Acquisition > Centre d'Apprentissage

L'OCR a pour rôle d'importer les informations contenues dans un document à l'intérieur d'une table. En vue de cette importation, l'OCR a besoin de savoir quelle information mettre dans quel champs. Les champs d'extraction servent à guider cette opération.

Ils permettent de détecter la donnée recherchée, de la traiter et de la valider avant récupération et insertion en base de données. Ces trois premières phases étant configurables dans le Centre d'apprentissage, dans cette page, sont décrites les différentes fonctions disponibles.

S'il manque une maquette, cliquez sur le bouton loupe à droite de la liste des maquettes.

Champs d'analyse

Si l'indication [*] est présente devant le nom du champs, cela indique que le champs doit être remplis dans la table (SQL).

Options de base

- Type d'extraction : Voir les types d'extractions
- Description : Nom de la valeur qui ne sert que de repère
- Champ de destination : permet de choisir dans quel champs de la table cible doit être enregistrée la valeur extraite
- Requis : si cette case est cochée et qu'aucune valeur n'est trouvée dans le fichier lors de l'extraction, l'opération sortira en erreur

Options de capture

Les options de captures dépendent du type d'extraction. Voir les types d'extractions

Options de post formattage

Article dédié au post formattage

- Post formattage :

@self	permet d'obtenir une référence sur le champ courant (celui sur lequel on travail) de l'analyse COLD et ainsi d'obtenir sa valeur. **@self** ne s'utilise que dans le cas d'un Post formatage en tant que paramètre d'une macro de formatage.
@<*groupname*>	fait référence à un groupe de capture d'une expression rationnelle, voir sur l'Article dédié au post formattage
@smartdate	-
@smartfloat	-

@getinlist (Assistant Source de données interne)	Permet de récupérer / retourner la valeur d'un champs en indiquant la table ciblée / recherchée et le champs recherché / qui sert à repérer l'enregistrement concerné. En 3.3+, son usage est simplifié par l'assistant “Source de données interne”. En savoir plus sur la syntaxe

- Gestion des espaces :

Supprimer aux extrémités	-
Supprimer à droite	-
Supprimer à gauche	-
Supprimer tous les espaces	**Attention** sur une donnée de type date, il est déconseillé de choisir cette option, cela peut invalider le format de la date et provoquer une erreur d'importation. **Préférez l'option supprimer aux extrémités**

Contrôle de validité

- Filtre de validation : Les filtres permettent de valider que la valeur extraite est conforme à certains critères. Il peut s'agir d'un format de données ou même d'une information située dans une table.

Tous ces filtres sont des regex ou plus communément appelées expressions régulières

- Ignorer la casse : cocher cette case fait que le filtre ignorera la différence entre majuscules et minuscules lors de la validation

Types d'extractions

Dans la section **Options de base**, se trouve la liste des types. Certains types sont des catégories contenant d'autres fonctions. Pour plus de détails sur un type, cliquez sur le lien correspondant.

Types	Exemples de macros	Description
Macros après un texte	@after	Capture après un texte...
Macro entre du texte et/ou des champs balise	@between	Retourne le texte compris entre une chaîne de début et de fin
Position relative à un champ balise	@relativeto	Permet de trouver du texte en se positionnant relativement à la position d'un autre champ du modèle d'analyse.
Fonctions spécifiques	@concat, @dateecheance, ...	
Fonctions d'évaluation	@year, ...	
Macros identité	@nom_passeport, @nom_permis, ...	
Macros Adullact	@adullactchorusengagement, ...	
Macros date	@lastdate, ...	
Macros courriels	@mailfrom, @mailto, ...	
Macros montants	@totalht, ...	
Macros système	@year, @jobid, ...	
Fonctions d'identification Fournisseur	@societynotwww	Cherche le nom de la société (*avec internet* optionnel)

Types	Exemples de macros	Description
Position Variable (Champ balise)		Récupère une valeur en se basant sur une position qui dépend d'un filtre de recherche
Position fixe		Récupère une valeur située à une position constante dans la grille
Position relative à un champ balise		Cherche une valeur en se positionnant par rapport à une balise
Valeur constante spécifiée		Récupère une valeur qui reste constante. Pour en savoir plus sur cette macro, cliquez à gauche

Fonctions d'évaluation

Aller sur l'Article dédié

Aller sur la partie Comment utiliser l'assistant de Fonctions avancées sur le champ ?

Les méthodes listées dans le tableau qui suit, sont soit des macros EzGED, soit des fonctions python.

