

# Extraire l'intelligence des documents

- Extrait l'intelligence à partir de contenus 'plats' et/ou non-structurés
- Transforme et valorise votre contenu existant
- Génère des métadonnées 'métier'
- Réutilise votre contenu historique en le convertissant dans un format XML structuré
- Offre une granularité à géométrie variable dans la structure de votre contenu
- Offre des fonctionnalités de validation des schémas DTD/XML
- Offre des capacités de classification, d'indexation et de recherche poussées
- Automatise la structuration de votre contenu
- Réduit vos coûts et améliore la productivité de vos processus de conversion XML
- Traite tout type de documents (presse quotidienne, magazines, livres, documents légaux, publications scientifiques, etc.)
- Traite tout type de formats (Word, Quark Xpress, PDF, HTML, papier, n'importe quel format OCR, SGML, XML, etc.)
- Fournit des données en sortie qui s'intègrent directement dans n'importe quel type de base de données

## Des clients qui nous font confiance ....

- Lexis Nexis
- Wolters Kluwer
- Groupe Lagardère
- Lexbase
- Les Editions Francis Lefebvre

## Khemeia™ – Définition

La solution Khemeia™ est une technologie développée par Stelae Technologies qui permet d'extraire et de structurer l'« intelligence » du contenu des documents. Elle est utilisée pour revaloriser un contenu qui doit être reconstruit ou transformé:

Khemeia™ analyse, extrait et balise les données « plates » ou non structurées d'un document en fonction de sa structure, de son style et du positionnement de son contenu.

Khemeia™ permet aussi d'industrialiser ces mécanismes de conversion - tâches généralement coûteuses le plus souvent exécutés manuellement par des sous-traitants - afin d'en améliorer la productivité aussi bien lors d'une première numérisation que lors d'une extraction des métadonnées de contenus existants.

Khemeia™ gère plusieurs types de formats en entrée comme en sortie. L'extraction de l'intelligence génère un jeu complet de métadonnées, qui sont convertis en fichier XML, HTML ou PDF afin d'être publiés en ligne ou d'alimenter une base de données.

## A ne pas confondre avec ....

... les plug-ins pour Microsoft Word ou Adobe Acrobat, ou même avec des solutions de comparaison sémantique. Khemeia™ est doté d'une technologie révolutionnaire d'extraction d'information à partir de documents « plats » et représente une classe de solution logicielle à part entière.

## Usages

### Applications

Voici la liste des tâches de transformation et de traitement des contenus numériques automatisées par Khemeia™:

- Conversion des contenus historiques en XML
- Génération des métadonnées décrivant et définissant ces contenus
- Fourniture de contenus indexés et recherchables - l'utilisateur tire le meilleur parti des données
- Production d'information structurée et enrichie de styles, pour des sorties en PDF
- Génération de styles HTML complets pour publier les contenus via le Web
- Fourniture de sorties XML pour les moteurs de recherche des métadonnées
- Reconstruction des structures XML historiques qui doivent être analysées/validées pour répondre à un schéma DTD/XML nouveau ou standardisé

## Formats traités

### Formats d'entrée

- Word, HTML, PDF, Quark Xpress, Adobe InDesign, tout format OCR (Reconnaissance Optique de Caractères) XML, SGML, etc.

### Formats de sortie

- XML, DTD, PDF, HTML.
- MySQL, SQL Server, Oracle, GED, etc.

---

## Fonctionnement

Le principe de fonctionnement de Khemeia™ s'appuie sur la détection de la structure et des styles repérés par l'œil lors de la lecture d'une page.

Le postulat de départ est que les êtres humains créent des documents en se fondant principalement sur une logique visuelle, même s'ils utilisent les styles ou les feuilles de style. Khemeia™ arrive à interpréter cette logique visuelle.

Pour chaque type de publication, les règles sont configurées en fonction des exigences précises de la structure des contenus. Pendant le traitement, une segmentation automatisée du fichier d'entrée est réalisée qui est ensuite affinée en fonction des exigences des métadonnées spécifiées par le client.

Voici quelques exemples des fonctionnalités offertes par la technologie de Khemeia™ :

- Identification des numéros de page, de chapitre, titres, en-tête, notes, notes de bas de page, notes de fin, références « voir », listes à points, listes à numéros, images, légendes des images, etc.
  - Définition d'entités homogènes à partir de :
    - leur position dans le contenu
    - leur police et taille de caractères
    - leurs coordonnées sur la page
  - Gestion des tableaux
  - Identification des catégories particulières du contenu (à partir desquels les métadonnées sont générées)
  - Identification des métadonnées dans les documents
- .... Avec, en sortie, ces données structurées par exemple en XML, PDF ou HTML.

