# Considérations sur les thesaurus et mots-clés

## Contexte et objectifs

Lors de la réunion du GT Standard Paysage du 27 Janvier 2023, il a été proposé d’utiliser un thesaurus commun pour exprimer les valeurs possibles des mots-clés portés par les unités paysagères. Plus exactement, suite à cette réunion, il est envisagé que la classe d’objets « Unité Paysagère » ait 2 attributs :

* Mots-clés communs (issus du thesaurus)
* Autre mots-clés

Les mots clés issus du thesaurus pourront servir à des requêtes communes alors que les autres mots-clés ne seront utiles qu’en consultation.

L’objectif de ce document est d’étudier la pertinence des thesaurus existant pour le standard données du paysage. Une liste des vocabulaires du paysage a été fournie par une des participantes du GT Standard et sert de base à cette étude. Cette liste est disponible sur le GiTub du GT : xxx

Cette liste est très riche mais comprend une grande variété de documents (guides, rapports, etc). Seuls les documents correspondant effectivement à glossaires et disponibles sur Internet ont été pris en compte. L’étude a été complétée par quelques recherches sur Internet.

Le chapitre 2 passe en revue les thesaurus potentiels et évalue leur adéquation aux besoins du standard sur les données paysages. Un thesaurus est proposé.

Le chapitre 3 propose un exercice d’appariement entre une sélection de mots-clés faite pour un atlas existant et le thesaurus proposé.

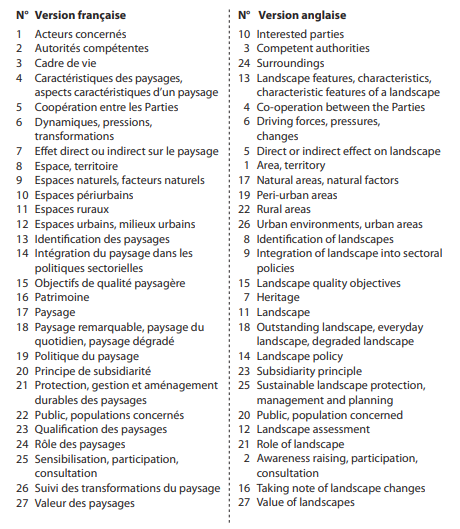
Le chapitre 4 propose quelques enseignements tirés de cet exercice et les conclusions qu’on peut en tirer pour la modélisation et le choix des mots-clés.

NOTE : Les participants au GT avaient aussi évoqué des thesaurus en lien avec la photographie ou le patrimoine mais je n’ai rien trouvé à ce sujet sur Internet. Comme l’étude qui suit a permis de trouver un thesaurus dédié au paysage et convenant globalement aux besoins du GT Standard, cela ne semble pas nécessaire d’aller chercher des thesaurus sur des sujets certes en lien avec le paysage mais pas dédiés à ce thème.

## Description et analyse des sources

#### 2.1 Glossaire du Système d’information de la Convention du Conseil de l’Europe sur le paysage

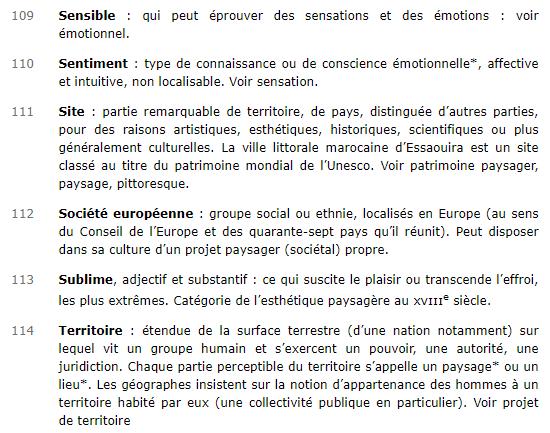
Le glossaire a été préparé par un Groupe de travail du Conseil de l’Europe. En identifiant aussi clairement que possible les principes généraux, les stratégies et les orientations, ainsi que les mesures particulières adoptées aux niveaux national, régional et local en faveur du paysage, le système d’information facilite les échanges d’expériences et d’informations entre les autorités publiques, la société civile et les organisations privées. Son glossaire se focalise sur des termes clés mentionnés dans la convention.

