

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

ISO03 - Découvrir les connaissances dans les sources de données textuelles

Conception d'annotateurs sémantiques avec IBM Content Analytics Studio

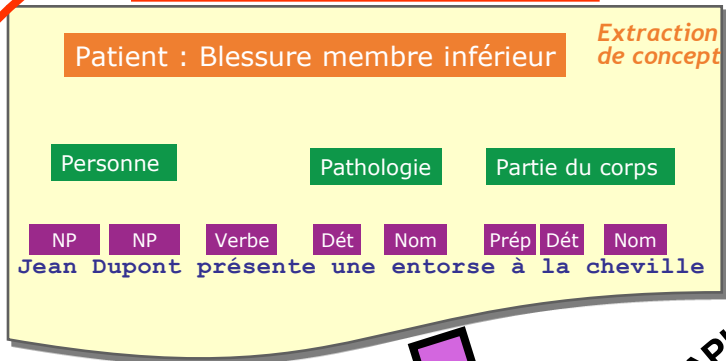
Jean-Marc Langé

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

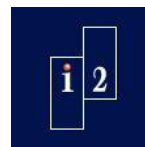
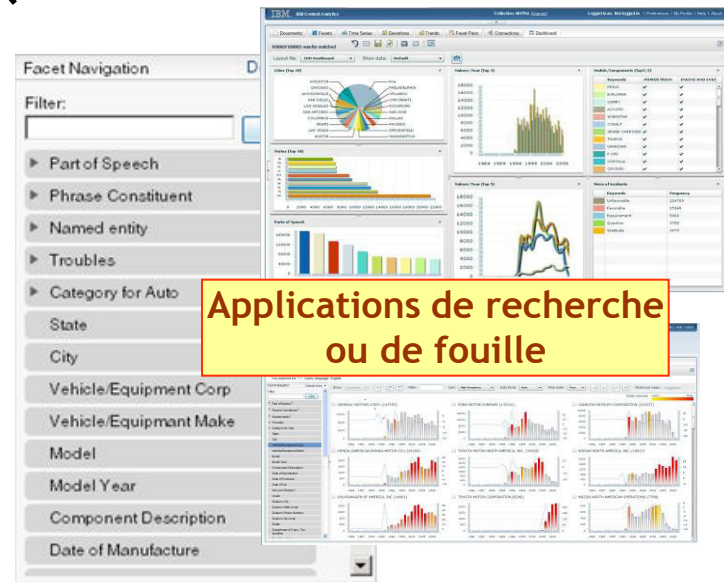
Documents et données analysés
Avec identification de concepts

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:

GRANDMERE 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 Studio Explorer with a tree view containing:

- Annotateurs généraux
- Medical
 - Configuration
 - Documents
 - Resources
 - Break Rules
 - Character Rules
 - Dictionaries
 - Parsing Rules
 - Semantic
 - SNOMED
 - Dictionaries
 - Parsing Rules
 - X_CIM-10
 - Results
- Projet ICPA-PR

The right sidebar shows the Outline view with a list of annotations:

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 a table with columns "Property" and "Value".

Configuration de la chaîne d'annotations

The screenshot displays the UIMA Pipeline Configuration interface. 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 list of 'Dictionaries for language French' such as 'Built in Lex dictionary: fr-XX-LLex-7009.dic' and various SNOMED dictionaries. A blue callout bubble points to the 'Lexical Analysis' stage in the list, and another points to the dictionary 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

