

# IBM SolutionsConnect 2013

L'IBM TechSoftware nouvelle génération

28, 29 et 30 août  
IBM Client Center Paris



#solconnect13

*Transformez vos opportunités en succès*



# IBM SolutionsConnect 2013

L'IBM TechSoftware nouvelle génération

## IND06P1 SmartCare ICA Studio

Jean-Marc Langé

**IBM Content Analytics with Enterprise Search**

2

# ICA Studio

- **Un environnement de conception permettant de paramétrer les annotateurs chargés d'identifier des concepts métier à partir des sources textuelles, sans nécessité d'utiliser un quelconque langage de programmation.**
- **Permet de gérer les projets, de concevoir et tester les annotateurs dans un environnement graphique, et de les déployer vers le moteur ICA.**
- **Utilise deux ressources essentielles pour générer des annotations:**
  - **dictionnaires** qui contiennent la terminologie pertinente pour l'objectif métier visé;
  - **règles** qui combinent des mots et des annotations déjà identifiées pour générer une annotation de niveau supérieur;

# Comment ça marche?

## Content Analytics Studio



Annotateurs (UIMA)

API REST Temps Réel

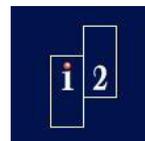
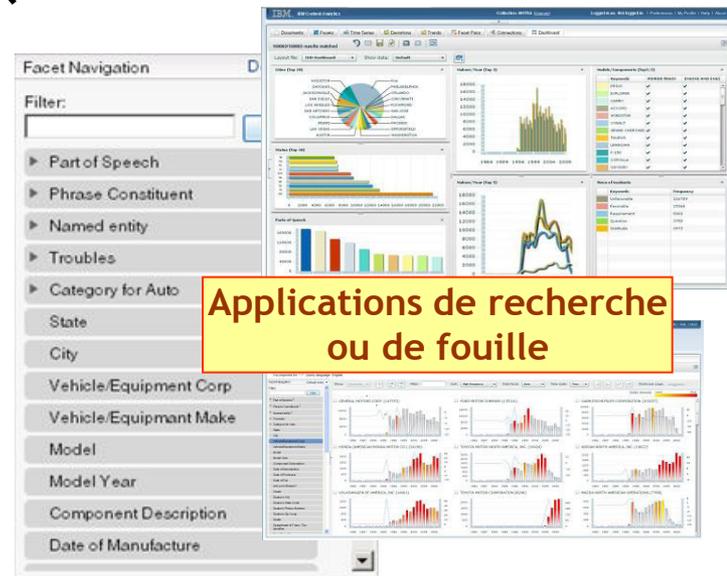
Documents et données analysés  
Avec identification de concepts

Crawlers

Crawlers

### Données non structurées

Entreprise (Centre de Contact, logs de test, notes des concessionnaires / conseillers, messagerie, ECM, etc.) et externes (Web, forums utilisateurs, réseaux sociaux, etc.)



Cognos software

Applications tierces d'analyse / exploration



## Qu'est-ce que **TAL (NLP)** et **UIMA**

- Traitement automatique du langage naturel (TAL) ou *Natural Language Processing (NLP)*, **le lien qui rend possible l'interaction entre les ordinateurs et les langages humains**
  - Watson utilise **IBM Content Analytics** pour les fonctions NLP
- **Unstructured Information Management Architecture (UIMA), une architecture ouverte pour le traitement des données textuelles et la définition de solutions d'analyse de texte**
  - Standard Open Source OASIS
  - Projet supporté par Apache Software Foundation
  - Mis en oeuvre dans plusieurs solutions IBM



# Aperçu général de ICA Studio

Projets:  
configuration,  
ressources  
(dictionnaires,  
règles), documents  
de test...

Documents testés  
avec une  
configuration  
d'annoteurs

Annotations  
identifiées dans le  
document

The screenshot displays the ICA Studio interface. The main window shows a document titled "antécédents\_familiaux.txt" with the following text:

GRANDMÈRE AVEC PR.  
père : K generalise.  
FILLE AVEC ASTHME.  
absent.  
CARDIOPATHIE GLAUCOME ASTHME POLYARTHRITE.  
fille a un Lupus , LED, suivit au CHU ( Pr Jorgensen).  
Néant.  
MERE AVEC PR.  
O.  
Néo pulmonaire chez la mere.  
ostéoporose maternelle.  
- gd-mère paternelle : cancer hépatique. -  
gd-père paternel : cancer du rein. - père :  
hypercholestérolémie, cancer de prostate. -  
mère : hypothyroïdie, ostéoporose.  
RAS.  
père pontages.

