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

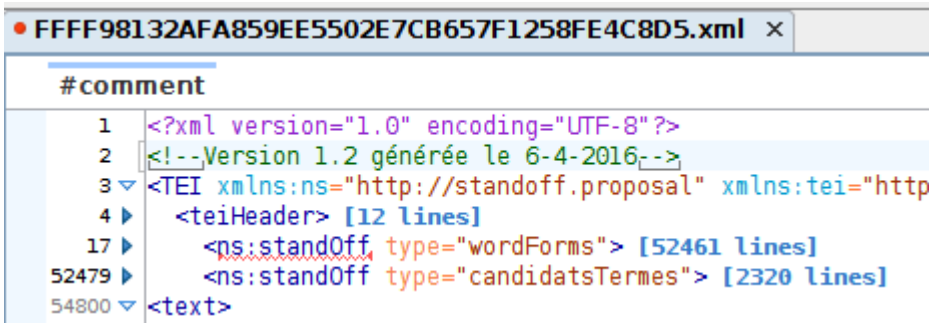
- Automatisation de l'extraction des textes : API-ISTEX
  - ✔Mémorisation des identifiants ISTEX des textes
  - ✔Vérification des textes au format XML TEI ISTEX
    - Métadonnées conformes TEI
    - Corps de texte présent et non vide, conforme TEI mais non structuré
