RITUEL

SESSION DU 13/12/18

INSTITUT CARNOT brgm

Géosciences pour une Terre durable
brgm

RITUEL

SESSION DU 13/12/18

INSTITUT CARNOT brgm

Géosciences pour une Terre durable
brgm

Différents sujets de travail

- Support utilisateur (Teneur des messages)
- Text Mining pour l'environnement
- Indexation automatique
- Extraction de connaissance
  - Entités métier
  - Entités nommés & géolocalisation du document
- Ergonomie de solutions semi-automatique : format PDF
- Autres ;)

Données générées par l'activité de support 1/3

Pérennisation du travail réalisé dans le cadre du stage de S. Lefevre

- Externalisation des fichiers de paramétrage des traitements pour :
  - Faciliter leur mise à jour

- Développement d'une interface web pour :
  - Faciliter l'exécution des traitements
  - Améliorer la restitution graphique des résultats

Données générées par l'activité de support 2/3

Poursuite de l'analyse du corpus

- Analyse de sentiments :
  - Objectif initial : tirer du corpus des demandes un indicateur de satisfaction globale des utilisateurs
    - ✓ 37% d'utilisateurs satisfaits : et alors ?
  - Nouvel objectif : identifier les sujets qui impactent le plus la satisfaction utilisateur
    - ✓ Nécessite de calculer la notion de tonalité non plus à l'échelle du message mais à l'échelle de chaque sujet abordé au sein des demandes
    - ✓ Nécessite d'identifier plus précisément à l'intérieur des messages quels sujets sont abordés positivement et négativement
      - Réclame une classification approfondie des messages
      - Meilleure prise en compte des modes d'expression des locuteurs (diversité d'expressions pour désigner une même fonctionnalité)

Données générées par l’activité de support 2/3

Poursuite de l’analyse du corpus

- Bonjour L'impossibilité d'obtenir les données georisques concerne tout les géorisques. Après vérification auprès de notre service assistance GENAPI, il s'avère qu'il ne s'agit pas d'un problème interne, mais bien d'un souci entre le site notaire et le site Georisques. **Nous ne pouvons plus obtenir avoir les données** pour les dossiers de vente ! problème à régler urgemment.

- Message: Madame, Monsieur, Je me retrouve dans l'impossibilité d'imprimer dans plus de 95% des cas **les pièces choisies**. Que dois-je faire ? Merci.
- Bonjour, **Je n'arrive pas à accéder à votre site. J'arrive péniblement à aller jusqu'à l'adresse du bien, sans jamais réussir à la valider.** Merci de m'indiquer comment procéder.

- bonjour, Je suis notaire et **j'ai besoin régulièrement des infos géorisques**. Cependant je dois attendre des heures pour avoir l'information en impression. C'est impossible à gérer... Il faut régler ce problème au plus vite. C'est un outil de travail indispensable pour moi. Merci

- Bonjour, Depuis plusieurs jours déjà, **nous n'arrivons plus à télécharger et obtenir l'imprimé Géorisques**, Basias et autres que nous consultons bien évidemment à plusieurs reprises par jour.

Données générées par l'activité de support 2/3

Poursuite de l'analyse du corpus

- Evaluation ergonomique des interfaces utilisateurs
  - Affiner le processus de classification des demandes relevant du support applicatif pour :
    - ✓ identifier les fonctionnalités des outils sur lesquels portent les demandes
    - ✓ distinguer les différents niveaux de besoins d'assistance

- Détection de verbatims « emblématiques »
    - ✓ Messages présentant une félicitation, un problème grave ou un autre fait marquant qui les font se différencier des autres verbatims au sein du corpus

Données générées par l'activité de support 3/3

Conception de nouveaux services à l'utilisateur

- Recommandation de contenus
  - Objectif : produire des recommandations de contenu (FAQ, tutos, ressources documentaires) en fonction des demandes utilisateurs

- FAQ dynamique / Agent conversationnel
  - Objectif : selfhelp : offrir un accès à des réponses / ressources par l'interrogation en langage naturel d'une base de connaissance

Documenter les sources historiques de pollution

- BRGM = référence nationale pour la conception de diagnostics, la mise au point ou l'évaluation de procédés de surveillance et de réhabilitation d'environnements pollués par d'anciennes activités économiques (friches industrielles, anciens sites miniers…)

