:value-of select="replace(substring-before(@ln, '.'), 'E', '')"/></xsl:when>
          <xsl:when test="contains(@ln, 'P')"><xsl:value-of select="substring-after(@ln, 'P')"/></xsl:when>
        </xsl:choose>
      </xsl:attribute>

      <xsl:attribute name="lem"><xsl:value-of select="@lem"/></xsl:attribute>

      <xsl:attribute name="orderNum">
        <xsl:choose>
          <xsl:when test="contains(@ln, '.')"><xsl:value-of select="($tome*100000000)
            + number(replace(substring-before(@ln, '.'), 'E', '')) * 1000
            + number(substring-after(@ln, '.'))"/></xsl:when>
          <xsl:when test="contains(@ln, 'P')"><xsl:value-of select="($tome*100000000)
            + number(substring-after(@ln, 'P')) * 1000
            + count(preceding-sibling::art[@ln=$artPage])+1"/></xsl:when>
        </xsl:choose>
      </xsl:attribute>

      <xsl:attribute name="redacteur"><xsl:value-of select="."/R//text()"/></xsl:attribute>
      <xsl:text disable-output-escaping="yes">&lt;![CDATA[&lt;article&gt;</xsl:text>
      <xsl:apply-templates/>
      <xsl:text disable-output-escaping="yes">&lt;/article&gt;]]&gt;</xsl:text>
    </article>
  </xsl:template>
```

Balises des lemmes

```
<!--Lemme-->
<!-- Français -->
<xsl:template match="//lem1/AN/b/text()">
  <span data-gpsr-section="varpho">
    <span class="enteteFrancais">
      <xsl:value-of select="replace(., '-', '##TIRET##')"/>
    </span>
  </span>
</xsl:template>
<!-- non rédigé -->
<xsl:template match="//lem2/AN/b/text()">
  <span data-gpsr-section="varpho">
    <span class="enteteFrancais">
      <xsl:value-of select="."/>
    </span>
  </span>
</xsl:template>
<!-- Patois -->
<xsl:template match="//lem3">
  <span data-gpsr-section="varpho">
    <xsl:apply-templates/>
  </span>
</xsl:template>
```

### 6.2.3 Etape 3 : Chargement des articles dans la base de données Oracle

Après avoir été transférés sur le serveur hébergeant le SGBD-R Oracle, les articles sont extraits du fichier xml obtenu à l'étape précédente et insérés dans la table *Articles*.

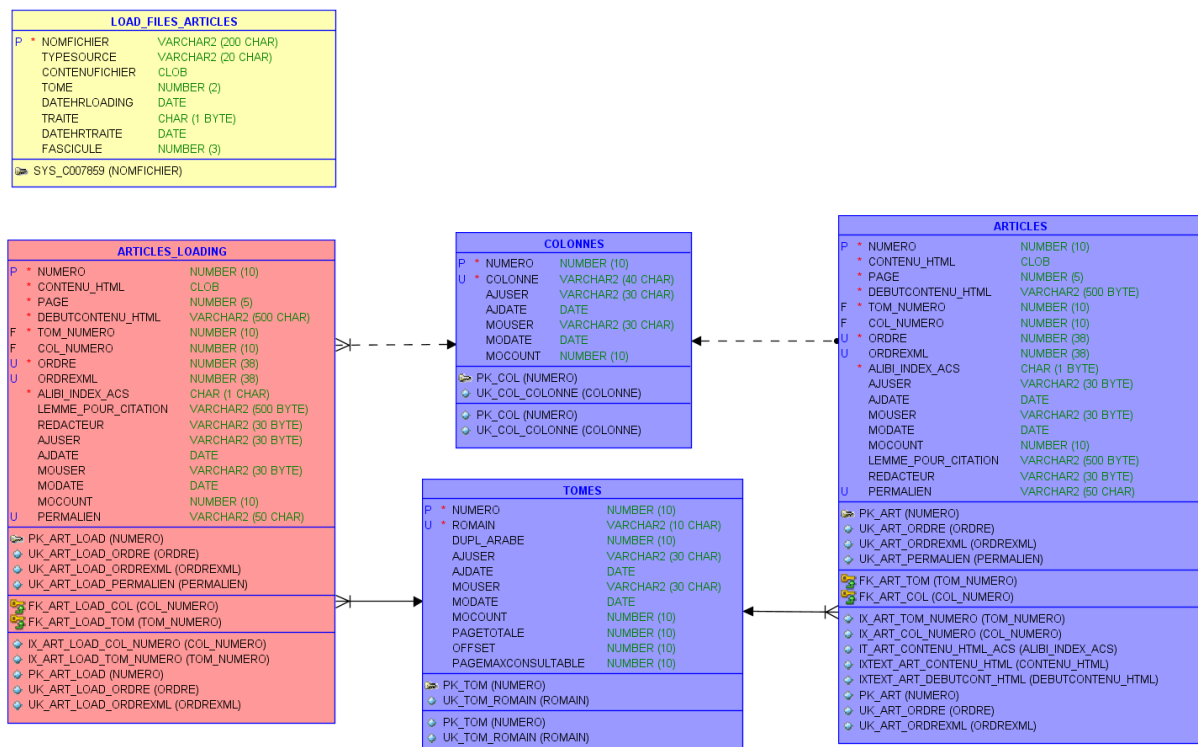
#### 6.2.3.1 Processus de création des articles dans la base de données

Le processus de chargement des articles se déroule en plusieurs temps. En effet, certains traitements doivent encore être réalisés afin de produire les articles dans leur version définitive.

