Projet ISTEX-R
2016 - 2017

ATILF - INIST - LORIA

ISTEX

Recherches exploratoires sur la base textuelle ISTEX

- Partenaires
  - ✔INIST
    - Pascal Cuxac, Nicolas Thouvenin, Sabine Barreaux
  - ✔LORIA
    - Jean-Charles Lamirel, Yannick Toussaint et 2 ans CDD ingé
  - ✔ATILF
    - Laurence Kister, Evelyne Jacquey, Etienne Petitjean et 2 ans CDD ingé

Objectifs 2016 - 2017

- Exploitation d'un corpus sur la thématique du vieillissement
  - ✔Enrichissements linguistiques et terminologiques des données textuelles
    - ➔ Extraction de contenus terminologiques et mise en forme des textes pour d'autres traitements
  - ✔Fouille de données
    - ➔ Extraction de contenus récurrents et Modélisation du domaine
- Articulation plus étroite avec les volets ISTEX Enrichissement et ISTEX Data

Le corpus vieillissement

- 8707 documents
  - ✔11 revues ELSEVIER, 4 revues OUP
    - 1995 - 2010
    - Mechanisms of Ageing and Development, Archives of Gerontology and Geriatrics, Neurobiology of Aging, Geriatric Nursing, Maturitas, Hearing Research, Experimental Gerontology, Clinics in Geriatric Medicine, Journal of Aging Studies, The American Journal of Geriatric Psychiatry, Journal of the American Medical Directors Association, Age and Ageing, The Gerontologist, The Journals of Gerontology: Series A, The Journals of Gerontology: Series B
  - ✔Sélection des textes ayant une version XML dans l'API

Enrichissements des textes

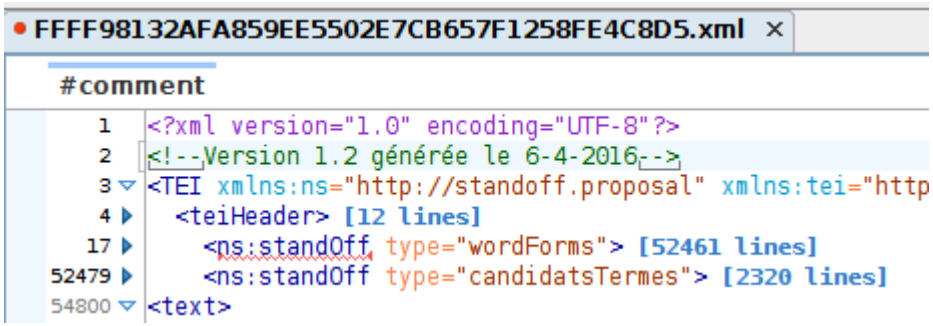
- Automatisation de l'extraction des textes : API-ISTEX
  - ✔Mémorisation des identifiants ISTEX des textes
  - ✔Vérification des textes au format XML TEI ISTEX
    - Métadonnées conformes TEI
    - Corps de texte présent et non vide, conforme TEI mais non structuré
- Adaptations de la chaîne de traitement d'un projet connexe (TermITH)
  - ✔Compatibilité de la chaîne avec l'anglais
  - ✔Modification de l'utilisation de l'extracteur terminologique TermSuite
  - ✔Optimisation

Traitements réalisés

- Définition et application d'un schéma XML pivot
  - ✔Métadonnées et Corps de texte
    - TEI-all
  - ✔Enrichissements linguistiques et terminologiques
    - StandOff proposal
- Enrichissements
  - ✔Linguistiques
    - POS-tagging (TreeTagger)
  - ✔Terminologiques
    - Extraction terminologique (TermSuite), détection des occurrences de candidats termes (chaîne TermITH)

Corpus ISTEX vieillissement traité

- Organisation des enrichissements (standOff)



- Tokenisation du texte intégral

```
<w xml:id="t372">Recruitment</w> <w xml:id="t373">Methods</w>

<w xml:id="t374">women</w> <w xml:id="t375">participate</w> <w xml:id="t376">in</w> <w xml:id="t377">bereavement</w> ·
<w xml:id="t384">participation</w> <w xml:id="t385">related</w> <w xml:id="t386">to</w> <w xml:id="t387">the</w> <w xr
<w xml:id="t398">Stroebel</w> <w xml:id="t399">found</w> <w xml:id="t400">that</w> <w xml:id="t401">those</w> <w xml:i
<w xml:id="t404">research</w> <w xml:id="t405">did</w> <w xml:id="t406">not</w> <w xml:id="t407">differ</w> <w xml:id=
<w xml:id="t421">their</w> <w xml:id="t422">review</w> <w xml:id="t423">of</w> <w xml:id="t424">the</w> <w xml:id="t42
<w xml:id="t428">studies</w> <w xml:id="t429">utilizing</w> <w xml:id="t430">death</w> <w xml:id="t431">certificates</
<w xml:id="t433">obituary-related</w> <w xml:id="t434">letters</w> <w xml:id="t435">had</w> <w xml:id="t436">the</w> ·
<w xml:id="t438">rates</w> <w xml:id="t439">of</w> <w xml:id="t440">participation</w><w xml:id="t441">,</w> <w xml:id=
<w xml:id="t461">participation.1</w> <w xml:id="t462">This</w> <w xml:id="t463">finding</w> <w xml:id="t464">may</w> <
<w xml:id="t466">that</w> <w xml:id="t467">bereavement</w> <w xml:id="t468">intervention</w> <w xml:id="t469">researc
```