****

C’est un glossaire très limité, avec de peu d’entrées (les 27 chapitres ci-dessus), chacun étant décrit en détail. Il n’est pas adapté comme source de mots-clés pour le standard Paysage.

#### 2.2 Petit lexique de géomédiation paysagiste - Varia N°3 - (24/12/2009) par Pierre Donadieu

Ce lexique peut se lire de A à Z comme un hypertexte en utilisant les astérisques qui renvoient à de nouvelles définitions. Les dates sont celles de l'apparition de certaines notions. Se rapportant à des textes d'enseignement dont il explique quelques termes, ce lexique n'est pas exhaustif.

****

# Extrait du Petit lexique de géomédiation paysagiste

Ce glossaire contient 121 termes surtout reliés à la « philosophie » ou à la politique du paysage. Il ne contient quasiment rien sur les structures ou les éléments de paysage. Peu adapté aux besoins du standard Paysages.

#### 2.3. Dictionnaire du paysage - Michel Godron et Hubert Joly - Conseil international de la langue française (CILF) puf - Paris – 2008

Ce dictionnaire est né d’un besoin ressenti en 1994 lors du projet de restauration du parc paysager à la française, du domaine de la Bucherie dans le Vexin français. Les deux auteurs ont voulu rassembler les mots qui concernent le paysage pour en cerner le sens. Leurs propositions ne sont nullement gravées dans le marbre parce qu’ils considéraient que le vocabulaire lié aux paysages allait continuer à évoluer.

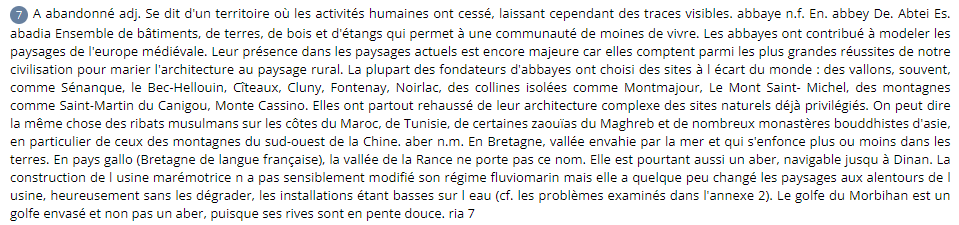
URL : <https://docplayer.fr/23453937-Dictionnaire-du-paysage.html>

Ce dictionnaire est disponible en ligne sur Internet mais les conditions d’accès ne sont pas très ouvertes :

* Inscription soi-disant gratuite sur doc player mais il faut donner ses coordonnées bancaires
* Cette inscription donne droit à 5 j gratuits
* Puis ça déclenche un abonnement automatique à 49 € /mois.

Le texte est disponible sur Internet sans téléchargement.

Néanmoins, il apparaît dans un format qui ne facilite pas une lecture facile car les transitions entre mots n’apparaissent pas clairement et sont noyées dans la masse du texte.



Par exemple, à la lettre B, on trouve environ 80 termes. Ceux-ci désignent :

* Surtout des structures ou éléments de paysage
  + Couverture géographique
    - Génériques , présentes en France : ballon, belvédère, bois, bosquet , bocage, beffroi…
    - Exotiques : banquise, bush, bled, bir ….
  + Granularité diverse des objets
    - Bouclier, bassin versant, banlieue, baie …
    - Bâtiment, brisant, bosquet ….
    - Bain d’oiseau, borie, bonzai
  + Granularité diverse des termes
    - Des termes génériques : bocage, brulis, belvédère
    - Des termes très spécifiques : bois marmenteau, bétulaie , …
* Des mots évoquant l’ambiance d’un paysage :
  + Végétaux : blé, buis, bruyère
  + Matières : basalte, bauxite, brique
  + Climat : brume, brouillard, bise
  + Autres ; brutalisme (style d’architecture), bruit des paysages
* Des concepts plus généraux
  + Subjectif : beauté des paysages
  + Scientifiques : biodiversité, biotope, biomasse ….