- Adaptations de la chaîne de traitement 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

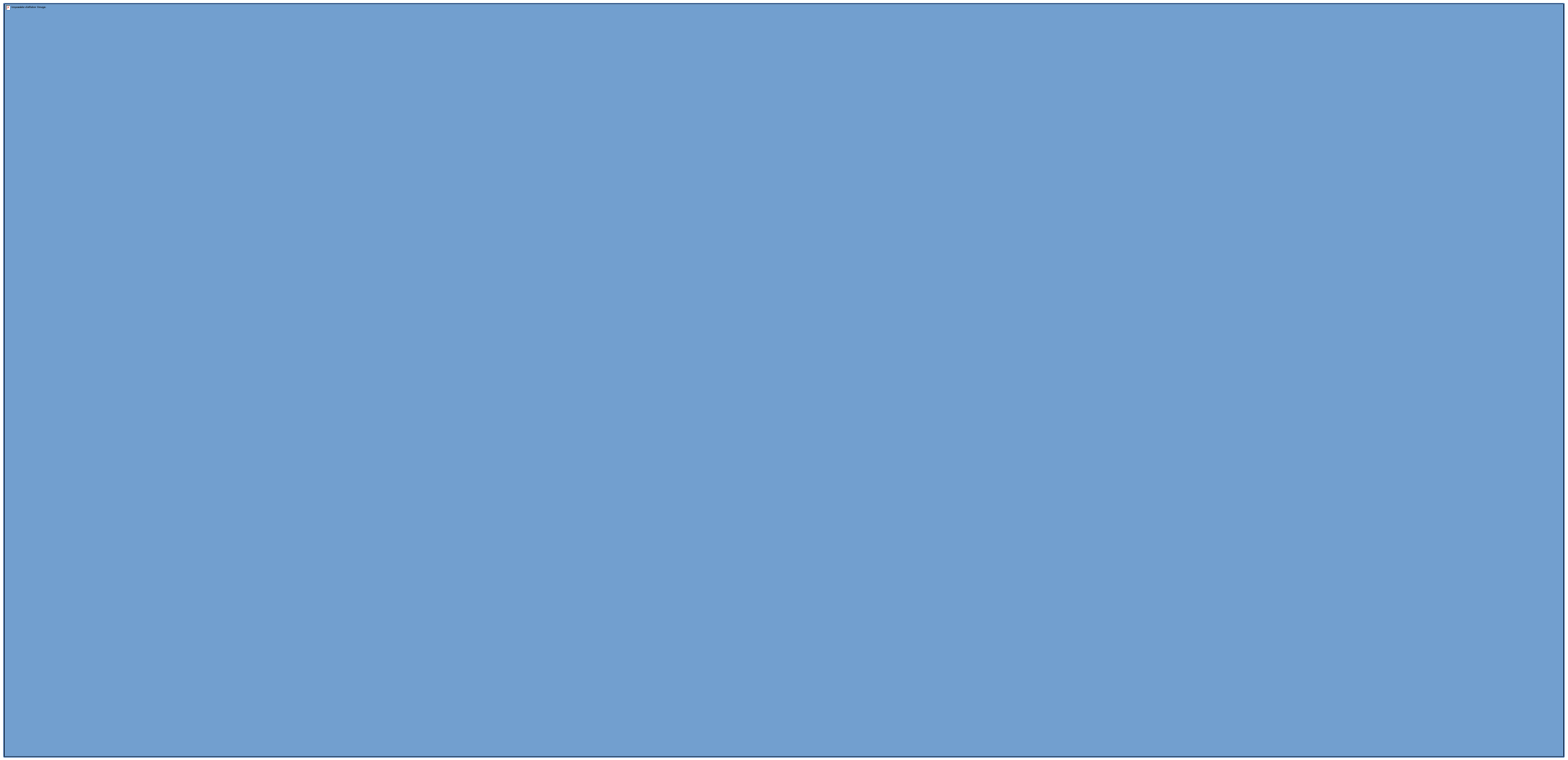
- Terminologie extraite
  - ✔TBX (ou json)

```
</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

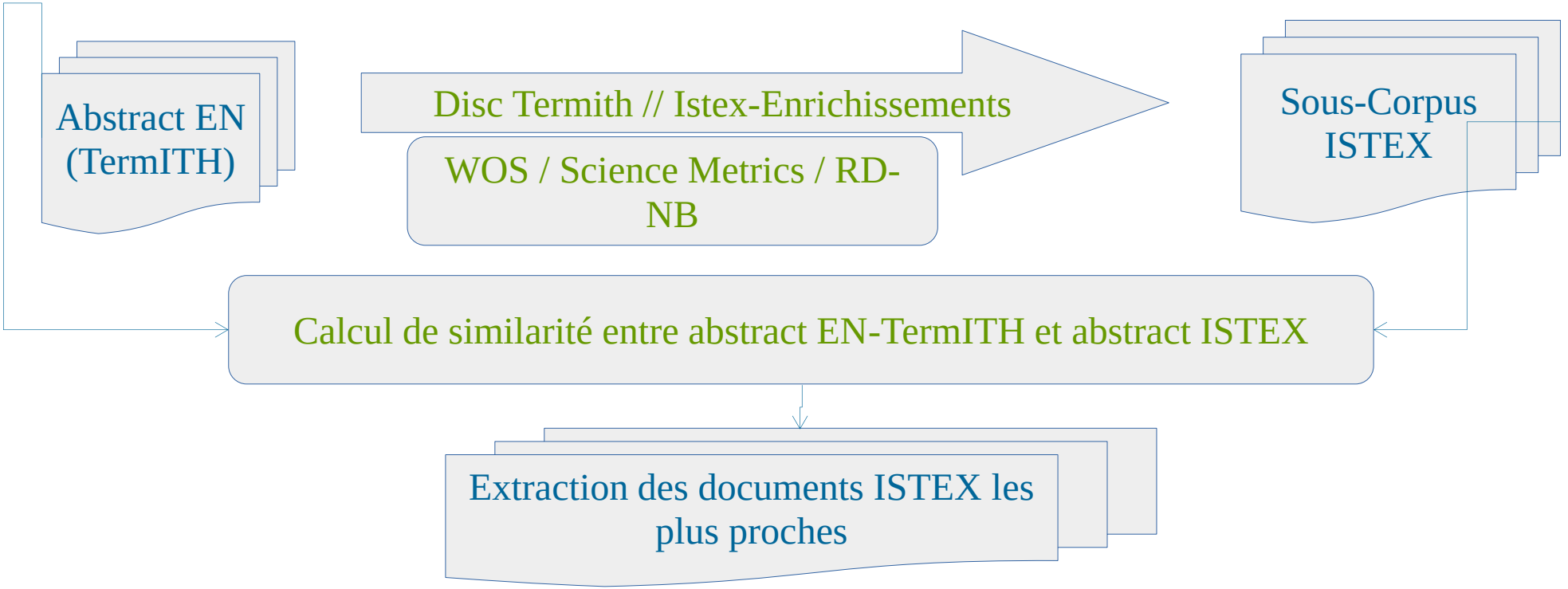
- Avec les enrichissements en cours d'intégration : un exemple de TermITH

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).