Corpus ISTEX vieillissement traité

- Pos-tagging
  - ✔ standOff Proposal
  - ✔ recommandation MAF

```
<span target="#t374">
    <fs>
        <f name="lemma">
            <string>woman</string>
        </f>
        <f name="pos">
            <symbol value="NNS"/>
        </f>
    </fs>
</span>

<span target="#t375">
    <fs>
        <f name="lemma">
            <string>participate</string>
        </f>
        <f name="pos">
            <symbol value="VVP"/>
        </f>
    </fs>
</span>
```

```
<span target="#t376">
    <fs>
        <f name="lemma">
            <string>in</string>
        </f>
        <f name="pos">
            <symbol value="IN"/>
        </f>
    </fs>
</span>

<span target="#t377">
    <fs>
        <f name="lemma">
            <string>bereavement</string>
        </f>
        <f name="pos">
            <symbol value="NN"/>
        </f>
    </fs>
</span>
```

Corpus ISTEX vieillissement traité

- Candidats termes
  - ✔ 10.000 candidats maximum
  - ✔ Classés par spécificité décroissante
  - ✔ Extraits par TermSuite2.0

```
<span target="#t413" corresp="#TS2.0-entry-4484">
    <fs>
        <f name="inflexionWord">
            <string>sex</string>
        </f>
    </fs>
</span>

<span target="#t430 #t431" corresp="#TS2.0-entry-8771">
    <fs>
        <f name="inflexionWord">
            <string>death certificates</string>
        </f>
    </fs>
</span>
```

```
<span target="#t381 #t382" corresp="#TS2.0-entry-24168">
    <fs>
        <f name="inflexionWord">
            <string>gender differences</string>
        </f>
    </fs>
</span>

<span target="#t409" corresp="#TS2.0-entry-103372">
    <fs>
        <f name="inflexionWord">
            <string>nonparticipants</string>
        </f>
    </fs>
</span>

<span target="#t411" corresp="#TS2.0-entry-127">
    <fs>
        <f name="inflexionWord">
            <string>age</string>
        </f>
    </fs>
</span>
```

Corpus ISTEX vieillissement traité

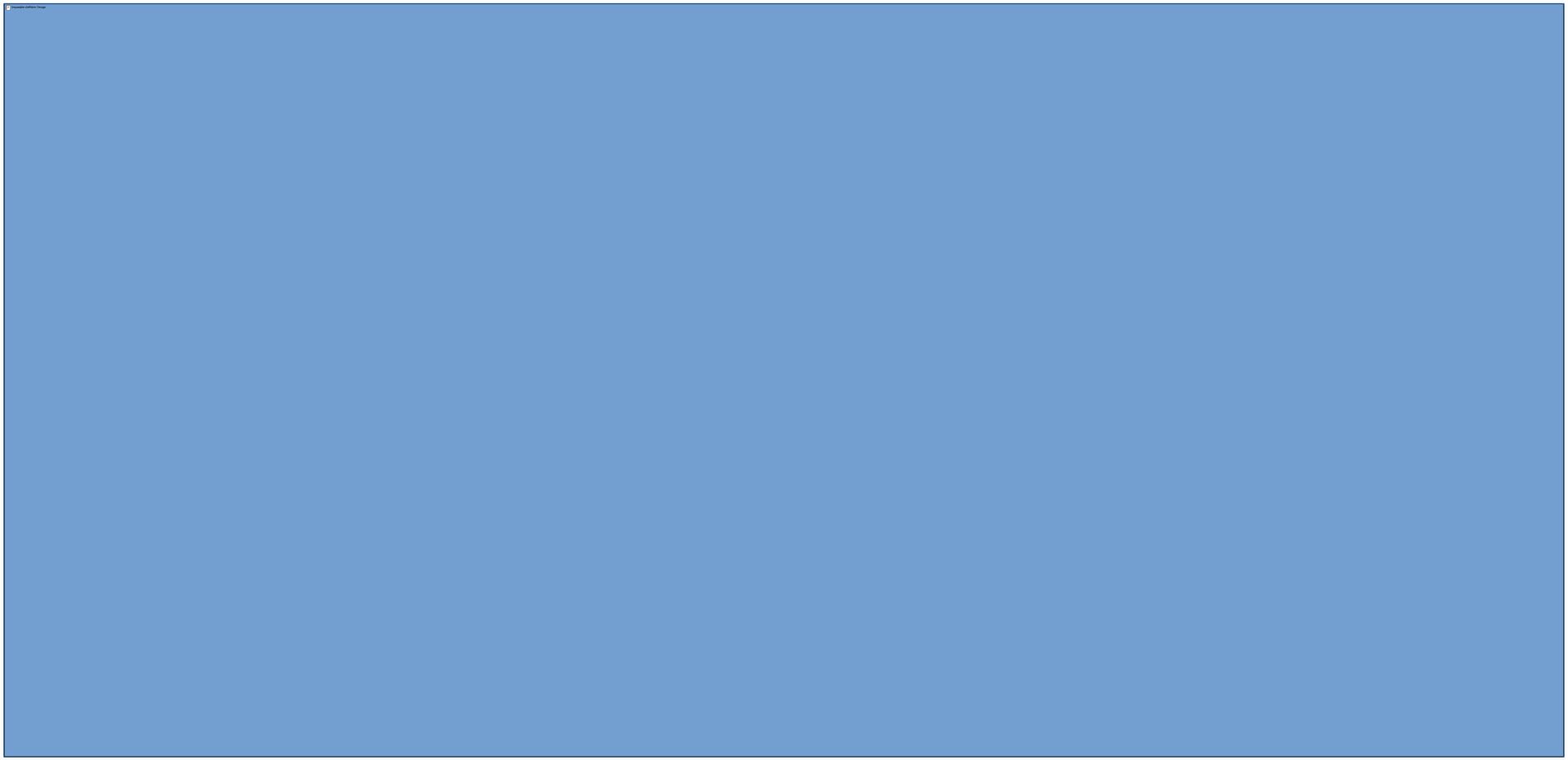
- Terminologie extraite
  - ✔TBX (ou json)

```
</termEntry>
<termEntry xml:id="entry-24168">
  <langSet xml:id="langset-24168" xml:lang="en">
    <descrip type="nbOccurrences">1961</descrip>
    <tig xml:id="term-24168">
      <term>nn: gender difference</term>
      <termNote type="termPilot">gender differences</termNote>
      <termNote type="termType">termEntry</termNote>
      <termNote type="partOfSpeech">noun</termNote>
      <termNote type="termPattern">N</termNote>
      <termNote type="termComplexity">multi-word</termNote>
      <descrip type="termSpecificity">595.0723</descrip>
      <descrip type="nbOccurrences">1961</descrip>
      <descrip type="relativeFrequency">1961.0000</descrip>
      <descrip type="formList">[{term="gender differences", count=1206}, {
term="Gender differences", count=368}, {term="gender difference", count=357}, {t
erm="Gender Differences", count=14}, {term="Gender difference", count=12}, {term
="GENDER DIFFERENCES", count=4}]</descrip>
      <descrip type="domainSpecificity">595.0722880688032</descrip>
    </tig>
  </langSet>
</termEntry>
```