- Dans ce domaine, le BRGM est souvent amené à réaliser des recherches pour documenter les **sources historiques de pollution**

- Plusieurs cas récents ont montré l'intérêt d'exploiter des sources complémentaires à la base de l'inventaire des anciens sites industriels (Basias) et notamment des ressources en presse ancienne numérisées (disponibles en ligne sur Gallica ou autres portails de bibliothèques numériques)

Text mining pour l'histoire de l'environnement

Documenter les sources historiques de pollution :

Le mercure à Paris 1/3

- Dans le cadre du 2ème plan national santé environnement, le BRGM intervient dans la validation des **diagnostics des sols des établissements accueillant des enfants et adolescents,** situés sur des sites potentiellement pollués du fait d'anciennes activités industrielles (ETS)

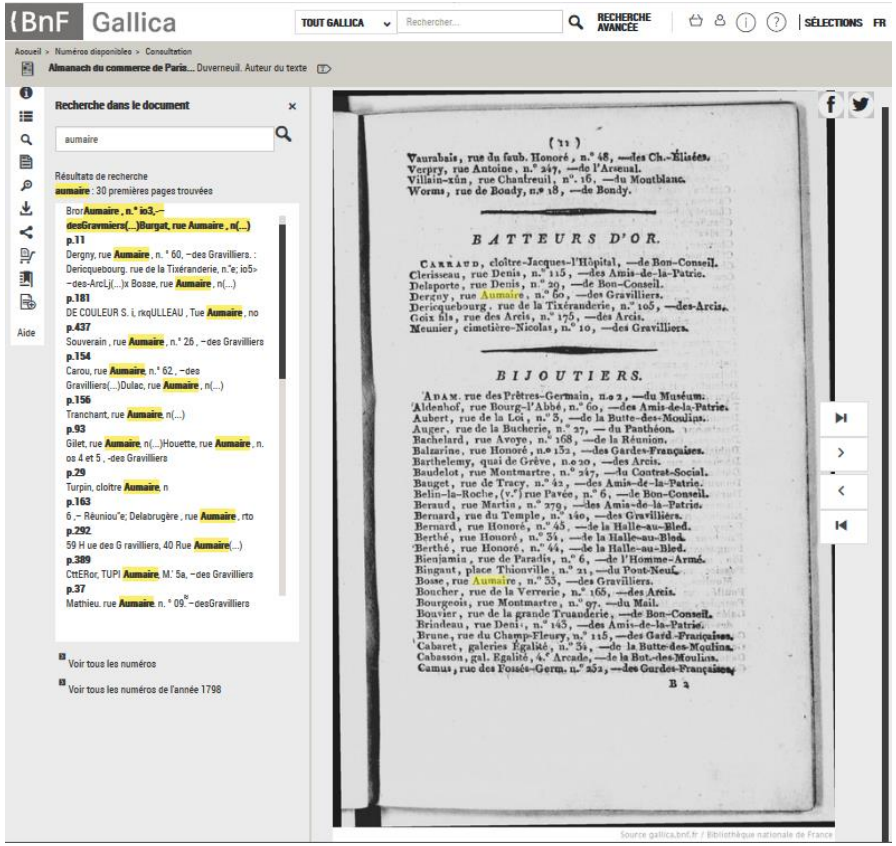
- Recherches en cours pour expliquer la présence de mercure dans l'air du sol de 4 établissement à Paris
  - présence restée inexpliquée malgré l'étude historique menée dans le cadre du diagnostic des sols
  - Objectif : identifier si des activités potentiellement émettrices de mercure (hors activités inventoriées dans BASIAS) ont eu lieu à ces 4 adresses et dans un périmètre de 30 mètres autour de ces adresses

Text mining pour l'histoire de l'environnement

Documenter les sources historiques de pollution :

Le mercure à Paris 2/3

- Source : Almanach du commerce de Paris :
  - 63 numéros disponibles sur Gallica couvrant la période de 1798 à 1907



- ✓ Echantillonnage sur 12 n°
- ✓ Recherche manuelle des adresses via le module de recherche plein texte
- ✓ Report manuel des informations dans un fichier Excel

Documenter les sources historiques de pollution :