- Chargement du fichier XML obtenu à l'étape précédente dans la table **LOAD\_FILES\_ARTICLES**.
- Extraction de chaque article stocké dans le fichier XML et alimentation de la table **ARTICLES\_LOADING**
- Pour chaque article :
  - Extraction du tome, de la page, de l'auteur
  - Extraction du lemme pour permettre une recherche ciblée sur celui-ci uniquement et créer le lemme pour sa citation
- Mise à jour de la table **ARTICLES**
  - Insertion des nouveaux articles
  - Mise à jour des articles existants

Tous ces traitements sont réalisés en PL/SQL, dans un package créé dans le schéma Oracle.

Ci-dessous, le modèle de données correspondant, réalisé à l'aide de la notation Barker (notation utilisée par le logiciel Oracle Data Modeler)



## 7 Développement de la recherche textuelle dans les articles

### 7.1 Création d'un tri linguistique propre au GPSR

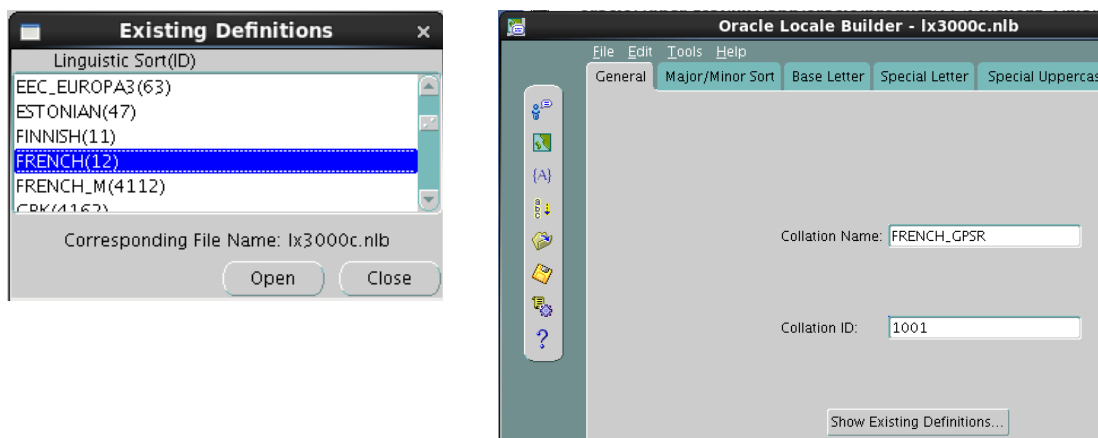
Comme mentionné au chapitre 5.2 il est capital que le processus d'indexation et la fonctionnalité de recherche plein texte indexe correctement les signes spéciaux, en particuliers ceux définis dans la zone privée Unicode.

Comme nous voulons pouvoir offrir une recherche sans tenir compte des diacritiques, il faut par exemple que la voyelle *ā* doit être indexée avec la lettre de base *a*.

Ci-dessous, de l'article *gatòlyāo*, le mot *rālyā* doit être indexé en *ralya*.