Adaptations réalisées pour ISTEX

- Langue anglaise
  - ✔Tree-Tagger EN
- Extraction terminologique (collaboration étroite avec le LINA)
  - ✔Externalisation du POS-Tagging
    - Meilleur contrôle de la tokenisation
    - Utilisation possible d'autres étiqueteurs
- Optimisations et curations de la chaîne de traitement
  - ✔Passage à l'échelle
    - TermITH : 1726 documents / ~ 11 millions de tokens
    - ISTEX-vieillissement : 8707 documents / ~ 55 millions de tokens

Temps de traitement



Fonctionnalités TermITH non encore intégrées dans les traitements ISTEX

- En cours
  - ✔Phraséologie de langue générale EN (collaboration avec Mathieu Constant, ATILF)
    - *point of view* est un phrasème => élimination des occurrences de *point*, *view* des distributions de fréquence
    - DELAC anglais
  - ✔Lexique transdisciplinaire en anglais (collaboration avec Patrick Drouin)
    - LexiTrans : http://olst.ling.umontreal.ca/lexitrans/
- Exploratoires
  - ✔Désambiguïsation terminologique non supervisée

Exploitations dans ISTEX

- ISTEX-1
  - ✔Enrichissements au niveau du document
- ISTEX-2 ?
  - ✔Enrichissements au niveau des contenus (tokens) du document ?
    - ☐ Visualisation dans le PDF des documents
    - ☐ Enrichissements cliquables des tokens
      - ✗ Création de sous-corpus en fonction des annotations

Un exemple de visualisation

- Avec les enrichissements actuels

women participate in bereavement re-search, suggesting [gender differences]#TS2.0-entry-24168 in participation related to the presence or absence of depression. Also, Stroebe and Stroebe[1] found that those participating in research did not differ from [nonpartici-pants]#TS2.0-entry-103372 in [age]#TS2.0-entry-127, [sex]#TS2.0-entry-4484, or years married. Last, their review of the literature suggested that studies utilizing [1death [2certificates2]#TS2.0-entry-87451]#TS2.0-entry-8771 and/or obituary-related letters had the lowest rates of participation, whereas studies us-ing referral sources (e.g., physician or per-sonal contact) had the highest rates of participation.[1] This [finding]#TS2.0-entry-1022 may suggest that bereavement [intervention research]#TS2.0-entry-66377 may face a trade-off between a represen-tative [sampling frame]#TS2.0-entry-2490 (e.g., death [certifi-cate]#TS2.0-entry-11037) with a [1low [2response1]#TS2.0-entry-78681 rate2]#TS2.0-entry-4466, vs. a more questionably representative [sampling frame]#TS2.0-entry-2490 (e.g., recruiting through referral or personal contacts) but possibly with a higher [participation rate]#TS2.0-entry-43396. In 21 studies of conjugal bereavement cited in the Stroe-bes' review, [response rate]#TS2.0-entry-4466 ranged from 35% to 67%.