Pour chaque terme, le dictionnaire indique le terme en anglais, allemand et espagnol, il donne une définition et il peut fournir des commentaires, des citations. Ces définitions ne paraissent pas toujours assez objectives.

Exemple 1 : *baie n.f. En. bay De. Bucht Es. bahia Rentrant plus ou moins marqué dans le tracé d'un littoral, généralement de forme arrondie, offrant des vues variées sur les paysages mixtes de terre et de mer ou de lac. "La très chaste baie Lovée dans la côte, S'arrondit en cœur Pour s'ourler sans faute." Jean Quintains. La baie de Rio de Janeiro est l'une des plus fameuses.*

Exemple 2 :  *banlieue n.f. En. suburbs De. Vororte, Stadtrandzone, Aussenbezirke Es. extrarradio, alrededores It. periferia Ensemble de communes urbaines ou en cours d'urbanisation, présentant une structure propre assez faible parce qu'elles sont dépendantes de la structure de la ville autour de laquelle elles se sont constituées ou qui a fini par les absorber. Souvent dépourvues des éléments structurants qui sont l’apanage de la ville dont elles dépendent, elles ne sont trop souvent que la juxtaposition de quartiers d’habitations disparates et d’ensembles commerciaux et industriels disposés en vrac dans le paysage. Les banlieues n’ont ni centre ni périphérie sauf lorsqu’elles noient un village ancien qu’elles ont englobé. Le plus souvent, elles se succèdent de commune en commune sans frontières précises, avec les mêmes zones indifférenciées, les mêmes plantations sans imagination et parfois les mêmes noms de rues. "Les Bacchantes des banlieues sanglotent, et la lune brûle et hurle." Arthur Rimbaud*

Malgré ces quelques défauts (conditions d’accès, qualité des définitions), ce dictionnaire paraît globalement adapté pour le GT Paysages :

* A utiliser tel quel en tant que thesaurus pour mots-clés
* Peut donner une liste de termes candidats pour une liste des éléments –structures de paysage

Proposition n° 1 :

Pour les données géomatiques, utiliser le dictionnaire du paysage de Michel Godron et Hubert Joly en tant que thesaurus pour documenter les mots-clés des unités paysagères.

NOTE : pour permettre une utilisation effective, ce dictionnaire devra être rendu gratuitement et facilement accessible. Idéalement, il faudrait qu’il puisse être disponible à partir du site du CNIG.

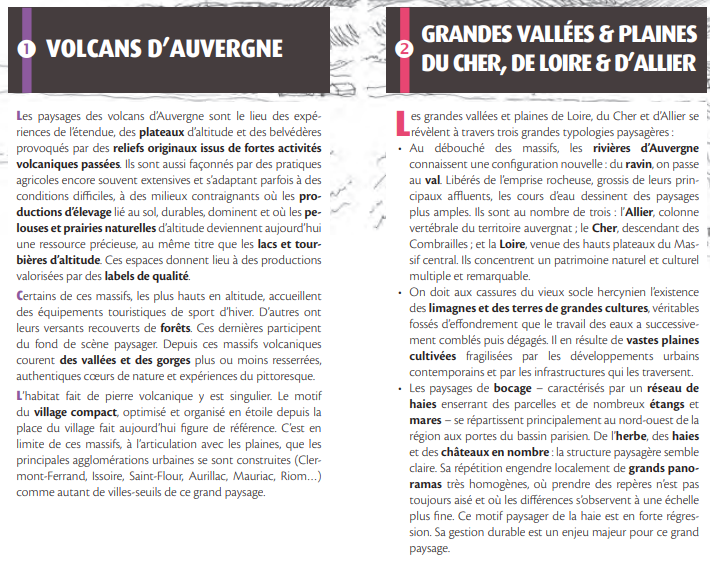
#### 2.4 Atlas d’Ile-et-Vilaine

Cet atlas comprend une trentaine de termes : alignement d’arbres, bassin versant, batholite, belvédère, bourg, bourg sommital, cluse, cheminement-promenade, conurbation, corridor écologique, coteau, échappée visuelle, espace de respiration, frange urbaine, glacis, hameau, village, ville, lande, lisière, lotissement, openfield, point de vue, ragosse, ripisylve, site, slikke-schorre, talus, talweg.