The left sidebar shows the project structure:

- Projets
- Configuration
- Documents
- Ressources
  - Break Rules
  - Character Rules
  - Dictionaries
  - Parsing Rules
  - Semantic
  - SNOMED
    - Dictionaries
    - Parsing Rules
  - X\_CIM-10
- Results
- Projets ICPA-PR

The right sidebar shows the "Outline" view with "Annotations By Type":

- com.ibm.ICA.general.DictUnitésMesu
- com.ibm.ICA.general.Famille
- com.ibm.ICA.general.fr.valeurMesuré
- com.ibm.ICA.general.fr.valeurNuméri
- com.ibm.ICA.medical.Dict\_SNO\_Diag
- com.ibm.ICA.medical.Dict\_SNO\_Fonc
- com.ibm.ICA.medical.Dict\_SNO\_Mod
- com.ibm.ICA.medical.Dict\_SNO\_Morq
- com.ibm.ICA.medical.Dict\_SNO\_Phys
- com.ibm.ICA.medical.Dict\_SNO\_Sociz
- com.ibm.ICA.medical.Dict\_SNO\_Topc
- @ poumon
- @ cutanée
- @ colique

The bottom status bar shows "Property" and "Value" fields.

# Configuration de la chaîne d'annotations

The screenshot displays the UIMA Pipeline Configuration window. On the left, a file explorer shows a project structure with folders like 'Languages' and 'Resources', and files like 'Antécédents.annoconfig'. The main window is titled 'UIMA Pipeline Configuration' and shows a list of 'UIMA Pipeline Stages' including 'Document Language', 'Lexical Analysis', 'Parsing Rules', and 'Clean Up'. The 'Lexical Analysis' stage is selected, and its configuration panel is visible on the right. This panel includes a 'Languages' list with 'French [set]' selected, and a 'Dictionaries for language French' list containing several built-in and external dictionaries. A blue callout bubble points to the 'Lexical Analysis' stage in the list, and another points to the 'Dictionaries for language French' list. A third callout bubble points to the 'Dictionaries for language French' list.

**Etape d'analyse lexicale (dictionnaires)**

**Etape d'analyse par les règles**

**Configuration des dictionnaires utilisés pour l'analyse de mots et termes**

UIMA Pipeline Configuration

UIMA Pipeline Stages

Select a stage to see the details

- Document Language
- Lexical Analysis**
- Parsing Rules
- Clean Up

Lexical Analysis

Set the lexical analysis configuration.