Le mercure à Paris 3/3

- Potentiel d'information sur les activités artisanales et industrielles à Paris très important (localisation et datation des activités) au-delà du seul cas d'étude relatif au mercure

- Volonté d'explorer les techniques de text mining pour exploiter ce corpus numérique

- Verrous :
  - Exploitation possible des données Gallica

Documenter les sources historiques de pollution :

Les perchlorates dans les eaux souterraines 1/3

- Dans le cadre de l'Objectif n°2 du plan micropolluants 2016-2021 pour préserver la qualité de l'eau et de la biodiversité : besoin d'améliorer la connaissance et de prédire la présence de perchlorate dans les eaux souterraines.
- Recherches en cours pour documenter les pressions polluantes en perchlorate liées à l'usage de nitrate de soude chilien pour l'amendement des sols agricoles durant la seconde moitié du XIX^{ème} et la première moitié du XX^{ème}
  - Objectif : quantifier et spatialiser l'usage du nitrate de soude chilien sur le territoire national
  - Hypothèse : la presse locale ancienne pourrait contribuer à répondre à la question de la localisation des usages des nitrates chiliens sur le territoire français

Text mining pour l'histoire de l'environnement

Documenter les sources historiques de pollution :

Les perchlorates dans les eaux souterraines 2/3

- Presse locale ancienne
  - 4 136 journaux anciens sur le thème de l'agriculture (de 1760 à 1944) dont 430 disponibles numériquement recensés par le site « Presse locale ancienne »

AVIS AU COMMERCE ET A L'AGRICULTURE.

H. et J. Deconinck à Dunkerque et à Arras, ont présentement à vendre 41 variétés de BLES DE SEMENCE anglais et français ; agents de *frédéric f^rs. halett* (blés généalogiques). Achats faits directement sur les lieux de production. Même maison : nitrate de soude, (importation directe) et tous autres engrais chimiques sur dosage garanti.

27.254

MAISON SPÉCIALE pour produits destinés à L'AGRICULTURE

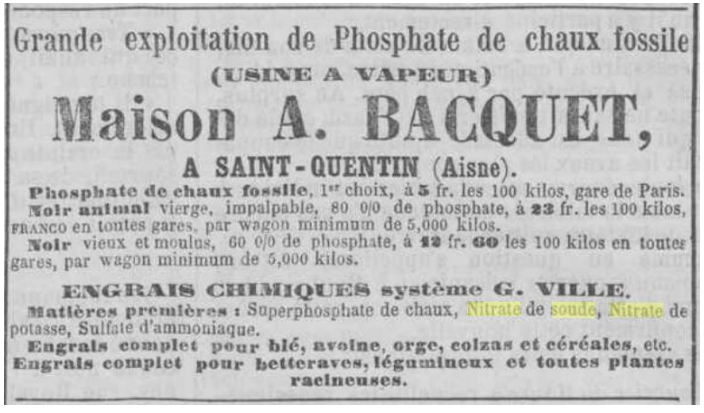
H. & J. DECONINCK à Arras et à Dunkerque ont présentement à vendre 43 variétés de BLÉS de SEMENCE anglais et français. — Achats faits directement sur les lieux de production. Agents de FRÉDÉRICK F. HALLETT (Blés généalogiques). Orges et Avoines de semence, etc.

Même Maison : Tous Engrais chimiques, dosages garantis sur analyse.

NITRATE DE SOUDE des mers du Sud, pour engrais (importation directe).

TOURTEAUX de toutes espèces et provenances pour nourriture et pour engrais.

40.364



Grande exploitation de Phosphate de chaux fossile

(USINE A VAPEUR)

Maison A. BACQUET,

A SAINT-QUENTIN (Aisne).

Phosphate de chaux fossile, 1er choix, à 5 fr. les 100 kilos, gare de Paris.

Noir animal vierge, impalpable, 80 0/0 de phosphate, à 23 fr. les 100 kilos, FRANCO en toutes gares, par wagon minimum de 5,000 kilos.

Noir vieux et moulus, 60 0/0 de phosphate, à 12 fr. 60 les 100 kilos en toutes gares, par wagon minimum de 5,000 kilos.

ENGRAIS CHIMIQUES système G. VILLE.