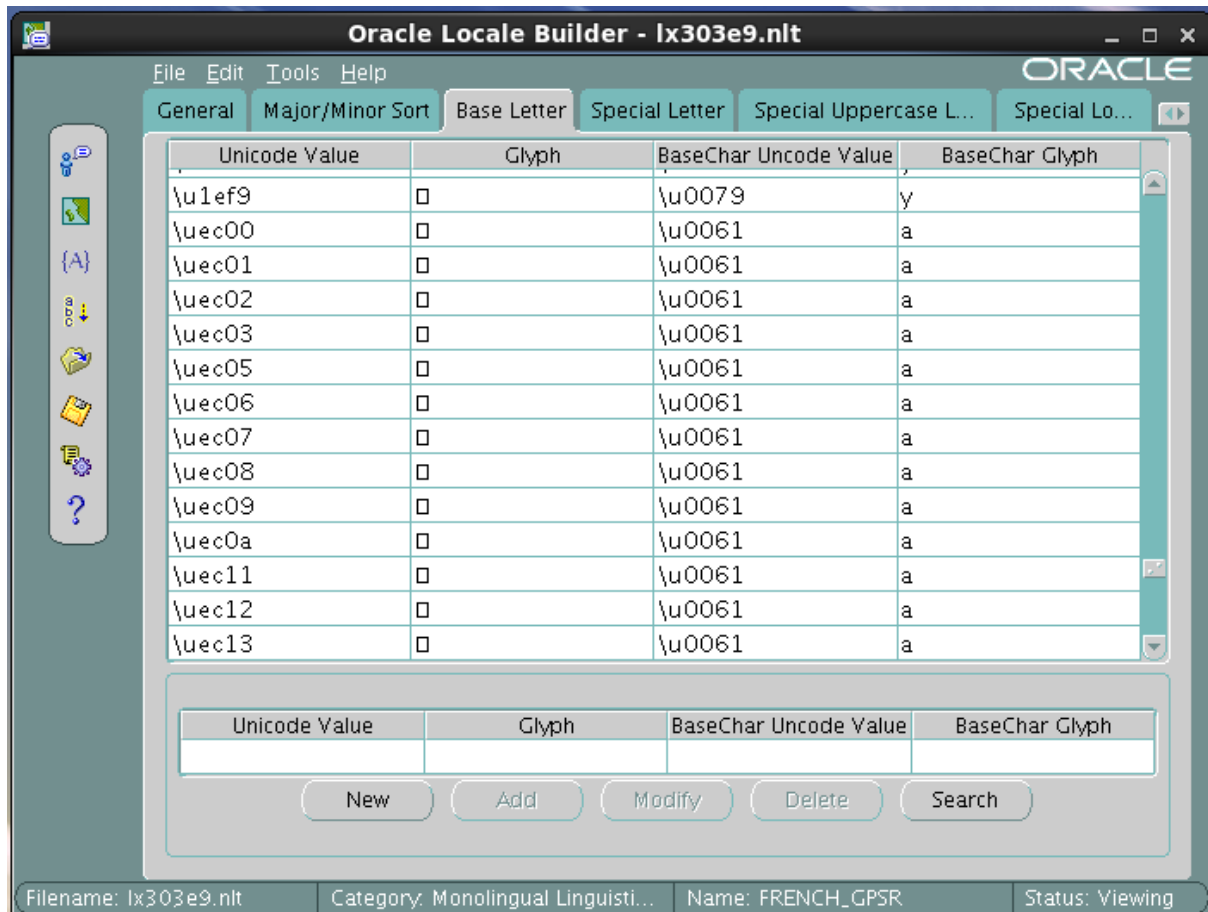
|| Adj. et s. **1°** Chatouilleux, sensible au chatouillement, ou sans précisions (Vd-V et F spor., N C.-aux-F., Cern.-Péq., Brév., Ch.-de-F., J spor.; anc. J; fr. rég. F GR.). *Lé fémalə son din la rālyā* [règle] *plyə gatòlyāuzə tyé léz òmo*, les femmes sont en général plus chatouilleuses que les hommes (Vd Blon.). *L è tan gatòyā dējə lè pi!* il est si chatouilleux

Pour ce faire, il faut utiliser le logiciel Oracle Locale Builder et créer un tri linguistique spécifique à ce projet. Ce nouveau tri linguistique sera créé sur la base du tri linguistique standard « FRENCH »



Ce tri linguistique sera complété en indiquant pour chaque caractère, y.c ceux de la zone privée, sa lettre de base





Les caractères ayant le point de code Unicode ECxx correspondent bien à des caractères placés dans la zone privée Unicode. Notre caractère *à* se trouve à EC09 que l'on peut voir ici, avec la lettre de base 'a'.