Languages

- French [set]**
- German
- Greek
- Hebrew
- Italian
- Japanese

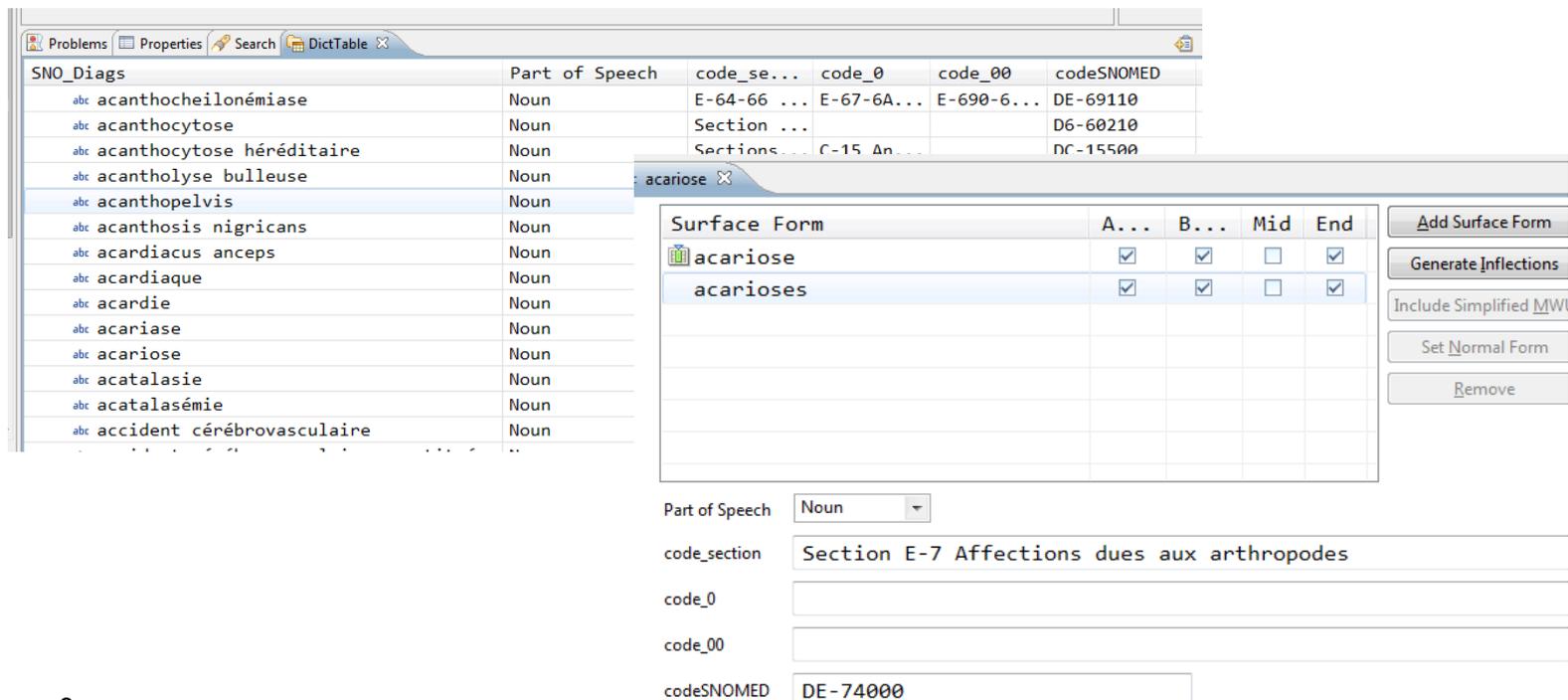
Dictionaries for language French

- Built in Lex dictionary: fr-XX-LLex-7009.dic
- Built in OOV dictionary: fr-XX-OOV-7002.dic
- /Medical/Resources/SNOMED/Dictionaryes/Dict\_SNO\_Diags.dic
- /Medical/Resources/SNOMED/Dictionaryes/Dict\_SNO\_Fonc.dic
- /Medical/Resources/SNOMED/Dictionaryes/Dict\_SNO\_Modif.dic
- /Medical/Resources/SNOMED/Dictionaryes/Dict\_SNO\_Morph.dic
- /Medical/Resources/SNOMED/Dictionaryes/Dict\_SNO\_Phys.dic
- /Medical/Resources/SNOMED/Dictionaryes/Dict\_SNO\_Social.dic

Use F2 to display a description of the selected dictionary.

# Paramétrage des dictionnaires

- Les dictionnaires peuvent être remplis « à la main » ou par importation de fichiers en format délimité
- Ils contiennent les entrées, leur catégorie grammaticale, et les variantes «de surface»: formes fléchies (pluriel, féminin, formes conjuguées), abréviations...
- On peut associer à une entrée des attributs définis par l'utilisateur (type, classification, codification, ... )



SNO_Diags	Part of Speech	code_se...	code_0	code_00	codeSNOMED
abc acanthocheilonémiase	Noun	E-64-66 ...	E-67-6A...	E-690-6...	DE-69110
abc acanthocytose	Noun	Section ...			D6-60210
abc acanthocytose héréditaire	Noun	Sections...	C-15 An...		DC-15500
abc acantholyse bulleuse	Noun				
abc acanthopelvis	Noun				
abc acanthosis nigricans	Noun				
abc acardiacus anceps	Noun				
abc acardiaque	Noun				
abc acardie	Noun				
abc acariase	Noun				
abc acariose	Noun				
abc acatalasie	Noun				
abc acatalasémie	Noun				
abc accident cérébrovasculaire	Noun				