Cette liste est clairement trop limitée pour servir de thesaurus pour les mots-clés. Par contre, elle fournit une sélection des composantes du paysage (éléments, structures).

## Mise en œuvre : exercice de correspondance

L’atlas des grands paysages d’Auvergne-Rhône-Alpes contient, pour chaque **grand paysage**, un texte descriptif avec mise en évidence de certains mots ou expressions. Ces expressions (en gras dans le texte) sont des candidates pour l’attribut mots-clés.



Le tableau suivant indique les correspondances avec les termes du thesaurus proposés pour le GT Paysage. Ces correspondances peuvent être totales, partielles ou inexistantes.

|  |  |  |
| --- | --- | --- |
|  | **Volcans d’Auvergne** | **Grandes vallées & plaines du Cher, de Loire et d’Allier** |
| Correspondance totale avec le thesaurus | Plateau  Pelouse  Lac  Forêt  Vallée  Gorge | Ravin  Val  Bocage  Etang  Mare  Herbe  Haies |
| Correspondance partielle avec le thesaurus | Reliefs originaux issus de fortes activités volcaniques passées (**relief & volcan**)  Production d’**élevage**  **Village** compact  **Prairie** naturelle  **Tourbière** d’altitude | **Rivières** d’Auvergne  **Terre** de grandes **cultures**  Vastes plaines cultivées (**plaine & culture)**  Réseau de **haies**  **Châteaux** en nombre  Grand **panoramas** |
| Pas de correspondance avec le thesaurus | Label de qualité | Allier  Cher  Loire  limagne |

Constatations :

* Sur ces 2 exemples, il y a peu de cas où aucune correspondance n’est possible. Ces cas correspondent à :
  + Des toponymes (Allier, Loire, Cher)
  + Des éléments de paysage spécifiques (limagne)
  + Des termes un peu éloignés du paysage (labels de qualité)
* Les cas de correspondance partielle posent davantage problème :
  + On peut difficilement garder dans une base de données des formulations longues de type « Reliefs originaux issus de fortes activités volcaniques passées » ; abréger en « volcan » ou « volcanisme » semble raisonnable
  + De même, on peut se limiter sans gros inconvénient à « élevage » au lieu de « production d’élevage » voire à « panorama » au lieu de « grand panorama «  ou à « châteaux » au lieu de « châteaux en nombre » (s’il n’y en avait qu’un ou deux, il faudrait donner leur nom)
  + Par contre, « plaine ; culture » n’est pas équivalent à « vaste plaine cultivée » ; on perd aussi de l’information si on utilise « village » au lieu de « village compact » …
  + On a des cas de formulation avec un mot-clé du thesaurus (rivière) et un mot-clé toponymique (Auvergne).

## Conclusions et propositions

L’exercice ci-dessus a bien entendu ces limites : il prend en compte des grands paysages et pas des unités paysagères, il a été effectué sur un échantillonnage très limité (2 entités paysagères). Néanmoins, on peut en tirer les enseignements suivants :

* **Pertinence du thesaurus proposé**

Conclusion n°1 :

A l’exception incontournable des termes toponymiques, il y a très peu de mots-clés nécessaires qui ne soient pas dans le dictionnaire du paysage. Cet exercice confirme que le dictionnaire proposé est suffisamment exhaustif pour les besoins du standard paysage.

* **Finalité et format des mots-clés**

Conclusion n°2:

Cet exercice relance le débat sur la finalité et donc le format des mots-clés. Un mot-clé pris dans le thesaurus commun permet de faire des requêtes communes ; une expression permet de décrire plus justement l’ambiance, les caractéristiques de l’entité paysagère.