Matières premières : Superphosphate de chaux, Nitrate de soude, Nitrate de potasse, Sulfate d'ammoniaque.

Engrais complet pour blé, avoine, orge, colzas et céréales, etc.

Engrais complet pour betteraves, légumineux et toutes plantes racineuses.

COMPAGNIE DES ENGRAIS CONTROLÉS

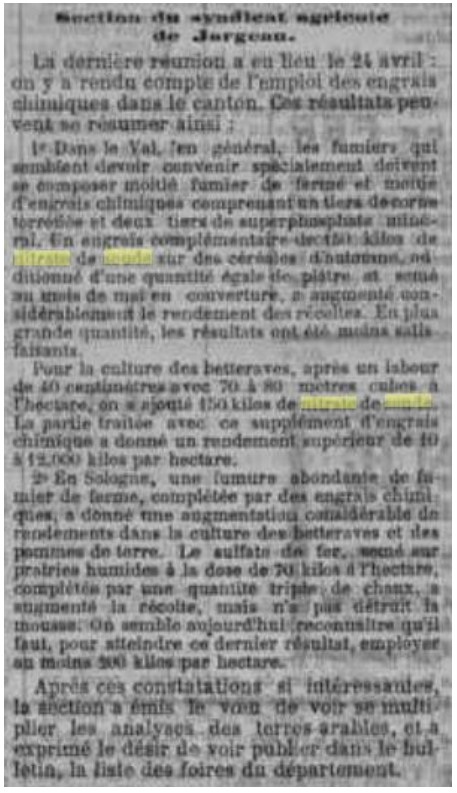
BERTHIER, SEURETTE & Cie

11, rue Boucry, Paris.

PHOSPHATE FOSSILE, SUPERPHOSPHATE, GUANO DU PÉROU

NITRATE DE SOUDE, ENGRAIS ÉQUILIBRE ET COMPLET POUR CÉRÉALES, ETC.

Envoi franco par la poste et sur demande de circulaires contenant les garanties de dosage et tous renseignements. 31.120



Section du syndicat agricole de Jargeau.

La dernière réunion a eu lieu le 24 avril : on y a rendu compte de l'emploi des engrais chimiques dans le canton. Ces résultats peuvent se résumer ainsi :

1° Dans le Val, en général, les fumiers qui semblent devoir convenir spécialement doivent se composer moitié fumier de ferme et moitié d'engrais chimiques comprenant un tiers de corne torréfiée et deux tiers de superphosphate minéral. Un engrais complémentaire de 150 kilos de nitrate de soude sur des céréales d'automne, additionné d'une quantité égale de plâtre et semé au mois de mai en couverture, a augmenté considérablement le rendement des récoltes. En plus grande quantité, les résultats ont été moins satisfaisants.

Pour la culture des betteraves, après un labour de 40 centimètres avec 70 à 80 mètres cubes à l'hectare, on a ajouté 150 kilos de nitrate de soude. La partie traitée avec ce supplément d'engrais chimique a donné un rendement supérieur de 10 à 12,000 kilos par hectare.

2° En Sologne, une fumure abondante de fumier de ferme, complétée par des engrais chimiques, a donné une augmentation considérable de rendements dans la culture des betteraves et des pommes de terre. Le sulfate de fer, semé sur prairies humides à la dose de 70 kilos à l'hectare, complétés par une quantité triple de chaux, a augmenté la récolte, mais n'a pas détruit la mousse. On semble aujourd'hui reconnaître qu'il faut, pour atteindre ce dernier résultat, employer au moins 300 kilos par hectare.

Après ces constatations si intéressantes, la section a émis le vœu de voir se multiplier les analyses des terres arables, et a exprimé le désir de voir publier dans le bulletin, la liste des foires du département.