Vous pouvez également dans la zone **Options de capture** > **Fonction**, appeler vos propres fonctions. **todo**

Extractions d'évaluation	Macro EzGED ou fonction python correspondante	Description
Concaténation d'éléments	@concat(,)	@concat(valeur1,valeur2,...,valeurN)
Date d'échéance	@dateecheance(,30)	@dateecheance(@fldX,jours)
Concaténation d'éléments et conversion en date	@concatdate(,)	@concatdate(valeur1,valeur2,...,valeurN)
Condition OU entre plusieurs champs	@or(,)	Retourne 0 ou 1
Condition évaluée	@condition(,)	Retourne 0 ou 1
Conversion du jour numérique en texte	@int2day(,)	int2day(3) ⇒ mercredi
Conversion du mois texte en mois numérique	@monthint(,)	@monthint('janvier') ⇒ 01
Conversion du mois numérique en texte	@int2month(,)	@int2month(11) ⇒ novembre
Filtrer la chaîne, ne garder que les lettres chiffres et tirets	@filter(,)	
Filtrer la chaîne, ne garder que les numérique	@keepdigits(,)	@keepdigits('abc123') ⇒ 123
Retourne le taux de comparaison entre deux textes en %	@taux_levenstein()	Plus les deux textes se ressemblent, plus le score est élevé

Extractions d'évaluation	**Macro EzGED ou fonction python correspondante**	**Description**
Retourne une comparaison approximative entre deux textes vrai si taux >70%	@cmp_levenstein()	Retourne vrai si les textes sont quasiment équivalents (plus de 70% identiques)
Récupère la chaine avant la première parenthèse ouvrante	@annuler_parenthese()	Exemple : “toto (tata)” ⇒ “toto ”
Récupère la chaine entre parenthèses	@get_parenthese(,)	Exemple : “toto (tata)” ⇒ “tata”
Valeurs alternatives	@alternative()	@alternative(@FLD4,@FLD3,'Non renseignée')
Extractions d'évaluation	**Macro EzGED ou fonction python correspondante**	**Description**
Conversion en entier (cast)	long()	
Conversion en majuscules	.upper()	
Conversion en minsucules	.lower()	
Conversion en nombre à virgule (cast)	float()	
Conversion en string (cast)	str()	
Couper sur séparateur (split)	.split('-')	
Extraction d'une partie de la chaine	[:]	
Nom du fichier d'origine sans le .txt	@filename[-4:]	
Nom du sous-répertoire d'origine (Niveau 1 par défaut)	@dir1	
Rechercher la position d'une chaine xxx depuis la fin	.rfind('xxx')	
Rechercher la position d'une chaine xxx depuis le debut	.find('xxx')	
Remplacer x par y	.replace('x','y')	
Suppression des blancs aux extrémités	.strip()	

Retour vers liste catégories

Macros identité

Extractions par identité	**Macro correspondante**	**Description**
Macros identité: Nom sur un passeport	@nom_passeport	
Macros identité: Nom sur un permis de conduire	@nom_permis	

Extractions par identité	Macro correspondante	Description
Macros identité: Nom sur une carte d'identité étrangère	@nom_ci_etrangere	
Macros identité: Nom sur une carte d'identité française	@nom_carte_identite	

Retour vers liste catégories

Macros Adullact

Extractions Adullact	Macro correspondante	
Macros Adullact: Chorus N° d'engagement	@adullactchorusengagement	Information requise par certaines collectivités
Macros Adullact: Chorus N° de document	@adullactchorusnumpiece X	Numéro de la facture
Macros Adullact: Chorus N° de marché	@adullactchorusmarche	
Macros Adullact: Chorus N° unique de Document	@adullactchorusfactid	Pour manipuler le document en son sein, Chorus génère un identifiant unique et l'affecte au document
Macros Adullact : Booléen : Document envoyé au tiers de télétransmission	@adullactsendtdt	ramène en booléen 0/1
Macros Adullact : Booléen : Le document a été signé	@adullactsigned	ramène en booléen 0/1
Macros Adullact : Booléen : Présence d'un accusé de réception	@addulacthasar	ramène en booléen 0/1
Macros Adullact : Chorus Date de dépôt	@adullactchorusdatedepot	
Macros Adullact : Chorus Date de mise à disposition	@adullactchorusdatemiseadispo	Il arrive, qu'entre son dépot et son accès en visibilité, le document passe par un processus, un autre traitement. D'ou l'existance de cette date.
Macros Adullact : Chorus Date du document	@adullactchorusdatepiece	
Macros Adullact : Chorus Total HT	@adullactchorustotalht	
Macros Adullact : Chorus Total TTC	@adullactchorustotalttc	
Macros Adullact : Chorus Total TVA	@adullactchorustotaltva	
Macros Adullact : Id du fournisseur dans societycache	@adullactsocietyid	
Macros Adullact : Nom du fournisseur	@adullactsociety	
Macros Adullact : Niveau 1 du plan de classement Acte	@adullactclassniv	
Macros Adullact : Niveau 2 du plan de classement Acte	@adullactclassniv	
Macros Adullact : Niveau 3 du plan de classement Acte	@adullactclassniv	
Macros Adullact : Niveau 4 du plan de classement Acte	@adullactclassniv	
Macros Adullact : Niveau 5 du plan de classement Acte	@adullactclassniv	