Une fois ce tri linguistique créé, il suffit de le déployer sur l'instance Oracle et y faire référence lors de l'indexation et la recherche.

## 7.2 Indexation des articles

L'indexation même des articles a été réalisée de la manière suivante :

```
ALTER SESSION SET NLS_SORT=FRENCH_GPSR;

BEGIN
  CTX_DDL.CREATE_PREFERENCE('LEXER_GPSR_ACI', 'BASIC_LEXER');
  CTX_DDL.SET_ATTRIBUTE('LEXER_GPSR_ACI', 'BASE_LETTER', 'YES');
  CTX_DDL.SET_ATTRIBUTE('LEXER_GPSR_ACI', 'MIXED_CASE', 'NO');
  CTX_DDL.SET_ATTRIBUTE('LEXER_GPSR_ACI', 'BASE_LETTER_TYPE', 'SPECIFIC');
  CTX_DDL.SET_ATTRIBUTE('LEXER_GPSR_ACI', 'SKIPJOINS', '()');
  CTX_DDL.CREATE_SECTION_GROUP('PATHGROUP_GPSR', 'PATH_SECTION_GROUP');
END;

CREATE INDEX ITEXT_ART_CONTENU_HTML ON articles(contenu_html)
INDEXTYPE IS ctxsys.context
PARAMETERS('DATASTORE ctxsys.default_datastore
FILTER ctxsys.null_filter
SECTION GROUP PATHGROUP_GPSR
LEXER LEXER_GPSR_ACI
STORAGE TABLESPACE_GPSR_ITEXT
STOPLIST CTXSYS.EMPTY_STOPLIST');
```

Dans les grandes lignes, on crée un index de type Oracle TEXT, en indiquant comme paramètres :

- BASIC\_LETTER YES → supprime les diacritiques
- MIXES\_CASE NO → insensible à la casse
- EMPTY\_STOPLIST → aucune *stop list* n'est indiquée, car tous les mots sont importants pour le Glossaire. En effet, le mot 'jamais', qui est un *stop word* dans notre langue, est aussi un article dans le Glossaire !

La 1<sup>ère</sup> instruction, ALTER SESSION, permet d'indiquer le tri linguistique à utiliser.

## 8 Développement de l'interface Web

### 8.1 Recherche par mots-clés

Destinée au grand public, la recherche rapide est une recherche "traditionnelle" par mots-clés. Les résultats sont les articles du Glossaire contenant les mots-clés, avec mise en contexte et surbrillance de ces mots-clés. L'affichage de l'article dans son intégralité contient aussi les mots-clés en surbrillance mais aussi une navigation entre ces mots trouvés.

Mots-clés

Si plusieurs mots sont recherchés, une opération ET est implicite : les guillemets permettent de rechercher une portion de texte. Il est possible de joker en saisissant l'astérisque.

Nombre de résultats trouvés : 12 article(s)

**ACCROCHER** (II, 95)  
..., *ma pour' anhyanna*, | *Akroutze tè à mon bredzon* », embrasse-moi, ma pauvre vieille

**armalyi** (I, 616)  
... et un gilet à manches très courtes, le *bredzon* : laissant les bras nus jusqu'au-dessus ...  
...courtes, parfois de velours, et un beau *bredzon* brodé, à manches bouffantes, sur lequel...

**bredzon** (II, 754)  
*bredzon* Vd 1 Ollon, 31, 32, 33 *Cont.* 1917 25, F...  
... formes *brezon* F 1722, *bergeon* 1719. Fr. *bredzon* Vd 31, F.  
..., 61. AC; douteux si sens 1<sup>o</sup> ou 2<sup>o</sup>). On *bredzon* *dè lan-na*, un gilet de laine (Montb.).  