Un exemple de visualisation

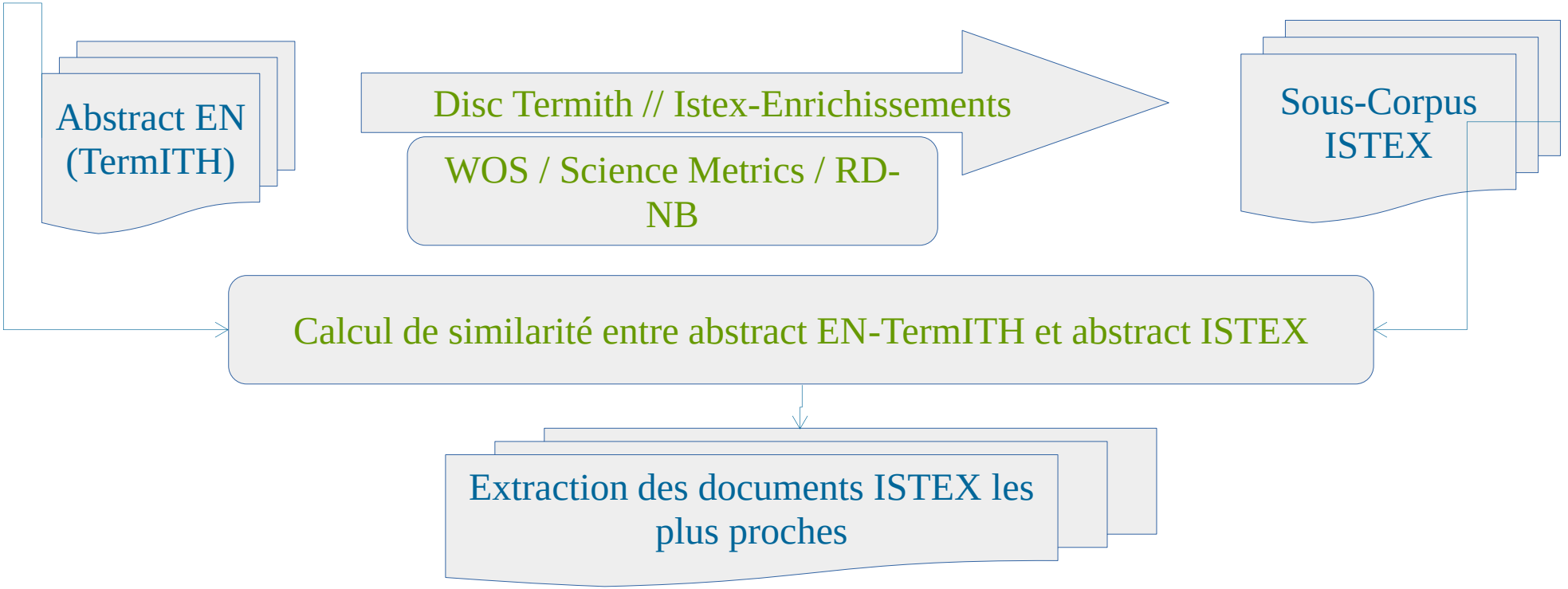
- Avec les enrichissements en cours d'intégration : un exemple de TermITH

Des ({ programmes }#lst de { recherche }#lst)#phraseo pluridisciplinaires sur l'[occupation du sol]#entry-8474 et le pastoralisme de la Préhistoire au (Moyen Âge)#phraseo dans le [1[2 sud 2]#entry-2620 du massif 1]#entry-23672 [alpin]#entry-48189 sont { menés }#lst, depuis 1998, sur les massifs du Haut Champsaur, de Freissinières et de l'Argentièrois (Hautes-Alpes). Des dix [{ phases }#lst d'occupation]#entry-1477 et d'{ activité }#lst [agropastorale]#entry-26722 (mises en évidence)#phraseo ([1 [2 prospections 2]#entry-3671 [3 pédestres 3]#entry-13190 1]#entry-13191 et fouilles), entre 1 600 et 2 700 m d'altitude, trois se { distinguent }#lst : la fin du [Néolithique]#entry-1542, l'[âge du Bronze]#entry-8318 et la [1 { période }#lst [2 médiévale 2]#entry-19069 1]#entry-19468. (Au travers des)#phraseo { premières }#lst [1{ données }#lst archéologiques 1]#entry-4742 et [environnementales]#entry-4205, cet { article }#lst { présente }#lst, depuis le { milieu }#lst du III e [millénaire]#entry-21627 au début du I er [millénaire]#entry-21627, les { grandes }#lst { caractéristiques }#lst de l'[occupation du sol]#entry-8474 mais aussi l'{originalité}#lst et l'{importance}#lst de l'[{activité}#lst {humaine}#lst]#entry-5368 dans cette {zone}#lst [alpine]#entry-48189.
[...]

Exploitations dans ISTEX

- Constitution d'un corpus comparable ISTEX en fonction du corpus TermITH

  ✔Désambiguïsation terminologique bilingue



Fouille de données

- Expérimenter les approches à base d'extraction de motifs sur du texte intégral à partir d'une description binaire

Objects / Items	a	b	c	d	e
o1		x	x		x
o2	x		x	x	
o3	x	x	x	x	
o4	x			x	
o5	x	x	x	x	
o6	x		x	x	

- L'extraction de motifs est une approche de fouille de données permettant :
  - ✔l'émergence de motifs simples, séquentiels, d'arbres ou de graphes
  - ✔l'extraction de règles d'association

Fouille de données

Itemsets of size 2 : {ab} (2/6), {ac} (4/6), {ad} (5/6), {bc} (3/6), {bd} (2/6), {cd} (4/6).