NOTE 1 : A mon avis, il ne faut pas privilégier exagérément la finalité de requête commune. Par exemple, la présence d’un phénomène géographique (lac, bois, marais, château …) peut être traitée de 3 façons différentes dans les données paysage :

* + Cette information n’est pas considérée comme une caractéristique du paysage et n’est pas présente du tout dans les données géomatiques
  + Cette information est considérée comme une caractéristique essentielle et est modélisée sous forme d’objets géographiques (classe « éléments de paysage »)
  + Cette information est considérée comme une caractéristique importante et est modélisée sous forme de mots-clés de l’entité paysagère.

Autrement dit, une requête de type «  donne-moi toutes les UP dont l’attribut [mot-clé] contient [lac] » ne garantira absolument pas d’obtenir toutes les UP ayant des lacs sur leur territoire. Pour avoir la garantie d’une réponse exhaustive, il vaut mieux croiser la couche des UP avec une couche de référence (ex : BD TOPO).

NOTE 2 : Les entretiens ont fait apparaître le besoin ou du moins l’intérêt de mots-clés toponymiques. En pratique, ces toponymes sont souvent composés d’un nom commun (trouvable la plupart du temps dans le thesaurus) et d’un nom propre. Ces mots-clés seraient difficiles à classer dans le modèle proposé lors de la réunion de Janvier (un attribut « mot-clé thesaurus » et un attribut « autre mot-clé »).

Exemples : lac des Settons, Mont Saint-Michel, Puy de Sancy, estuaire de la Gironde, château de Chantilly, plateau du Larzac

NOTE 3 : Lors de la réunion de Janvier, le GT avait opté pour de simples mots-clés (pas d’expressions avec plusieurs mots) mais je ne suis pas convaincue de la pertinence de ce choix. Outre les expressions toponymiques, je pense qu’il faudrait aussi autoriser des expressions à base de noms communs et d’adjectifs

* Avantage : l’expression permet de décrire plus justement la caractéristique de l’UP. Par exemple, « vaste plaine cultivée » est plus clair, plus parlant que « plaine ; culture » (les cultures sont-elles forcément dans la plaine ?). Le mot-clé « village » peut s’appliquer à de nombreuses UP et donner une information assez banale alors que « village compact » ou « village perché » indique une véritable caractéristique.
* Risques :
  + On court le risque d’expressions exagérément longues, si on ne met pas une limite sur la longueur des expressions (et cette limite sera sujette à discussion).
  + On peut passer à côté de certains mots-clés (on ne trouve pas « culture » dans « vaste plaine cultivée »)
* **Quantification des mots-clés**

On peut utiliser les résultats de l’exercice précédent pour quantifier le nombre de mots-clés et la longueur des chaînes de caractères.

Scénario 1 : on garde toutes les expressions issues de l’atlas

|  |  |  |
| --- | --- | --- |
|  | **Volcans d’Auvergne** | **Grandes vallées & plaines du Cher, de Loire et d’Allier** |
| Correspondance totale avec le thesaurus | Plateau  Pelouse  Lac  Forêt  Vallée  Gorge | Ravin  Val  Bocage  Etang  Mare  Herbe  Haies |
| Correspondance partielle avec le thesaurus | Reliefs originaux issus de fortes activités volcaniques passées  Production d’**élevage**  **Village** compact  **Prairie** naturelle  **Tourbière** d’altitude | **Rivières** d’Auvergne  **Terre** de grandes **cultures**  Vastes plaines cultivées  Réseau de **haies**  **Châteaux** en nombre  Grands **panoramas** |
| Pas de correspondance avec le thesaurus | Label de qualité | Allier  Cher  Loire  limagne |
| Nombre mots ou expressions clés | 12 | 17 |
| Longueur chaîne de caractères | ≈ 190 | ≈ 190 |

NOTE 1 : Le comptage est fait de façon approximative avec un seul caractère comme séparateur.