...s, les domestiques (Vd, F). Dans F, le «*bredzon*» est devenu une pièce caractéristique d  
...relaz (voir *bolé* 1). On l'appelle aussi *bredzon* *a mandzèté*, br. à courtes manches (F Gru  
...*mè ly aruvè*, | *On va veyre montà* | *Di bi bredzon* *à mandzèté*, | *Di bi loyi brodâ*, le

**CORSELET** (IV, 345)  
... (Vd Orm. pat. et fr. rég.); cf. synon. *bredzon* 1<sup>o</sup>. On *kòrsalè dè tridzo blu*, un «corsel...

**Tome II page 754**

*bredzon* > Vd 1 Ollon, 31, 32, 33 *Cont.* 1917 25, F 1, 21, 30, 4 Épendes, 61, *bré-* Vd Dum., *bér-* F 63. Anc. formes *brezon* F 1722, *bergeon* 1719. Fr. < *bredzon* > Vd 31, F.  
|| S. m. 1 ° Gilet (F Montb., Grandv., Joux, vieilli). «1 veste, un *bergeon*» (F Gru. 1719. *Reg. not.* 2826, 450. AC). «Une robettaz ou soit *brezon* rouges» (F Neirivue 1722. *Reg. not.* 2746, 61. AC; douteux si sens 1<sup>o</sup> ou 2<sup>o</sup>). On < *bredzon* > *dè lan-na*, un gilet de laine (Montb.). | Auj. spécialt. Veste à courtes manches bouffantes, aux revers parfois brodés, que portent les armailis, le personnel des alpages, les domestiques (Vd, F). Dans F, le «< *bredzon* >» est devenu une pièce caractéristique du costume national. Voir fig. Gl. I, 617 et synon. *dzapa* F Vaulruz (sous *jupe*), *dzapon* Vd P. d'E. (sous *jupon*), *bolé* Vd Forclaz (voir *bolé* 1). On l'appelle aussi < *bredzon* > *a mandzèté*, br. à courtes manches (F Gruy.). «Le vin *dè mè ly aruvè*, | *On va veyre montà* | *Di bi < bredzon > à mandzèté*, | *Di bi loyi brodâ*, le vingt mai approche, on va voir monter de beaux br., de belles sacoches à sel brodées (F Gruy. *Étr. frib.* 1878, 58). 2 ° *Jupe* (F Alb. Co., Sugiez). | *Jupon* (F Sugiez).  
AEB. identifie ce mot avec afr. *haubergeon* «cotte de maille à courtes manches ou sans manches» qui, dès le XVI<sup>e</sup> s., a désigné des pièces d'habillement civil; cf. p. ex. GAY, *Gloss.* II, 15 et anc. F: «Porte des chausses jaunes et ung *aubergeon* de trige [triège] blanc» (F 1564. *Livre noir*, VI. AC). Cf. *haubergeon*. De.

La requête SQL pour cet exemple est très simple :

```
SELECT ... FROM ARTICLES WHERE contains(contenu_html,'bredzon')>0;
```

La mise en contexte et en surbrillance sont réalisées par des fonctions fournies par Oracle TEXT.

## 8.2 Recherche personnalisée

Plutôt destinée au monde scientifique (rédacteurs), cette recherche permet entre autres l'utilisation d'opérateurs logiques (AND, OR, NOT) et la recherche de proximité entre mots.

Recherche personnalisée : "v saxon" PROX raccard

Nombre de résultats trouvés : 1 article(s)

---

**étro** (VI, 924)  
 ... grange (V Mart. BA.). || Espace sous le **raccard**, utilisé comme étable, remise ou chambr...  
 ...étable, remise ou chambre à provisions (**V Saxon** et Sembr. Hunz.). **5.** Corridor d'une mai...