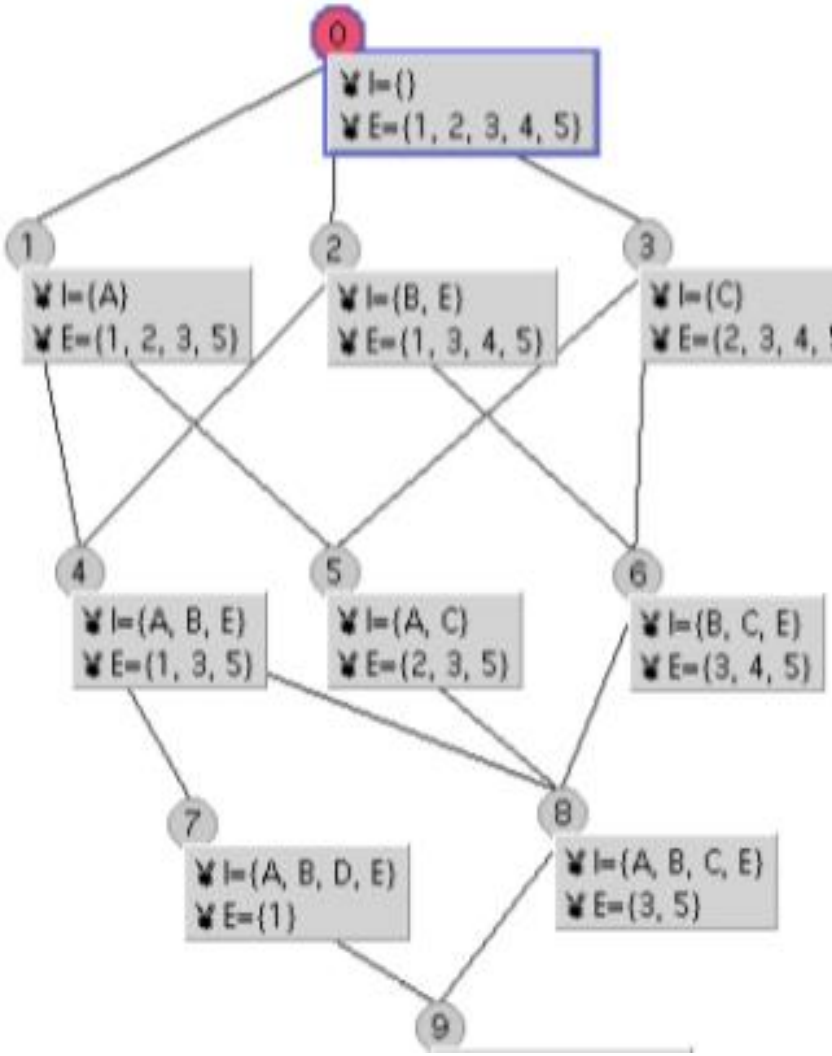
Itemsets of size 3 : {abc} (2/6), {abd} (2/6), {acd} (4/6), {bcd} (2/6).

Itemsets of size 4 : {abcd} (2/6)

{ab} ⟶ {c} (2/6,1),
{ac} ⟶ {b} (2/6,1/2),
{bc} ⟶ {a} (2/6,2/3),
{c} ⟶ {ab} (2/6,2/5),
{b} ⟶ {ac} (2/6,2/3),
{a} ⟶ {bc} (2/6,2/5) ...

Fouille de données

- ✔La construction d'un treillis
- Les usages de ces approches :
  - ✔Apprentissage (extraction d'information)
  - ✔Classification (treillis)
  - ✔La construction de connaissances
    - Lien avec les Logiques de Descriptions
    - Intégrations de connaissances dans le processus de fouille



La fouille en texte intégral

- Exemples de motifs :

La fouille en texte intégral

- Exemples de motifs :
  - ✔{decline, cognitive, growth, trajectory} (11) +
    - Rate of decline, linear decline, cognitive decline
    - growth models, growth curves
    - Aging trajectories
  - ✔{adulthood, longitudinal} (10) +
    - Longitudinal (cognitive) changes | longitudinal data analysis
    - Across|in adulthood

La fouille en texte intégral

- Les avancées : le défi d'identifier des connaissances dans le texte intégral.
  - ✔Poser les bases d'une plateforme devant intégrer de nombreux outils et interfaces :
    - Définir la granularité du traitement (document, paragraphe, phrase)
    - Méthodes de classification
    - Développement de modules de navigation dans les résultats pour l'analyse des motifs extraits
      - ✗Quantités (et similitudes) des motifs
      - ✗Quantités des objets
  - ✔La mise en place d'une classification de textes sur la base de motifs extraits
  - ✔Articulation avec d'autres approches : topic models,

Les méthodes à base de motif en texte intégral

- Les difficultés :
  - ✔ Encore beaucoup de traitement de bas niveau
  - ✔ Des motifs très courts en comparaison aux abstracts pour un même support
  - ✔ Des contextes plus riches, mais au final plus de dispersion, donc plus de données
  - ✔ Le besoin de ressources externes
  - ✔ Des textes pluri-domaines (corpus veillissement) mais difficultés à trouver les interactions entre domaines
  - ✔ Beaucoup de bruit dans les étapes à base d'apprentissage :
    - Classification (K-NN) *en pré-traitement* pour réduire le bruit et obtenir des motifs plus pertinents
    - Association d'une mesure de qualité aux motifs (la stabilité)

Les méthodes à base de motif en texte intégral

- Les perspectives :
  - ✔Prise en compte de ressources extérieures
    - Terminologie + variation
    - Des connaissances en lien avec les domaines considérés et/ou conaissances communes
    - Les Linked Open Data
      - ✗ Une complexification de l'analyse (notamment lié à la hiérarchie des concepts)
      - ✗ Difficulté à gérer les multiples points de vue sur le connaissances externes, beaucoup plus divergents avec les LOD qu'avec les ontologies d'un domaine spécifique
  - ✔Un besoin majeur de visualisation des résultats pour faire une synthèse entre plusieurs paragraphes.