Surface Form	A...	B...	Mid	End
acariose	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
acarioses	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Part of Speech: Noun

code\_section: Section E-7 Affections dues aux arthropodes

code\_0:

code\_00:

codeSNOMED: DE-74000

## Création de règles

- **Les règles sont créées en sélectionnant dans un document une séquence de mots correspondant au concept que l'on souhaite annoter, et en la faisant glisser dans l'espace de création des règles**
- **Dans l'espace de création, on retrouve les différentes annotations déjà identifiées (dictionnaires ou règles) et des mots « bruts » (non annotés)**
- **Pour chacun de ces éléments (annotations ou mots bruts), on peut sélectionner des critères (maj/min, catég. grammaticale, valeur, nbre de caractères, etc), ainsi qu'une possible répétition.**
- **Une fois définie la séquence à annoter et les critères, on lui associe une nouvelle annotation, et optionnellement des attributs, qu'on peut piocher dans les mots ou annotations composant cette séquence**

# Création de règle: antécédents médicaux familiaux

## Etape 1: drag&drop dans la zone de création de règles

antécédents\_familiaux.txt

néant.  
pas de PR, pas de maladie auto immune  
RAS.  
Ø.  
Rhumatisme inclassable chez la  
mère, spondylarthrite chez l'oncle, syndrome de  
Sharp chez la soeur, syndrome grave des  
antiphospholipides chez la cousine germaine.  
1979.  
tante : PR.  
Mère atteinte de PR.  
RAS.  
GRAND MERE : PR ?.  
Grand mère avec rhumatisme psoriasique.  
RAS.  
Père : IDM (décès).  
PR chez une soeur, mère et grand-mère.  
MERE DE 92 ANS "polyarthrosique".  
Mère porteuse d'une PR et d'une ostéoporose.  
Grand mère polyarthrite rhumatoïde .  
2 cousins germains ont une PR.  
ras.  
1 nièce SEP , fille de sa soeur.  
Pere: Behcet.  
père dcd d'un cancer de la prostate à 70 ans.  
pr chez une soeur.  
neant.  
PR chez une tante maternelle.  
Mère diabète insuliné. .  
2006-09-12 00:00:00.0.  
aucun.  
arthrose.  
père:eczéma de contact, LLC.. Mère: maaldie de

Outline \*Create Parsing Rules

Rule Type: Phrases  Fire all rules at this level

Using Config: Antécédents.annoconfig

Annotation Constraints Properties

Rule Config

Rule Set: Default

Input Text: Grand mère avec rhumatisme psoriasique

Add Annotations

Type:

- 1: Famille
  - Features
    - Value = Grand mère [Special]
    - Lemma = grand-mère [Special]
    - Part of Speech = Noun [Special]
    - Length = 10 [Special]
  - Subtree
- 2: Token
  - Features
    - Type = LowercaseAlphabetic [Special]
    - Value = avec [Special]
    - Lemma = avec [Special]
    - Part Of Speech = Noun (Common) [Special]
    - Length = 4 [Special]
    - dictionaryMatch = true [Boolean]
- 3: Dict\_SNO\_Diags
  - Features
    - Value = rhumatisme psoriasique [Special]
    - Lemma = rhumatisme psoriasique [Special]
    - Part of Speech = Noun [Special]
    - Length = 22 [Special]

drag & drop de la séquence textuelle

Annotation résultant du dictionnaire de liens familiaux

« mot brut » (sans annotation particulière)

Annotation résultant du dictionnaire de diagnostics médicaux

# Création de règle: antécédents médicaux familiaux

## Etape 2: spécification de contraintes sur les éléments de la règle

...tisme psoriasique.

...re et grand-mère.  
...arthrosique".  
...R et d'une ostéoporose.  
...te rhumatoïde .  
...it une PR.

...le sa soeur.

...de la prostate à 70 ans.

...ernelle.

...  
...).

...t, LLC.. Mère: maaldie de

The screenshot shows a rule editor interface. On the left, a list of rule elements is displayed, including '1: Famille' and '2: Taken'. A context menu is open over '2: Taken', showing options like 'Repeats...', 'Replace With', 'Remove...', 'Add Annotation', 'Move', and 'Group...'. The 'Repeats...' option is selected, and the 'Advanced Repeat Options' dialog box is open. The dialog box has a title bar 'Advanced Repeat Options' and a main area with the following options:

- time(s) exactly
- time(s) or more 3
- to

There is a numeric input field with the value '0' and a 'Repeat' icon. At the bottom of the dialog are 'OK' and 'Cancel' buttons.