---

**1**

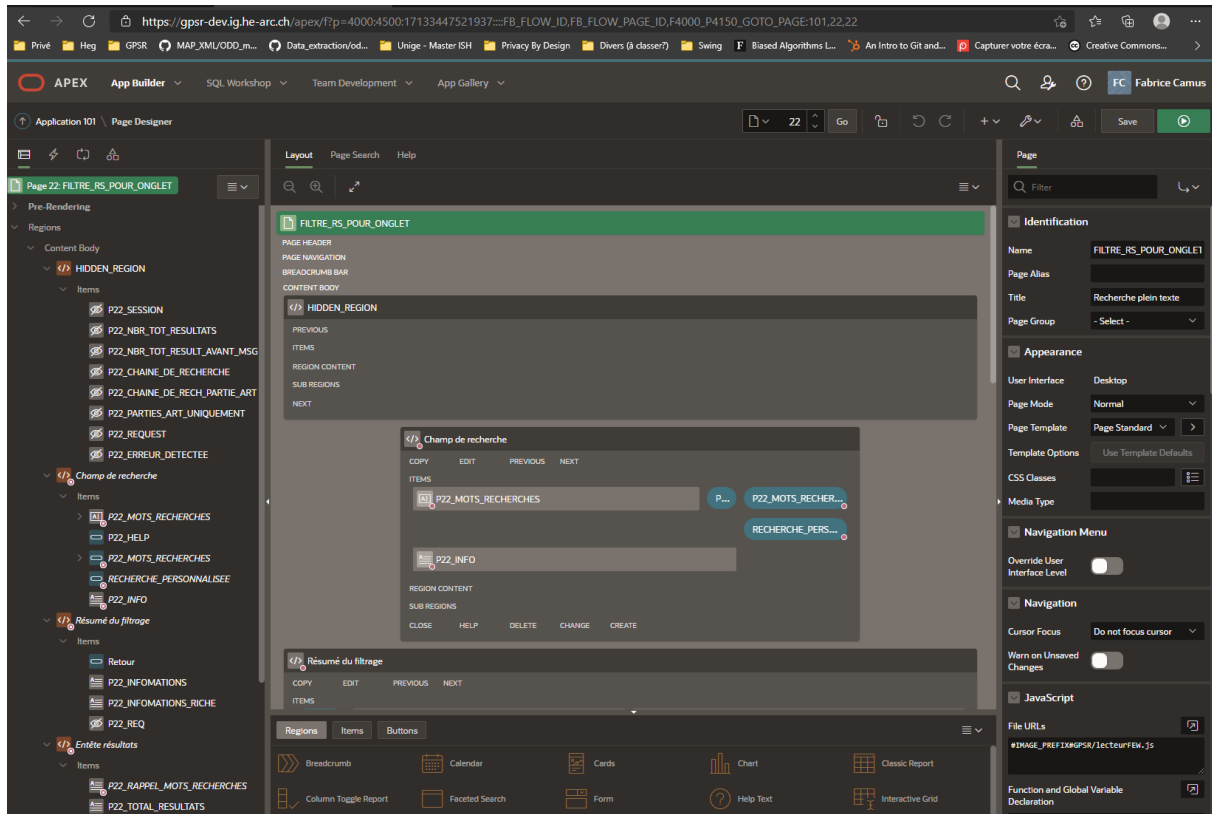
La requête SQL pour une recherche de proximité dans Oracle TEXT est :  
 SELECT ... FROM ARTICLES WHERE contains(contenu\_html,'near((v  
 saxon),raccard,10)')>0;

L'application web est disponible à l'adresse suivante : <https://portail-gpsr.unine.ch/>

### 8.3 Interface web de développement d'application

Apex est un RAD permettant de développer des applications web. Son interface de développement est elle-même développée en « Apex ». Il s'agit d'un développement RAD relativement traditionnelle, avec une interface graphique permettant de créer des pages et manipuler des composants graphiques.

Ci-dessous, la page de résultats de la recherche vue en mode développement :



## 9 Conclusion

Ce projet de grande ampleur s'est révélé être semé de challenges technologiques. Les choix technologiques et stratégiques furent les éléments les plus problématiques, car il ne fallait pas se fourvoyer dans ces choix vu les impacts sur tout le projet.

La mise en œuvre de la chaîne de traitement fut certes difficile, car relativement complexe de part le volume de données à traiter et les problèmes qui sont apparus au fur et à mesure. Fort heureusement, une méthodologie – en partie automatisée – de contrôle qualité réalisé en amont et durant le développement m'a permis de détecter le maximum de cas à corriger.

Ce projet a été un très grand défi technique mais ce que je retiens et que je n'avais pas perçu durant, c'est la plus-value qu'il a amené au Glossaire. Les rédacteurs m'ont fait comprendre, avec émotion pour certains, que c'était pour eux la plus grande avancée qu'ils avaient eue dans leur institution, car ce qui n'était tout simplement pas possible avant l'est devenu.

A peine la recherche textuelle mise à disposition que de nouvelles demandes d'évolutions ont vu le jour et sont en cours de réalisation ou planifiées.