### Exemple - Magazines et journaux

- Dans le cas d'un magazine constitué de plusieurs pages PDF, Khemeia™ segmente les données par article en reconnaissant le début et la fin de chacun d'eux. Ensuite, chaque article est encore segmenté de manière plus fine en titre, sous-titre, auteur, date, etc. Cette segmentation reconnaît aussi les images, les légendes, les crédits photographiques et les encadrés par exemple.

### Exemple - Livres

- Un livre contient plusieurs chapitres. Khemeia™ segmente les données par chapitre, en reconnaissant le début et la fin de chacun d'eux. Ensuite, chaque chapitre est segmenté de manière plus fine en titre, paragraphes, numéros de pages, images, légendes, etc.

## Exemple – documents juridiques

- Un document juridique en PDF peut contenir une centaine de cas de jurisprudence. Khemeia™ segmente ce document cas par cas, puis chaque cas de manière plus détaillée avec le nom de la cour, le numéro du cas, la juridiction, le nom du président du tribunal, le nom du plaignant/partie civile, du prévenu/défenseur, du jugement, etc.

## Exemple – Extraits d'une revue scientifique

- Une revue scientifique contient plusieurs articles. Khemeia™ les segmente et à l'intérieur de chacun d'eux, identifie l'information nécessaire à la constitution des extraits : titres, auteur, éditeur, date, sujet, classification, édition, extrait, identifiant, format, étendue, langue, disponibilité, lieu, source, champ d'application, droits, conditions d'utilisation, citations, etc.

---

## Quels utilisateurs pour Khemeia™ ?

Toute entreprise désireuse :

- d'extraire la structure de ses contenus et générer des métadonnées
- de réduire les coûts d'enrichissement de ses contenus numériques

### Editeurs et fournisseurs d'information

Voici quelques types d'entreprises qui peuvent rapidement profiter des fonctions de Khemeia™

- Editeurs de publications juridiques
- Editeurs de magazines
- Editeurs de journaux
- Editeurs de livres
- Editeurs de revues scientifiques, techniques, médicales, etc.
- Editeurs de matériels pédagogiques pour les solutions d'e-learning
- Secteur public (publications, décrets, etc.)
- Agrégateurs de contenus en ligne (portails, moteurs de recherche, etc.)

### Sociétés de services de numérisation

Les sociétés de services – y compris les entreprises d'externalisation – désireuses de répondre au plus près aux besoins de leurs clients :

- Entreprises d'externalisation des processus métier (BPO – Business Process Organisation)
- Entreprises d'externalisation des processus de gestion la connaissance (KPO– Knowledge Process Organisation)
- Entreprises d'externalisation des processus juridiques (LPO– Legal Process Organisation)