AVANT-PROJET : CITATION FOCUSER (CITEFOX)

- But : mettre en place des méthodes de ciblage de citation entre documents citant et documents cités,
- Mettre en place des techniques à la frontière de l'état de l'art dans le domaine du résumé de communauté:
  - Résumé du contenu des documents basé sur la compétition de blocs

  ⇒ Théorie de la maximisation des traits,
  - Mesures de mise en correspondance textuelles

  ⇒ Etat de l'art,
  - Expansion de requêtes

  ⇒ Méthodes de propagation d'activation,
  - Extraction des citations

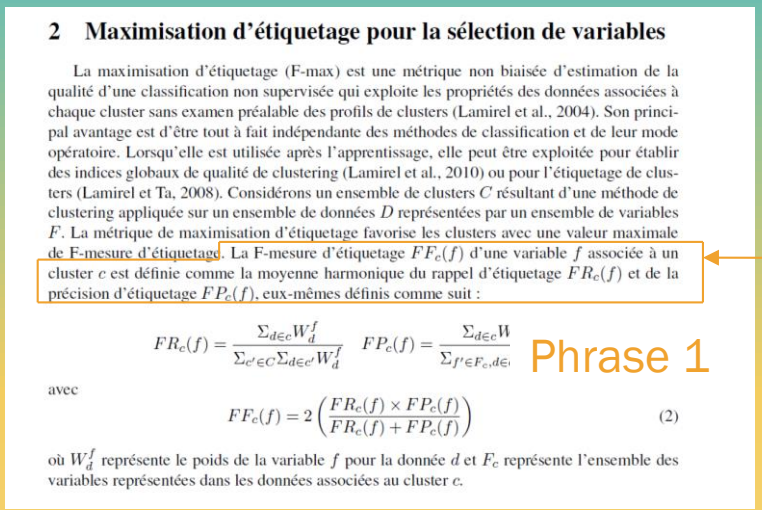
  ⇒ Méthodes d'extraction d'entités nommées ? (à voir).

- Travailler à titre d'exemple sur des données-test issues d'un challenge international (CL-SCISumm 2016).

Intervenants : Jean-Charles LAMIREL (Synalp-LORIA), Hazem AL ZIED (ATILF), Nicolas DUGUE (Université du Mans).

EXTRACTION ET CIBLAGE DES CITATIONS

EGC 2014, Lamirel et al.



2 Maximisation d'étiquetage pour la sélection de variables

La maximisation d'étiquetage (F-max) est une métrique non biaisée d'estimation de la qualité d'une classification non supervisée qui exploite les propriétés des données associées à chaque cluster sans examen préalable des profils de clusters (Lamirel et al., 2004). Son principal avantage est d'être tout à fait indépendante des méthodes de classification et de leur mode opératoire. Lorsqu'elle est utilisée après l'apprentissage, elle peut être exploitée pour établir des indices globaux de qualité de clustering (Lamirel et al., 2010) ou pour l'étiquetage de clusters (Lamirel et Ta, 2008). Considérons un ensemble de clusters C résultant d'une méthode de clustering appliquée sur un ensemble de données D représentées par un ensemble de variables F. La métrique de maximisation d'étiquetage favorise les clusters avec une valeur maximale de F-mesure d'étiquetage. La F-mesure d'étiquetage $FF_c(f)$ d'une variable f associée à un cluster c est définie comme la moyenne harmonique du rappel d'étiquetage $FR_c(f)$ et de la précision d'étiquetage $FP_c(f)$, eux-mêmes définis comme suit :

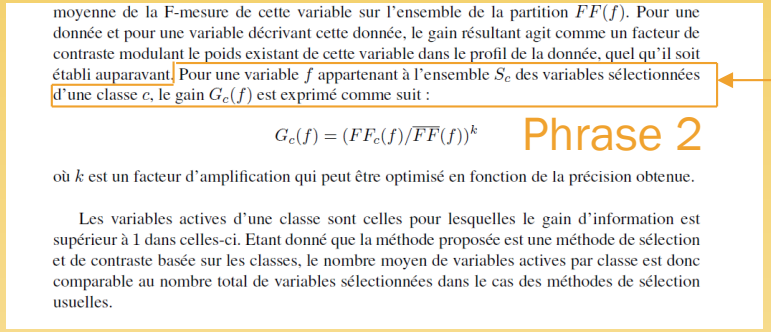
$$FR_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{c' \in C} \Sigma_{d \in c'} W_d^f} \qquad FP_c(f) = \frac{\Sigma_{d \in c} W}{\Sigma_{f' \in F_c, d \in}}$$

Phrase 1

avec

$$FF_c(f) = 2\left(\frac{FR_c(f) \times FP_c(f)}{FR_c(f) + FP_c(f)}\right) \quad (2)$$

où W_d^f représente le poids de la variable f pour la donnée d et F_c représente l'ensemble des variables représentées dans les données associées au cluster c.



moyenne de la F-mesure de cette variable sur l'ensemble de la partition $FF(f)$. Pour une donnée et pour une variable décrivant cette donnée, le gain résultant agit comme un facteur de contraste modulant le poids existant de cette variable dans le profil de la donnée, quel qu'il soit établi auparavant. Pour une variable f appartenant à l'ensemble S_c des variables sélectionnées d'une classe c, le gain $G_c(f)$ est exprimé comme suit :

$$G_c(f) = (FF_c(f)/\overline{FF}(f))^k$$

Phrase 2

où k est un facteur d'amplification qui peut être optimisé en fonction de la précision obtenue.