Text mining pour l'histoire de l'environnement

Documenter les sources historiques de pollution :

Les perchlorates dans les eaux souterraines 3/3

- Techniques de text mining pour :

  - Détecter et comptabiliser les occurrences du sujet « nitrate de soude » sur la période d'usage connue des nitrates de soude chilien
  - Les associer à un lieu géographique grâce à la détection et à la reconnaissance d'entités nommées
  - Pour avoir un aperçu via le prisme de la presse ancienne des usages historiques du nitrate chilien

Indexation automatique

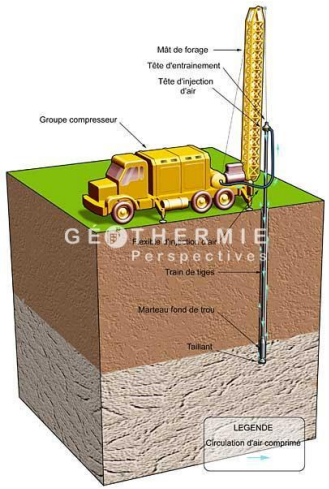
Introduction : Centre de documentation du BRGM

Rapports produits par le BRGM (Environ 30k rapports)

- Démarche de dépôt
  - Indexation de la part des documentaliste (titre, auteur, résumé, pages, …)
  - Mot clés de la part des auteurs avec thésaurus dans l'outil de gestion PMB

D'autres documents (rapports extérieurs)

- Récupération opportune / motivée par un projet
  - Exploitation du contenu technico-scientifique (run once)
  - Pérenniser le contenu scientifique du document (run several times)
  - Cataloguer le document (PDF)
    - Automatique
    - Identification titre, auteur, commanditaire, caractéristique
    - Mot-clés : Thésaurus et Libre (RAKE)
    - Géolocalisation – toponyme et géométrique



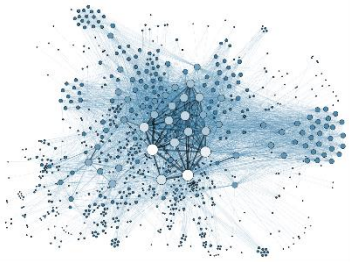
Extraction de connaissance

Comment valoriser la connaissance du corpus BRGM

Retrouver nos objets métiers : ce qui a déjà été fait :

→ Démarche automatique

- → Monde d'identifiants
- → **Expressions régulières** + contrôle dans le SI
- → Bancarisation → Linked Data et Web sémantique
  - → Ex : BSS (03676X0054/PUITS ou BSS001AQPK), Basias (CEN1800312)



→ A faire

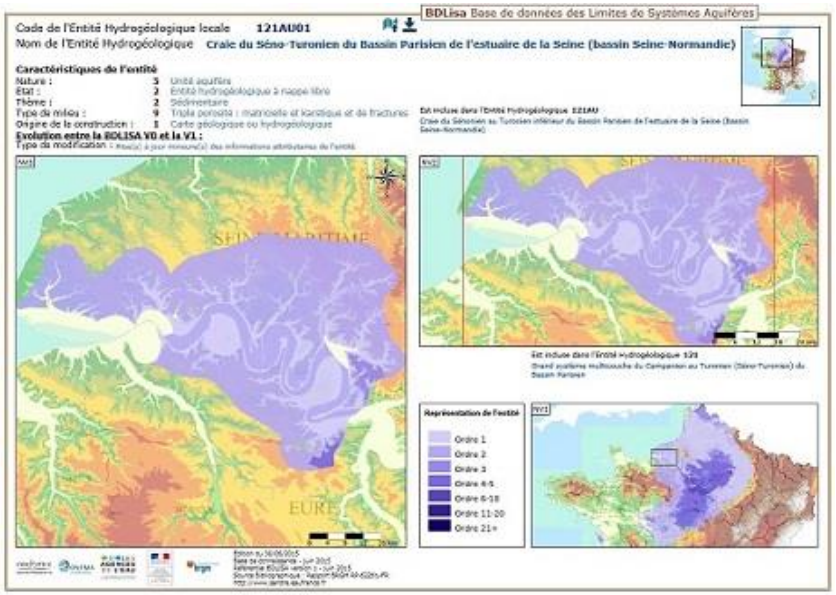
- → Lien entre les rapports. Comment nos rapports s'appuient les uns sur les autres.
- → La BDLISA : Limite des systèmes aquifères