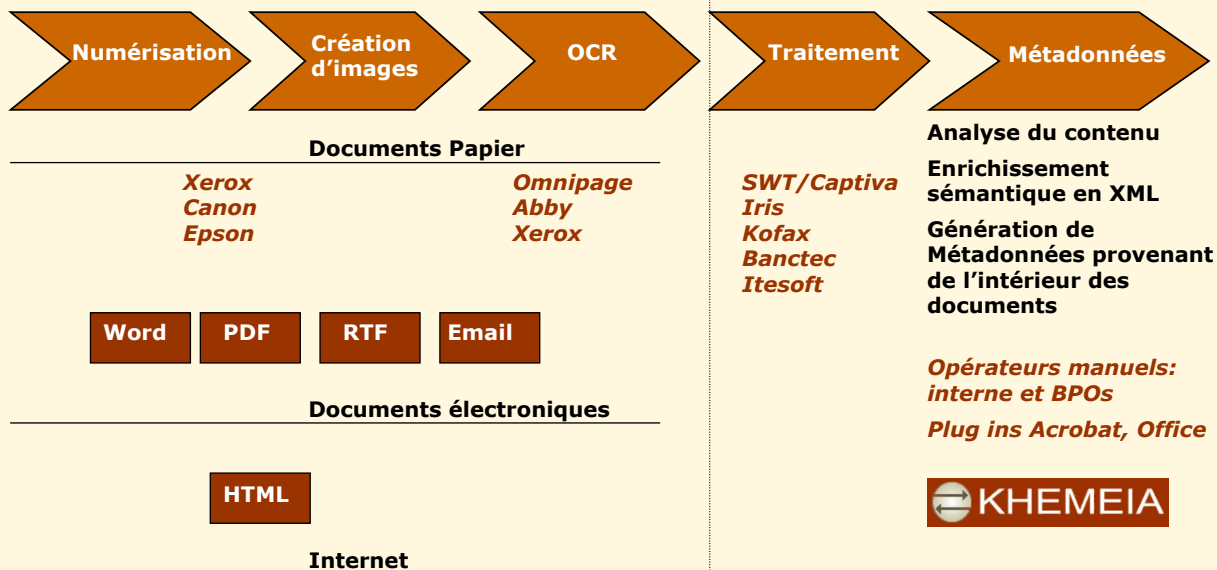
### Solutions d'archivage

La récupération de l'information souhaitée au sein des données archivées peut être une tâche complexe pour une entreprise.

Khemeia™ ajoute de l'intelligence aux fichiers électroniques non structurés (MS Office, Adobe PDF, etc.) qui pourront être indexés et consultés de manière plus fine et plus efficace par les utilisateurs lors de leurs recherches d'informations historiques dans leurs archives.

# Gestion de l'information: Paysage

Information Structurée, Semi Structurée, non structurée



Applications Professionnelles: ECM, EDM, Workflow, Moteurs de Recherche, Bases de données, XML Repositories

Documentum, Exalead, Filenet/IBM, Hummingbird, Ever, W4, Gamma-site, Oracle, Xyleme, Tamino...

## L'avantage concurrentiel de Khemeia™

Lorsqu'il s'agit de documents dont la structure peut être complexe, comme certaines publications scientifiques, techniques ou médicales, les solutions existantes exigent souvent une main d'œuvre importante et un coût unitaire souvent rédhibitoire qui empêchent les éditeurs de contenu de capitaliser au mieux sur l'information qu'ils génèrent.

Khemeia™ apporte une plus grande souplesse dans la structuration, et une automatisation plus efficace des processus permettant ainsi aux clients de Stelae Technologies :

1. d'augmenter la productivité dans l'extraction et la structuration de leurs contenus
2. et/ou d'augmenter la qualité et la finesse de cette structuration
3. et/ou investir plus de temps sur des tâches à plus grande valeur ajoutée (interprétation du contenu, indexation métier...)

## Les bénéfices : gains de productivité, de coûts et de qualité

### Gains de productivité pour l'éditeur de contenu (structuration interne)

Les fonctions d'automatisation de Khemeia™ rendent les tâches extrêmement efficaces, et ouvre des nouvelles perspectives à la réutilisation et à la valorisation des contenus.

Khemeia™ améliore de manière substantielle la productivité en permettant notamment:

- d'industrialiser le traitement de l'information avec une plus grande qualité ;
- d'automatiser la quasi-totalité du travail réalisé d'habitude manuellement ;
- de réduire le temps de contrôle qualité effectué généralement manuellement ;
- d'enrichir et de faciliter l'accès aux contenus (fonctions de métadonnées, d'indexation et de recherches par l'utilisateur).

De manière générale, Khemeia™ permet d'obtenir des gains de productivité et des économies de coûts de numérisation importants sur des contenus complexes : ces gains peuvent représenter jusqu'à 70% d'un workflow manuel.

### Gains de productivité pour une société de service de numérisation (externalisation)

Avec Khemeia™ l'externalisation de la numérisation de vos contenus devient une option très intéressante. En effet, en utilisant Khemeia™, votre prestataire peut bénéficier des avantages suivants :

- Conversion plus productive, et réduction des taux d'erreurs
- Gestion plus facile des contenus complexes
- Comparaison visuelle des résultats, original et XML généré
- Les ressources externes ne sont plus nécessaires pour vérifier les contenus ligne par ligne
- Le traitement des langues étrangères est facilité

Les gains de productivité peuvent être considérables. A titre d'exemple l'utilisation de Khemeia™ a permis à un de nos clients de réduire le temps de traitement par page de 2 heures à 6 minutes pour une même qualité en sortie.

---

## L'interfaçage avec votre système d'information

Khemeia™ ne perturbe pas le fonctionnement ni l'organisation historique des processus existants de gestion de contenu dans l'entreprise : c'est une solution indépendante, dite « stand alone », qui fonctionne sur la base d'un système de dossiers « surveillés » en entrée et en sortie.

Les fichiers d'entrée sont détectés au moment où ils arrivent dans le dossier entrée, et une fois traités, les fichiers de sortie sont placés dans le dossier sortie – et à partir de ce dossier, ils sont intégrés à votre système. Tout nouveau fichier de sortie peut être notifié de manière dynamique à votre système de gestion de contenu.