Scénario 2 : on fait un peu de ménage en enlevant les redondances et en simplifiant autant que possible, mais en gardant les expressions considérées comme indispensables ou très utiles. Par exemple, concernant la seconde UP :

* « Bocage » peut résumer : bocage, haies, herbes, réseaux de haies
* « vastes plaines cultivées » peut spécifier « terre de grandes cultures »

|  |  |  |
| --- | --- | --- |
|  | **Volcans d’Auvergne** | **Grandes vallées & plaines du Cher, de Loire et d’Allier** |
| Mots-clés thesaurus | Plateau  Pelouse  Lac  Forêt  Vallée  Gorge  Volcan  élevage | Ravin  Val  Bocage  Etang  Mare  Château  Panorama |
| Expressions (avec mots-clés du thesaurus) | **Village** compact  **Prairie** naturelle  **Tourbière** d’altitude | **Rivières** d’Auvergne  Vastes **plaines** cultivées |
| Pas de correspondance avec le thesaurus | Label de qualité ( ?) | Allier  Cher  Loire  limagne |
| Nombre mots ou expressions clés | 12 | 13 |
| Longueur chaîne de caractères | ≈ 130 | ≈ 130 |

NOTE 1 : cet exercice fait aussi apparaître le problème de comment écrire les mots-clés : singulier ou pluriel ? J’ai fait le choix du singulier car c’est ce qui apparaît dans le thesaurus mais « châteaux » serait probablement plus adapté pour résumer « châteaux en nombre ».

NOTE 2 : l’exercice de simplification a été fait de façon expéditive et ne prétend pas être optimum. Par exemple, faut-il garder « label de qualité » sans la mention de l’espace auquel il s’applique ?

Scénario 3 : on ne met que des mots-clés et aucune expression (sauf éventuellement les expressions toponymiques)

|  |  |  |
| --- | --- | --- |
|  | **Volcans d’Auvergne** | **Grandes vallées & plaines du Cher, de Loire et d’Allier** |
| Mots-clés thesaurus | Plateau  Pelouse  Lac  Forêt  Vallée  Gorge  Volcan  Elevage  Village  Prairie  Tourbière | Ravin  Val  Bocage  Etang  Mare  Château  Panorama  Plaine  Culture |
| Autres mots-clés |  | **Rivières** d’Auvergne ( ?)  Allier  Cher  Loire  limagne |
| Nombre mots ou expressions clés | 11 | 14 |
| Longueur chaîne de caractères | ≈ 80 | ≈ 125 |

Le scénario 1 ne paraît pas raisonnable pour les données géomatiques. Lors de la réunion du GT Standard, l’agence CAUDEX a signalé que les mots-clés devaient provenir du texte descriptif de l’entité paysagère, que ce texte devait être la source initiale (et pas le thesaurus), ce qui paraît une bonne méthodologie. Néanmoins, il faudra un travail de « nettoyage » et de sélection des mots-clés à mettre dans les données géomatiques.

Conclusion n°3

Les mots-clés sont à extraire du texte descriptif de l’entité paysagère. Il faudra prévoir un travail de simplification, de sélection et d’appariement avec les termes du thesaurus. Cet appariement est plus ou moins strict selon le choix du scénario (seulement des mots ou aussi quelques expressions).

La comparaison entre les scénarios 2 et 3 permet d’illustrer sur un cas concret la question de la finalité et du format des mots-clés (conclusion n°2). Il faut noter que le travail de « nettoyage » des expressions-clés du texte descriptif est plus radical mais aussi plus automatique, plus facile dans le scénario 3.

* **Modélisation des mots-clés**

On peut envisager 3 types de modélisation :

* 1. Un attribut « mot-clé générique » et un attribut « mot-clé toponymique » (proposition initiale) en recommandant l’utilisation du thesaurus pour les mots-clés génériques
  2. Un attribut « mot-clé thesaurus » et un attribut « autre mot-clé » (proposition lors de la réunion du GT Paysage)
  3. Un attribut unique pour tous les mots-clés en recommandant l’utilisation du thesaurus pour les mots-clés autres que la toponymie