Les variables actives d'une classe sont celles pour lesquelles le gain d'information est supérieur à 1 dans celles-ci. Etant donné que la méthode proposée est une méthode de sélection et de contraste basée sur les classes, le nombre moyen de variables actives par classe est donc comparable au nombre total de variables sélectionnées dans le cas des méthodes de sélection usuelles.

Les cahiers du numérique 2016, Dugué et al.

Ciblage

Rappel, Précision, F-mesure

Extraction terminologique

…. La F-mesure de traits combine les indices de Rappel et de Précision (Lamirel et al., 2014) ….

Ciblage

Contraste, Gain, Information, Classe

Extraction terminologique

…. Le gain d'information, ou contraste, exprime la capacite du traits a caractériser une classe (Lamirel et al., 2014) ….

Résumé de communauté

Phrase 1. Phrase 2.

CAS DU CHALLENGE CL-SCISUMM

- Corpus de 10 articles de référence dans le domaine biomédical,
- Un ensemble de plus de 10 articles citant accompagnés du contexte des citations est associé avec chaque article cité,
- Un Gold (citation-meilleur phrase/paragraphe du document cité) est construit par expertise.
- Extraction terminologique et résumé du document citant par maximisation d'étiquetage,
- Extraction terminologique des termes dans les phrases support des citations,
- (Propagation d'activation pour enrichir les termes des citations),
- Analyse de la densité de contraste générée par les termes des citations dans les blocs du document cité,
- Ciblage des meilleures paragraphes/phrases par mesure de similarité,
- Adjonction des phrases sélectionnées au résumé de communauté,
- Mesure de validité des résultats en exploitant des méthodes de la classe ROUGE.

Une expérimentation préalable a été menée sur du texte non structuré.

ETAT D'AVANCEMENT

- Nous avons présenté une méthodologie de trouver les points d'impacts des citations de documents citant dans les documents cités,
- Cette méthodologie a été testé avec succès dans le cadre du challenge CL-SCISumm 2016,
- Le système obtenu s'est avéré être le plus efficace parmi les 17 systèmes proposés dans le challenge (avec une très forte différence en terme de rappel),
- Le système ne nécessite pas de source de connaissance externe pour apprendre (contrairement aux systèmes concurrents) et possède des capacités naturelles d'élimination de la redondance,
- La méthodologie et les test ont été publiés dans un journal international (IJDL 2017),
- Un problème important dans le cadre ISTEX est celui du repérage cohérent des citations et d'extraction du contexte,
- Cette proposition de projet reste à financer (le financement de l'étude préalable a été opéré dans le cadre du CPER LCHN).

AVANT-PROJET : NOUVEAUX PARADIGMES SCIENTIFIQUES

- But : explorer un corpus de données scientifiques en mesurant les changements de sujets incluant la récurrence de sujets (alternance de citations et d'oublis,
- Travailler sur un corpus de données issues de la base mutti-éditeurs ISTEX gérée par l'INIST,
- Mettre en place des techniques à la frontière de l'état de l'art :
  - Distances de compromis entre la généralité et la discrimination

  ⇒ Théorie de la maximisation des traits,
  - Travailler avec des vues multiples et des mécanismes de généralisation en ligne

  ⇒ Paradigme MVDA,
  - Intégrer la visualisation

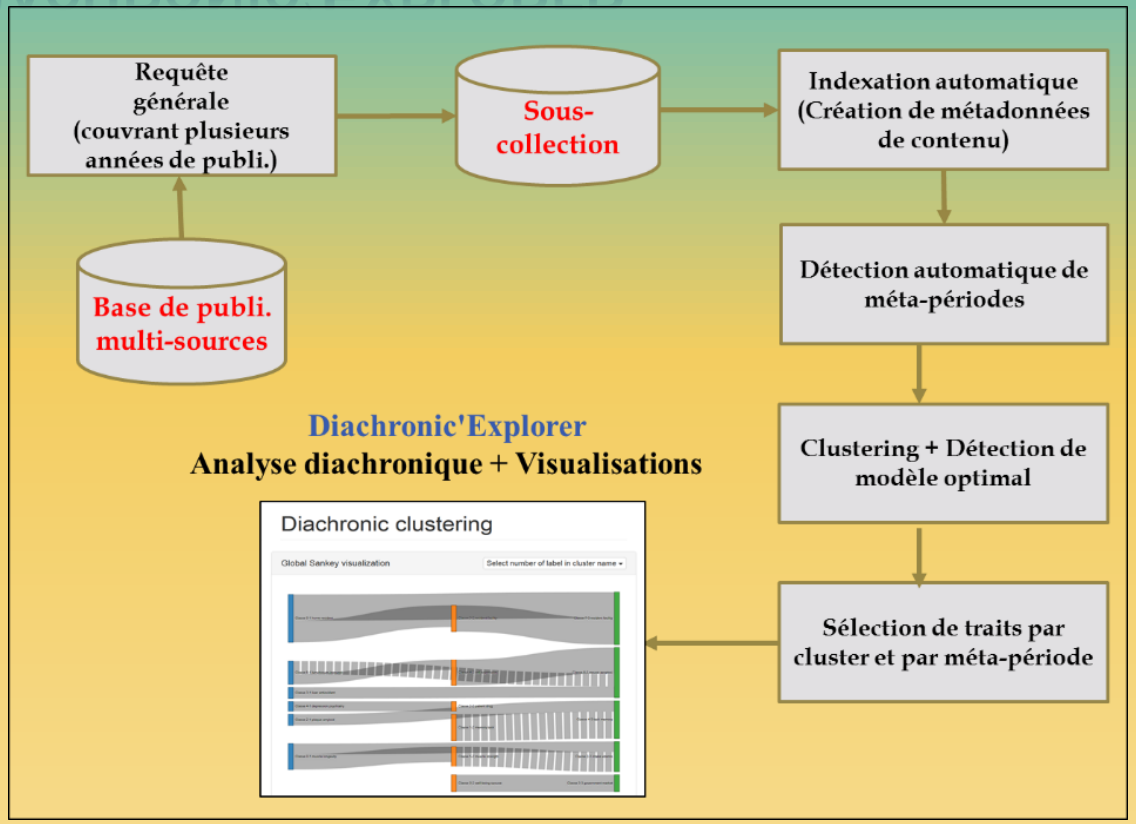
  => Approche Diachronic'Explorer,
  - Intégrer les informations produites par les entités nommées dans le processus (en cours),
- Travailler à titre d'exemple sur des données du domaine de l'astronomie (en cours).