Extraction de connaissance

La BDLISA : Subdivision du sous-sol en réservoirs - aquifères / ou non réservoir

Vue aérienne →

Niveau 0 - Formations superficielles		
Niveau 1 - National	**Niveau 2 - Régional**	**Niveau 3 - Locale**
080AA72-- Formations des Limons des plateaux (code géol.: LP) dans l'extension de l'entité régionale : 107AK		
107-- Grand système multicouche de l'Oligo-Miocène du Bassin Parisien	**107AK**-- Calcaires de Brie du Rupélien (Oligocène inf.) du Bassin Parisien (bassin Seine-Normandie et Loire-Bretagne)	**107AK01**-- Calcaires de Brie du Rupélien (Oligocène inf.) du Bassin Parisien (bassin Seine-Normandie et Loire-Bretagne)
110-- Grand domaine hydrogéologique de l'Oligocène inf. à l'Eocène sup. (Sannoisien au Ludien) du Bassin Parisien	**110AA**-- Marnes vertes et supra-gypseuses du Rupélien (Oligocène inf.) du Bassin Parisien (bassin Seine-Normandie majoritairement et bassin Loire-Bretagne)	**110AA01**-- Marnes vertes et supra-gypseuses du Rupélien (Oligocène inf.) du Bassin Parisien (bassin Seine-Normandie majoritairement et bassin Loire-Bretagne)
113-- Grand système multicouche de l'Eocène du Bassin Parisien	**113BA**-- Faciès marneux du Ludien moyen de l'Eocène sup. du Bassin Parisien (bassin Seine-Normandie)	**113BA01**-- Faciès de transition (marnes et calcaires) du Ludien de l'Eocène sup. du Bassin Parisien
	113AI-- Marnes Infra-gypseuses de l'Eocène du Bassin Parisien	**113AI01**-- Marnes Infra-gypseuses de l'Eocène du Bassin Parisien
	113AK-- Sables, calcaires et grès du Bartonien (Eocène) du Bassin Parisien	**113AK01**-- Sables de Monceau, de Marines, de Cresnes du Marinésien supérieur (Bartonien inf.) du Bassin Parisien
		113AK03-- Calcaires de Saint-Ouen du Bartonien inf. du Bassin Parisien
		113AK05-- Sables du Marinésien (sables de Mortefontaine, Calcaire de Ducy, Sables d'Ezanville) et de l'Auversien (Sables de BeauChamps, d'Auvers) du Bassin Parisien



← Vue verticale

107AK01 : Calcaires de Brie du Rupélien (Oligocène inf.) du Bassin Parisien (bassin Seine-Normandie et Loire-Bretagne)

Niveau 3 - Locale

xtension de l'entité régionale : 107AK

107AK01-- Calcaires de Brie du Rupélien (Oligocène inf.) du Bassin Parisien (bassin Seine-Normandie et Loire-Bretagne)

110AA01-- Marnes vertes et supra-gypseuses du Rupélien (Oligocène inf.) du Bassin Parisien (bassin Seine-Normandie majoritairement et bassin Loire-Bretagne)

113BA01-- Faciès de transition (marnes et calcaires) du Ludien de l'Eocène sup. du Bassin Parisien

113AI01-- Marnes infra-gypseuses de l'Eocène du Bassin Parisien

113AK01-- Sables de Monceau, de Marines, de Cresnes du Marinésien supérieur (Bartonien inf.) du Bassin Parisien

113AK03-- Calcaires de Saint-Ouen du Bartonien inf. du Bassin Parisien

113AK05-- Sables du Marinésien (sables de Mortefontaine, Calcaire de Ducy, Sables d'Ezanville) et de l'Auversien (Sables de BeauChamps, d'Auvers) du Bassin Parisien

113AO01-- Marnes et caillasses du Lutétien sup. du Bassin Parisien, contenant localement du gypse

113AQ21-- Calcaires grossiers du Lutétien du sud du Bassin Parisien

113AV03-- Sables de Cuise de l'Yprésien sup. du Bassin Parisien (bassin Seine-Normandie)

117AC03-- Argiles, sables et lignites de l'Yprésien inf. du Bassin Parisien (bassin Seine-Normandie et sud du bassin Artois-Picardie)

117AC05-- Argiles plastiques de l'Yprésien inf. du Bassin Parisien (bassin Seine-Normandie)

121AP03-- Craie du Sénonien au Turonien inférieur, partie sous recouvrement au centre du Bassin Parisien (bassin Seine-Normandie et bassin Loire-Bretagne)