La première modélisation est adaptée au scénario 2 (on peut garder quelques expressions) alors que la seconde modélisation est adaptée au scénario 3 (que des mots, extraits autant que possible du thesaurus). L’attribut unique peut s’adapter aux 2 scénarios.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Attribut** | **Volcans d’Auvergne** | **Grandes vallées & plaines du Cher, de Loire et d’Allier** |
|  | | | |
| Modèle 1  (scénario 2) | Mot-clé générique | *élevage, forêt, gorge, lac, pelouse, plateau,* ***prairie naturelle, tourbière d’altitud****e, vallée,* ***village compact****, volcan* | *bocage, château, étang,* ***limagne****, mare, panorama, ravin, val,* ***vastes plaines cultivées*** |
| Mot-clé toponymique |  | *Allier, Cher, Loire, rivières d’Auvergne* |
|  | | | |
| Modèle 1  (scénario 3) | Mot-clé thesaurus | *élevage, forêt, gorge, lac, pelouse, plateau, prairie, tourbière, vallée, village, volcan* | *bocage, château,* ***culture****, étang, mare, panorama, ravin,* ***plaine****, val* |
| Autre mot-clé |  | *Allier, Cher,* ***limagne,*** *Loire, rivières d’Auvergne* |
|  | | | |
| Modèle 3  (scénario 2) | Mot-clé | *élevage, forêt, gorge, lac, pelouse, plateau, prairie naturelle, tourbière d’altitude, vallée, village compact, volcan* | *bocage, château, étang, limagne, mare, panorama, ravin, val, vastes plaines cultivées* |
|  | | | |
| Modèle 3  (scénario 3) | Mot-clé | *élevage, forêt, gorge, lac, pelouse, plateau, prairie, tourbière, vallée, village, volcan* | *Allier, bocage, château, Cher, culture, étang, limagne, Loire, mare, rivières d’Auvergne, panorama, ravin, plaine, val* |

Constatations (sur cet exercice) :

* La quantité d’information est raisonnable : la liste des mots-clés pourrait apparaître de façon lisible dans une info-bulle de chaque UP
* Cette lisibilité est améliorée quand on utilise 2 attributs. La séparation entre mots-clés génériques et mots clés toponymiques paraît plus naturelle, plus immédiatement compréhensible que la séparation selon l’appartenance ou non au thesaurus. La dernière case du tableau (un seul attribut mêlant noms génériques et toponymie ) est peu agréable à consulter .
* Pour cet exercice, j’ai fait le choix de ranger les mots-clés selon un ordre alphabétique mais d’autres choix sont possibles
  + - Ne donner aucune consigne
    - Regroupements par thématiques : ex : mots-clés relatifs au relief, mots-clés relatifs à l’occupation du sol, …
    - Regroupements selon les expressions : ex : « plaine, culture »  pour « vastes plaines cultivées » ; néanmoins, cette correspondance n’a rien d’évident ; le lecteur ne saura pas forcément que « culture » s’applique à « plaine » et non pas au terme suivant ou à toute l’UP …

Proposition n° 2 :

Privilégier l’usage de consultation : les mots-clés sont destinés à donner l’ambiance de l’entité paysagère, pouvoir faire des requêtes communes est un plus mais ce n’est pas l’objectif principal

* Avoir 2 attributs, un pour les mots-clés génériques, l’autre pour les mots-clés toponymiques
* Autoriser quelques expressions
* Recommander l’utilisation du thesaurus

NOTE : A discuter (retour sur ma proposition initiale, remise en cause de qui avait été proposé lors de la réunion du GT Standard).

Proposition n°3

Pour une mise en œuvre efficace du standard, il faudrait donner des règles et/ou des exemples de sélection et de syntaxe des mots-clés.

NOTE : Certains points ont été évoqués lors de la réunion du GT, d’autres ont été mentionnés dans ce document :

* consignes pour les expressions : longueur raisonnable, compréhension immédiate
* éventuelle redondance avec les éléments de paysage
* singulier ou pluriel
* ordre des termes dans la liste des mots-clés
* nombre de caractères maximum pour les attributs « mots-clés »