Extractions Adullact	Macro correspondante	
Macros Adullact : N° de Tva du fournisseur	@adullactsocietynumtva	
Macros Adullact : Objet	@adullactobjet	récupère l'objet (titre du document télétransmis).
Macros Adullact : Siren du fournisseur	@adullactsocietysiren	
Macros Adullact : Siret du fournisseur	@adullactsocietysiret	
Macros Adullact : Type de document Hélios	@adullactheliosnature	indique la nature du document (id de liste spécifique, voir la tables liste des natures)

Retour vers liste catégories

Macros date

Extractions par date	Macro correspondante	Description
Macros date: Date actuelle	_common.tstamp()	
Macros date: Date d'échéance	@lastdate	
Macros date: Date de commande	@firstdate	
Macros date: Date de facture (de document)	@middledate	
Macros date: Date de fichier	@filedate	
Macros date: Première date dans le document	@firstdateindoc	

Retour vers liste catégories

Macros courriels

Extractions par courriels	Macro correspondante	Description
Macros courriels: Copie cachée	@mailcc	
Macros courriels: Date courriel	@maildate	
Macros courriels: Destinataires	@mailto (alias: @maila)	
Macros courriels: Expéditeur	@mailfrom (alias: @mailde)	
Macros courriels: Objet	@mailsubject (alias: @mailtitle, @mailtitre, @mailsujet)	

Retour vers liste catégories

Macros montants

Options de capture / Fonction avancée	Macro correspondante	Description
Total HT	@totalht	Cherche et retourne le total HT.
Total TTC	@totalttc	Cherche et retourne le total TTC.
Total TVA	@totaltva	Cherche et retourne le total TVA.

Retour vers liste catégories

Macros système

Macros système	Macro correspondante	Description
Macros système: Année actuelle	@year	Cherche et retourne l'année actuelle
Macros système: Identifiant de document (mode découpage)	@docid	?
Macros système: Identifiant du travail	@jobid	
Macros système: Nom du fichier d'origine	@filename	
Macros système: Nombre de page total	@nbtotalpage	
Macros système: Numéro de semaine actuel	@week	
Macros système: Séparateur	@separator (alias: @separateur)	Recherche “Code39: EZGED” dans le document
Macros système: Texte contenu dans la page	@page	

Retour vers liste catégories

Macros après un texte

Macro équivalente ancienne version :

```
@after
```

Capture un texte (sur une ligne) se situant après une chaîne de caractère fixée.

Si la chaine de caractères n'est pas trouvée dans le texte alors la macro retourne la chaîne vide.

Option	Description
Après ce texte	Chaine de caractère (pas de regex) après laquelle on va capturer
Nombre de caractères à extraire	Longueur de la chaine à capturer

Macro entre du texte et/ou des champs balise

Macro équivalente ancienne version :

```
@between
```

Retourne le texte compris entre un marqueur de début et de fin.

Cette extraction se base sur deux repères : un pour le début et l'autre pour la fin.

Pour chaque repère, vous pouvez attribuer une option à la fois. Celles-ci sont visibles dans l'interface et décrites ci-dessous :

Option	Description
Texte de début ou de fin	Texte qui s'il est détecté et trouvé va servir de repère à l'extraction
Balise de début ou de fin	La balise dans le fichier servant de repère si elle est trouvée.

Dans les champs Textes, vous pouvez mettre 2 types de valeurs :

- Une chaîne de caractère permettant de trouver le marqueur
- Une référence à un champ de l'analyse (@FLD*xxx*).

Fonctions d'identification Fournisseur

Macros équivalentes ancienne version :

```
@societynotwww
@society
@societyeu
```

Retourne le nom de la société ou “inconnu” si ne la trouve pas.

La première macro utilisée par défaut est **@societynotwww**, elle cherche un numéro de SIREN.

ATTENTION EzGED a besoin d'un accès internet pour utiliser la recherche à échelle Européenne et la recherche via internet car la recherche se fait dans des bases externes.

Si le serveur n'est pas joignable, vous verrez “NULL” dans la colonne correspondante.

ATTENTION La recherche à échelle européenne n'est pas disponible en version ONE.

Option	**Description**
Etendre la recherche à l'Europe	**@societyeu** cherche dans le document à analyser un **numéro de TVA** Intracommunautaire et retourne le nom de la société correspondante. Pour plus de détails, voir la partie ci-après “Recherche par Numéro de TVA“
Activer la recherche à internet	Enclenche l'utilisation de **@society** à la place de **@societynotwww** et cherche en utilisant la base + internet
Liste des SIREN ignorés	à séparer par des virgules

Recherche de la société par Numéro de TVA

ATTENTION cette macro n'est pas disponible en version ONE.