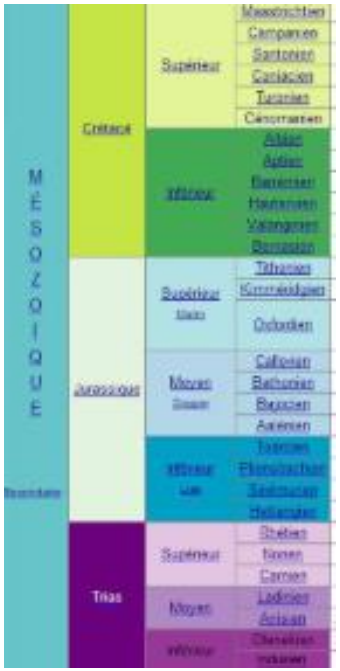
123BP01-- Marnes et craie

Extraction de connaissance

Ce qui a été envisagé :

Scénarios progressifs en 3 étapes …

- 1 - Index SolR des termes BD LISA + algo de fenêtre glissante sur le texte avec indicateurs par rapport à l'index
- 1 bis – considération de la localisation en baissant le niveau d'indicateur
- 1 ter – considération de l'échelle des temps géologiques pour la parenté des périodes géologiques
- 2 - Espace vectoriel de mots + fenêtre glissante sur texte
- 3 - Machine learning



Fait : 1

Géosciences pour une Terre durable
brgm

Extraction de connaissance

Extraire les descriptions lithologique complètes

Alternance de petits bancs silteux micacés,de petits bancs arénacés à lumachelles et des niveaux schisteux à silteux de couleur vert à gris clair avec altération des fossiles de couleur rouille + des petits bancs de grès de 5 à 6 cm d'épaisseur

Mais il y a des formes plus narratives

http://infoterre.brgm.fr/rapports//73-RME-008-FE.pdf

Les filons sont formés d'un quartz blanc opaque, imprégné par places de fer oligiste, qui affecte par endroits un aspect brechoïde très prononcé; les fragments de quartz sont alors ressoudés tantôt par de l'agate zonée, tantôt par du jaspe rouge ou violacé.

Extraction de connaissance

Extraire les descriptions lithologique complètes

alternance
- banc
  - petit
  - silt
    - mica
- banc
  - petit
  - lumachelles
- niveau
  - Variation
    - schiste
    - silt
  - couleur
    - Variation
      - vert
      - gris
        - clair
    - altération
      - fossiles
        - couleur
          - rouille
- banc
  - petit
  - grès
    - Variation
      - 5
      - 6
      - d'épaisseur

Calcaire à oolithes ferrugineuses et à petits nodules d'améthyste

Calcaire	→ Rang 1
oolithe	
fer	→ Rang 3
nodule	
petit	
améthyste	

Extraction de connaissance

Extraire les descriptions lithologique complètes

Et pour les formes narratives ….

Les filons sont formés d'un quartz blanc opaque, imprégné par places de fer oligiste, qui affecte par endroits un aspect brechoïde très prononcé; les fragments de quartz sont alors ressoudés tantôt par de l'agate zonée, tantôt par du jaspe rouge ou violacé.

Filon
  quartz
    blanc
    opaque
     place
       oxyde de fer – hématite
     aspect
        brèche
           ciment
             agate
             jaspe
                rouge
                violacé

→ Apprendre à la machine les références en se basant sur les structures grammaticales

→ *SEM*

→ *Stanford NLP*

Ergonomie avec le format PDF

- → Comment faire des chaines de traitement semi-automatique
  - → Restituer dans le look de l'affichage PDF les résultats de traitement.

Autres

- Anonymisation (gestion vers la diffusion)
- Fouille de Tweet – Suricate
- Extraction de mesures dans des tableaux
- Retranscription audio
  - Minutes des réunions avec de nombreux participants
  - Texte
  - Identification des individus

Cas d'utilisation 3

Extraction des valeurs analytiques dans les rapports

- Corpus des rapports ETS → « Format quasi standardisé »
- Extraction des valeurs analytiques dans les rapports

- Technique sur mesure ….
  - Identification de tableaux → Conversion HTML des documents (<table>)
  - Comment détecter tableau de résultat (liste de composés)
  - Comment détecter les valeurs et les unités de mesure
    - ✓ Unités – liste ➔ Identifier la colonne porteuse
    - ✓ Mesures ➔ Valeurs détectées