Intervenants : Jean-Charles LAMIREL (Synalp-LORIA), Denis MAUREL (Université de Tours), Anubhav GUPTA (DIST-CNRS & Université de Tours).

Analyse diachronique et navigation dans les données multisources

ISTEX-R - WP1 & Diachronic'Explorer

La figure présente le déroulement de l'approche Diachronic'Explorer complète jusqu'à la visualisation



La méthode ne présente pas les inconvénients des méthodes d'extraction de sujets usuelles, comme LDA (Blei et al. 2003) : sujets imprécis et dépendants du processus d'optimisation utilisé, non applicabilité à l'échelle des documents, ...

L'indexation automatique peut être remplacée par le processus d'extraction de métadonnées basé sur la maximisation des traits.

Cas des données astromoniques

- Corpus d'env. 500000 articles sur le thème général de l'astronomie de issues de la base ISTEX,
- Période couvrant 189 ans,
- Les données sont étiquetées par les entités nommées, noms de lieu, nom de personnes, dates
  => Unitex/CacSys [Maurel et al. 2016].
- Identification des articles les plus cités,
- Analyse du contenu par extraction automatique des métadonnées et isolement de sujet centraux (ex: big bang, théorie des cordes ….),
- Extraction du contexte des citations (phrases dans lesquelles les citations apparaissent) [Al Zied et al. 2017],
- Mesure du cumul de contraste/période généré sur les sujets centraux par le contexte des citations,
- Visualisation des variabilités temporelles,
- Mise en parallèle avec une analyse directe basée sur le clustering et sur MVDA.

Cas des données astromoniques

Un exemple d'annotations dans un article :

18CBC6B00F42A58322ECB8DA287F002C5D828ACD - Notepad

File Edit Format View Help

```
                <persName change="#Unitex-3.2.0-alpha" resp="istex-rd" scheme="http://persName-entity.lod.istex.fr
                        <term>Martin Heidegger</term>
                        <fs type="statistics">
                                <f name="frequency">
                                        <numeric>1</numeric>
                                </f>
                        </fs>
                </persName>
        </annotationBlock>
        <annotationBlock corresp="text" xmls="https://wwww.tei-c.org/ns/1.0">
                <persName change="#Unitex-3.2.0-alpha" resp="istex-rd" scheme="http://persName-entity.lod.istex.fr
                        <term>Marcus Hellyer</term>
                        <fs type="statistics">
                                <f name="frequency">
                                        <numeric>2</numeric>
                                </f>
                        </fs>
                </persName>
        </annotationBlock>
        <annotationBlock corresp="text" xmls="https://wwww.tei-c.org/ns/1.0">
                <persName change="#Unitex-3.2.0-alpha" resp="istex-rd" scheme="http://persName-entity.lod.istex.fr
                        <term>Francisco Suárez</term>
                        <fs type="statistics">
                                <f name="frequency">
                                        <numeric>9</numeric>
                                </f>
                        </fs>
                </persName>
        </annotationBlock>
        <annotationBlock corresp="text" xmls="https://wwww.tei-c.org/ns/1.0">
                <persName change="#Unitex-3.2.0-alpha" resp="istex-rd" scheme="http://persName-entity.lod.istex.fr
                        <term>Raymond Bullman</term>
```

- Nous avons présenté une méthodologie permettant d'analyser les données de manière diachronique à partir des données étiquetées par des entités nommées,
- Le principe de cette méthodologie a été présenté à la conférence ACFAS 2017 (Montréal),
- Le but est d'analyser les effets de récurrence liés aux nouveaux paradigmes scientifiques,
- Le principe général de l'approche repose sur des techniques récemment expérimentées avec succès,
- Les données étiquetées restent cependant en cours de traitement,
- Un problème important est celui du repérage cohérent des citations dont la syntaxe varie en fonction des périodes de temps,
- Le repérage du contexte des citations reste également à traiter,
- Cette proposition de projet reste à financer.