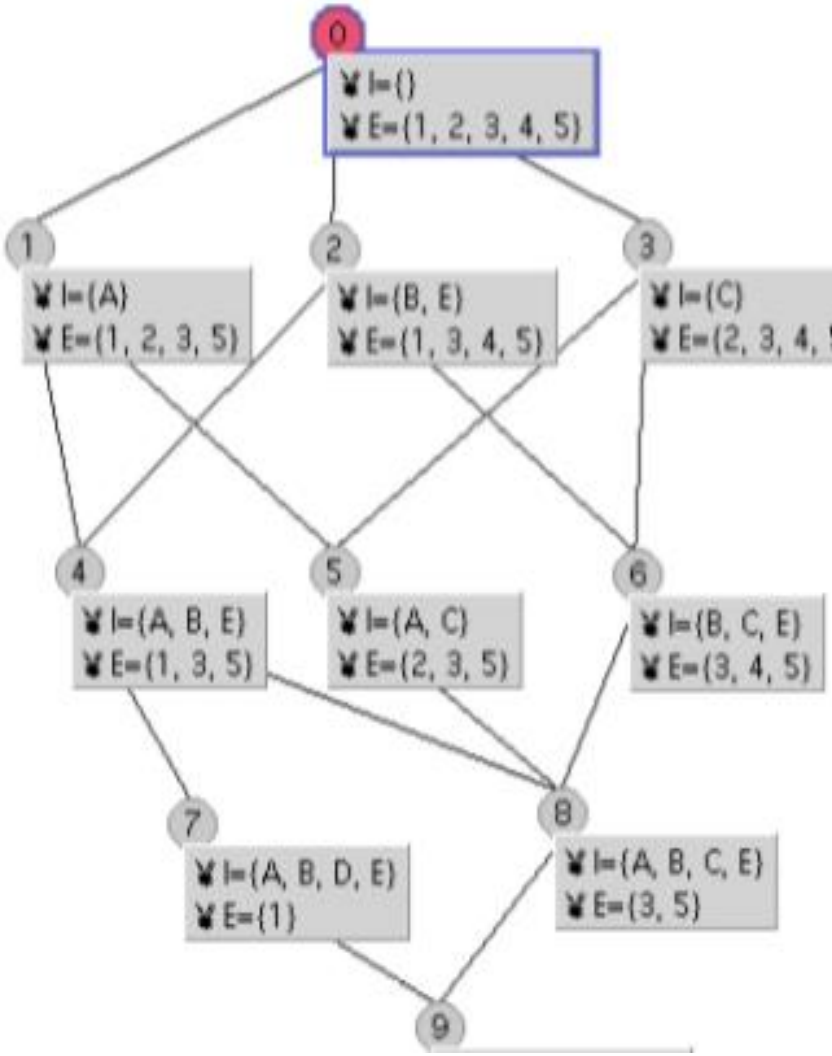
Itemsets of size 3 : {abc} (2/6), {abd} (2/6), {acd} (4/6), {bcd} (2/6).

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.