Khemeia™ peut alimenter et s'interfacer avec tout système d'information : des systèmes de gestion de contenu, des entrepôts XML, systèmes éditoriaux, des sites web, des bases de données (SQL Server, Oracle, MySQL), ou bien des systèmes de gestion documentaire.

---

## Configurations système requise

- Windows 2000, 2003 Server ou XP Professional
- 1 GO RAM minimum
- Résolution écran : 1024 x 768 ou plus
- Lecteur CD-ROM
- Configuration Disque dur recommandée :
  - Partition C: installation de Windows (4 GO)
  - Partition D: installation de Khemeia™ (150 MO)
  - Partition E (exemple): dossiers partagés (1 GO) :
    - Configuration
    - Dossiers d'entrée pour les documents source
    - Dossiers de sortie
    - Dossiers d'avancement pour les tâches non partagées
    - Configuré pour le partage de fichiers sur les réseaux MAC/PC

---

## Questions fréquemment posées (FAQ)

### ***Comment savoir si Khemeia™ convient à mes besoins ?***

Nous pouvons, à partir d'un échantillon de votre contenu, évaluer s'il est adapté à un traitement par Khemeia™ : nous lançons un test sur vos données et vérifions ensuite le résultat conformément à un cahier des charges que nous définissons ensemble.

### ***Et si j'utilise déjà un autre workflow ?***

Il est fort probable que votre workflow actuel soit très consommateur de ressources, ou qu'il n'offre pas les mêmes fonctions et bénéfices que Khemeia™. L'intégration de notre produit, en suivant des tests de viabilité, est habituellement sans douleur et ne peut qu'enrichir le traitement de l'information de votre système actuel.

### ***Et si je sous-traite actuellement ce travail ?***

Comme vous-même, il y a de fortes chances que votre sous-traitant apprécie les gains de productivité que lui fera réaliser l'intégration de Khemeia™.

### ***Qui doit faire fonctionner le logiciel ? Notre entreprise ou notre sous-traitant ?***

Les deux options sont intéressantes. Option 1 : vous utilisez le logiciel et votre sous-traitant fait le contrôle qualité du résultat. Option 2 : votre sous-traitant exécute l'intégralité du travail en utilisant le logiciel pour votre compte et en réalisant le contrôle qualité.

### ***Où s'insère Khemeia™ par rapport à l'OCR (reconnaissance optique de caractères) ?***

Khemeia™ prend le relais dans le processus de numérisation APRES L'OCR. Ce dernier prend son sens lorsque la source de votre information est sur papier

ou sur des images et des textes numérisés : il permet de convertir le document source en fichier numérique. A partir de cette étape, Khemeia™ prend le relais pour structurer le document sortant de l'OCR.

### ***J'ai compris que le logiciel traite des contenus de formats simples, mais Khemeia™ peut-il vraiment traiter les formats complexes d'un magazine par exemple ?***

Absolument. Les collaborateurs de Stelae ont auparavant travaillé dans l'édition de magazines et le logiciel a été initialement développé précisément pour ce type de contenu.

### ***Nous possédons notre propre taxonomie/thesaurus, et désirons l'utiliser ?***

C'est tout à fait possible. Khemeia™ permet la validation de termes spécifiques au sein d'une taxonomie, d'un dictionnaire ou d'une liste de définitions et d'intégrer les règles métier associées lors de la structuration des documents.

### ***Et si je possède déjà un système de gestion de contenus ou de documents ?***

A la différence d'un outil de gestion documentaire qui indexe des documents en fonction de critères variables, Khemeia™ rentre DANS le document et en extrait le contenu en fonction d'une structure XML bien définie.

Dans le cas d'un système de gestion de contenu, Khemeia™ pourra en alimenter la base avec du contenu extrêmement granulaire qui va constituer la matière première des tâches d'édition futures.

---

## En savoir plus

Contactez un de nos bureaux ci-dessous pour en savoir plus.

### France

Stelae Technologies, 19 rue de l'Echiquier, 75010 Paris, France

Tel: + 33 1 44 79 38 02

### UK

Stelae Technologies, Riverbank House, 1 Putney Bridge Approach, London, SW6 3JD, UK

Tel: +44 20 7736 2014

### India

Stelae Technologies, 204, Tower B, GBP, Gurgaon (Haryana), India

Tel: +91 92444 15 905

E-mail: [info@stelae-technologies.com](mailto:info@stelae-technologies.com)

Web: [www.stelae-technologies.com](http://www.stelae-technologies.com)