Le terme annoté «famille» et le terme annoté «diagnostic» peuvent être séparés par 0 à 3 mots (quels qu'ils soient)

D'autres contraintes pourraient être précisées, comme la valeur d'un mot particulier ou sa catégorie grammaticale

# Création de règle: antécédents médicaux familiaux

## Etape 3: création de l'annotation pour la séquence paramétrée

Outline \*Create Parsing Rules

Rule Type: Phrases  Fire all rules at this level

Using Config: Antécédents.annoconfig

Selection Annotation Constraints Properties

- 1: Famille
  - Features
  - Subtree
- 2: Token
  - Features
  - Subtree
- 3: Dict\_SNO\_Diags
  - Features
  - Subtree

Delete Annotation

Insert Annotation

### Insert Annotation

Enter an annotation type or select an existing name

Annotation Type: com.ibm.ICA.medical.Antécédent\_familial

Existing annotation types:

Outline \*Create Parsing Rules

Rule Type: Phrases  Fire all rules at this level

Using Config: Antécédents.annoconfig

Selection Annotation Constraints Properties

Antécédent\_familial [Created by this Rule]

- lien parenté = grand-mère [String]
- diagnostic = rhumatisme psoriasique [String]
- Subtree
  - 1: Famille
    - Features
      - Value = Grand mère [Special]
      - Lemma = grand-mère [Special]
      - Part of Speech = Noun [Special]
      - Length = 10 [Special]
    - Subtree
  - 2: Token
    - Features
      - Value = rhumatisme psoriasique [Special]
      - Lemma = rhumatisme psoriasique [Special]
      - Part of Speech = Noun [Special]
      - Length = 22 [Special]
    - Subtree
  - 3: Dict\_SNO\_Diags
    - Features
    - Subtree

Full

Des attributs peuvent être rajoutés à l'annotation par drag&drop d'attributs des composants de la séquence

# Création de règle: antécédents médicaux familiaux

## Etape 4: vérification de l'application de la règle à l'échantillon de texte

The screenshot displays a text editor window titled '\*antécédents\_familiaux.txt' containing medical text. Several lines are highlighted in red, indicating they are covered by a rule. A yellow callout box points to a specific instance of the annotation, showing its configuration details. A blue callout box points to the text editor, explaining that the new annotation appears in the preview with the covered text sequences. Another blue callout box points to the configuration window, explaining that hovering over an instance in the text reveals associated attributes like diagnosis and relationship.

1979.  
tante : PR.  
Mère atteinte de PR.  
RAS.  
GRAND MERE : PR ?.  
Grand mère avec rhumatisme psoriasique.  
RAS.  
Père : IDM (décès).  
PR chez une soeur, mère et grand-mère  
MÈRE DE 92 ANS "polyarthrosid  
ne porteuse d'une PR et d'u  
ostéoporose.  
grand mère polyarthrite rhuma  
cousins germains ont une PR  
as.  
aucun.  
arthrose.  
père:eczéma de contact, lkc..  
père:eczéma de contact  
Covered text = père:eczéma de contact  
Rule identifier = 4CCB98B30AC20C6EC8442E5EF13FF777  
diagnostic = eczéma de contact  
lien\_parenté = père

La nouvelle annotation apparaît dans l'aperçu, avec les différentes séquences textuelles couvertes dans le document testé

En passant la souris sur une instance de l'annotation dans le texte, on peut voir les attributs (diagnostic et lien de parenté) associés à l'annotation

Les sections non couvertes par la règle peuvent suggérer des améliorations de la règle, ou la création de variantes qui produiront la même annotation pour une séquence textuelle différente.

Value  
com.ibm.ICA.medical.fr.Anté

## Déploiement des annotateurs vers ICA

- **Les annotateurs sont déployés vers ICA sous forme d'un fichier .pear qui inclut toutes les ressources nécessaires (dictionnaires, règles, configuration...)**
- **Un assistant permet de sélectionner les annotations à exporter, si nécessaire les attributs de ces annotations, et les facettes que ces annotations vont générer à la suite de l'indexation dans ICA.**
- **Ce même assistant permet de déclencher jusqu'à la réindexation par ICA, ce qui fait qu'aucune intervention d'administration n'est nécessaire dans ICA pour que les utilisateurs bénéficient de ces nouvelles facettes.**