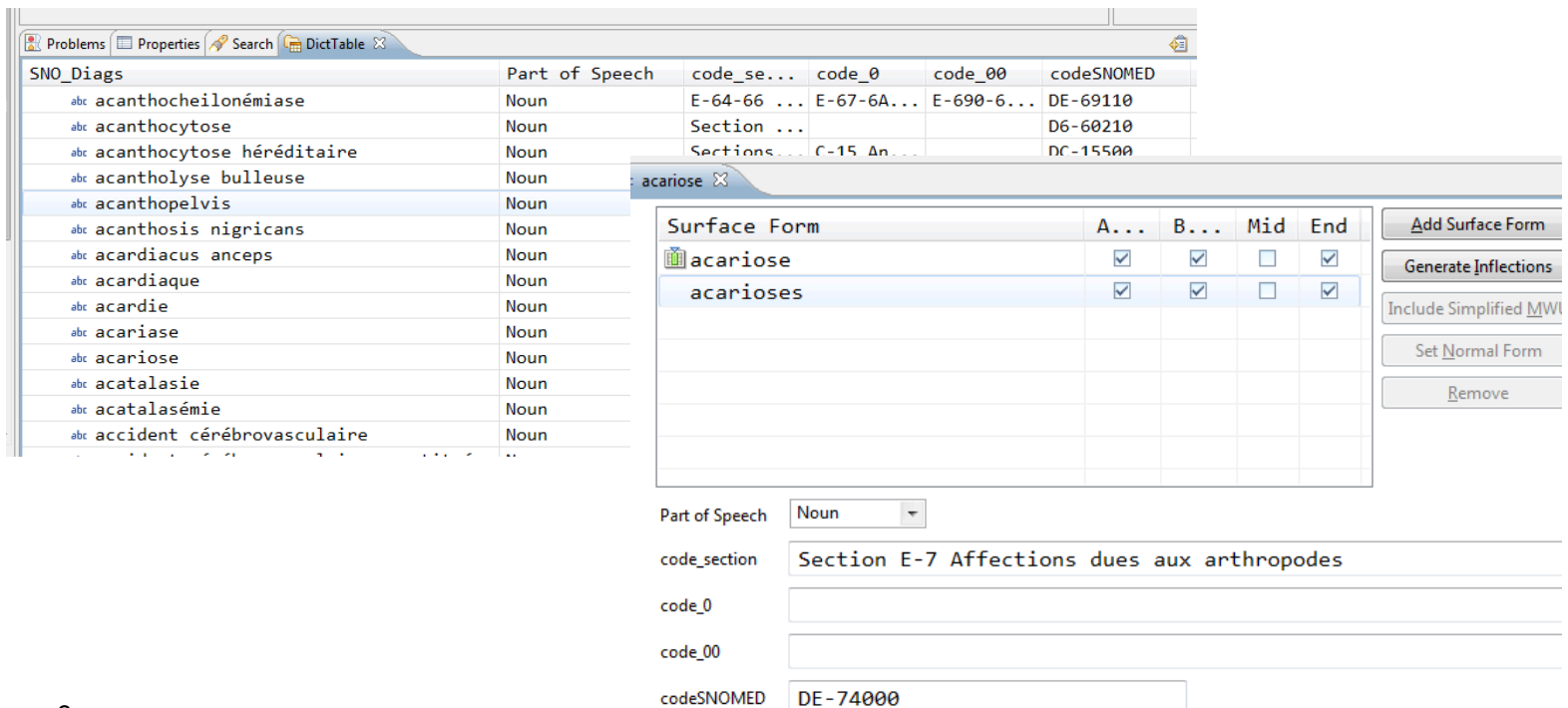
Dictionaries for language French

- 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

The screenshot shows the 'Create Parsing Rules' interface. On the left, a text file named 'antécédents_familiaux.txt' is open, displaying medical history text. The line 'Grand mère avec rhumatisme psoriasique.' is highlighted in blue. On the right, the 'Rule Config' panel shows a rule set named 'Default' with the input text 'Grand mère avec rhumatisme psoriasique'. The rule configuration includes three annotations:

- 1: Famille** (checked):
 - Value = Grand mère [Special]
 - Lemma = grand-mère [Special]
 - Part of Speech = Noun [Special]
 - Length = 10 [Special]
- 2: Token** (checked):
 - 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** (checked):
 - Value = rhumatisme psoriasique [Special]
 - Lemma = rhumatisme psoriasique [Special]
 - Part of Speech = Noun [Special]
 - Length = 22 [Special]

Annotations and callouts:

- drag & drop de la séquence textuelle**: Points to the highlighted text in the file.
- Annotation résultant du dictionnaire de liens familiaux**: Points to the '1: Famille' annotation.
- « mot brut » (sans annotation particulière)**: Points to the '2: Token' annotation.
- Annotation résultant du dictionnaire de diagnostics médicaux**: Points to the '3: Dict_SNO_Diags' annotation.

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

Type:

- 1: Famille
 - Features
 - Value = Grand mère [Special]
 - Lemma = grand-mère [Special]
 - Part of Speech = Noun [Special]
 - Length = 10 [Special]
 - Subtree
- 2: Taken
 - Features
 - Value = rhumatisme psoriasique [Special]
 - Lemma = rhumatisme psoriasique [Special]
 - Part of Speech = Noun [Special]
 - Length = 22 [Special]
- 3:
 - Features
 - Value = rhumatisme psoriasique [Special]
 - Lemma = rhumatisme psoriasique [Special]
 - Part of Speech = Noun [Special]
 - Length = 22 [Special]

Advanced Repeat Options

Advanced Repeat Options

0

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

OK Cancel

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]
 - Features
 - 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 attributes: 'Covered text = père:eczéma de contact', 'Rule identifier = 4CCB98B30AC20C6EC8442E5EF13FF777', 'diagnostic = eczéma de contact', and 'lien_parenté = père'. A blue callout box points to the 'Antécédent_familial' category in the 'Outliner' pane, which lists various medical conditions and relationships like 'NIECE SPA', 'MERE PSORIASIS', 'FRERE SPA', etc. Another blue callout box points to the text 'Grand mère avec rhumatisme psoriasique.' in the document.

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

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.

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

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

Outliner
View: Default View
com.ibm.ICA.medical.D...NO_Topo
com.ibm.ICA.medical.fr.Antéc.patho
com.ibm.ICA.medical.fr.Antécédent_familial
@ NIECE SPA
@ MERE PSORIASIS
@ Filles: arthrite
@ FRERE SPA
@ grand père: rhumatisme
@ FILLE AVEC ASTHME
@ mère: hvnothvro... ostéonorse

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.**

La plate-forme d'échange **Vivastream™**

- Développez votre réseau
- Découvrez les experts sur les sujets qui vous intéressent
- Echangez avec les speakers et les experts
- Regardez qui participe aux sessions auxquelles vous êtes inscrit
- Évaluez les sessions auxquelles vous avez participé



Inscrivez-vous sur le web ou avec votre smartphone

<http://www.vivastream.com/events/ibmimt-solutionsconnect-frfr>