Utilise la macro **@societyeu** qui cherche une société à l'échelle européenne. Nécessite internet pour fonctionner.

La macro cherche par ordre de préférence les numéro de TVA dans cet ordre par défaut,

“FR,BE,IT,GB,LU,DK,EL,IE,NL,AT,PT,FI,SE,HU,PL,CY,CZ,EE,LV,SI,RO,DE,ES,SK,BG,HR,MT,LT”

puis interroge un web service de la base européenne : Pour l'instant les bases des pays suivants ne sont pas interrogeables : DE,ES,SK,BG,HR,MT,LT c'est pour cela qu'ils sont en dernier dans l'ordre de recherche par défaut, toutefois vous pouvez utiliser la base de cache pour les faire fonctionner, pour cela modifier le champs societycache_code en Varchar de 20.

Vous pouvez changer l'ordre de recherche ou le limiter à certains pays uniquement en ajoutant la variable suivante dans instance.conf.

Exemple : ici, limitation aux pays limitrophes à la France avec francophones de préférence + Angleterre et Irlande:

[ezged]

```
mytvacountries = FR,BE,LU,IT,GB,IE,DE,ES
```

Position Variable (Champ balise)

Cette extraction se base avant tout sur une information ou un format de données, que vous pouvez choisir dans l'assistant dédié de l'option “**Filtre de recherche**”, pour trouver la valeur voulue.

Les options de positions servent simplement à préciser le rayon de recherche dans la page.

Option	**Description**
Position de départ X	La colonne sur laquelle doit commencer la recherche
Position de fin X	La colonne sur laquelle doit stop la recherche
Position de départ Y	La rangée sur laquelle doit commencer la recherche
Position de fin Y	La rangée sur laquelle doit stop la recherche
Filtre de recherche	Le filtre en question apparait sous forme d'une regex ou expression régulière. Vous pouvez écrire directement dans le champs texte si vous préferez utiliser une regex personnalisée.
Ignorer la casse pour l'expression régulière	Ignorera la différence entre majuscules et minuscules

Il est possible d'utiliser des positions négatives. La position sera alors considérée en partant de la dernière colonne ou rangée.

Exemple: Pour -2 en position de fin X

La recherche se fera jusqu'à l'avant-avant-dernière colonne. Donc à -2 en partant de la fin.

Position fixe

Cette macro permet de récupérer une valeur qui se trouve toujours au même endroit quel que soit le fichier passé dans l'OCR.

Les options doivent obligatoirement être réglés sur **des chiffres entiers réels**.

Option	**Description**
Position de départ X	La colonne sur laquelle doit commencer la recherche
Position de départ Y	La rangée sur laquelle doit commencer la recherche
Longueur d'extraction	Chiffre entier, longueur de la valeur extraite. Si vous en avez plusieurs et de longueurs différentes, mettez la longueur de la plus longue valeur et gérez la *suppression des blancs* dans les **options de Post-Formatage**

Position relative à un champ balise

Permet de trouver du texte en se positionnant relativement à la position d'un autre **champ du modèle d'analyse**.

Option	Description
Champs balise	**Obligatoire** pour que ça fonctionne bien. Liste des champs qui peuvent servir de repère à la détection
Position de départ X	La colonne sur laquelle doit commencer la recherche
Position de départ Y	La rangée sur laquelle doit commencer la recherche
Position de fin Y	La colonne sur laquelle doit stop la recherche
Longueur d'extraction	Chiffre entier, longueur de la valeur extraite. Si vous en avez plusieurs et de longueurs différentes, mettez la longueur de la plus longue valeur et gérez la *suppression des blancs* dans les **options de Post-Formatage**

La macro s'arrête à la première ligne de texte non vide rencontrée au sein de la zone de recherche cible (déterminée par les positions de démarrage et de fin). La macro ne permet donc pas de rechercher un résultat sur plusieurs lignes.

Valeur constante spécifiée

Permet d'ajouter une valeur constamment identique à l'extraction, en plus de toutes celles récupérées.
Par exemple, une information qui ne se trouverait pas dans les fichiers mais reste constante et dont vous avez besoin dans la table destinatrice.

Macros Adullact : Type de document Hélios

Retourne **2** s'il s'agit d'un document reçu donc correspondant à une dépense

OU

1 si c'est un document emis

From:
http://wiki.ezdev.fr/ - **EzGED Wiki**

Permanent link:
http://wiki.ezdev.fr/doku.php?id=doc:v3:acquisition:apprentissage&rev=1528272347

Last update: **2023/03/17 09:56**