

Stelae
Technologies

NOW WITH SUPPORT FOR
XBRL & iXBRL

Extracting the intelligence from content

- Extract intelligence from “flat” or seemingly unstructured data with Khemeia
- Transform and “enrich” your content
- Generate “deep” and rich metadata
- Re-purpose your legacy content converting it into XML
- Output as XML, PDF, HTML, XMP, XBRL, iXBRL, S1000D, NewsML, NITF, EPUB, other e-book formats, etc.
- Configurable levels of granularity
- Full DTD/XML schema validation
- Provide enhanced classification, index and search features for your users
- Automate work otherwise carried out manually by hand
- Reduce costs and increase productivity of the conversion process
- Digitize and transform data that you previously thought to be unviable
- Handles magazines & newspapers, books, legal documents, STM publications, financial accounts
- Input formats include PDF, Word, QuarkXPress, HTML, text, paper, any OCR format, SGML, XML, etc.
- Link directly to your CMS or database

Khemeia customers include

- LexisNexis
- Wolters Kluwer
- Lagardère
- Lexbase
- Justis

Khemeia – what is it?

Stelae’s Khemeia is a high-productivity tool for extracting the intelligence from documents.

What appears as “flat” or seemingly unstructured data, is normally for Khemeia a mine of intelligence - intelligence in terms of document structure, style information and metadata.

Used when content needs to be digitized, re-purposed or transformed in order to enhance the value of your content, Khemeia analyses this apparently unstructured content, extracts the intelligence from it - generating full metadata - and converts it into valid XML, HTML, PDF or XBRL to be published electronically.

Yes, finally there is a product, Khemeia, that redefines the mechanics of this whole process of converting and adding value to content - offering new opportunities in terms of transforming and enriching content with metadata.

The boundaries of what constitutes data viable for transformation, have now suddenly been extended with our unique technology.

Whether for first time digitization or extracting structure and metadata from existing content, Khemeia is the ideal tool of this job.

Not to be confused with

Khemeia should not be confused with available plug-ins for Word for Acrobat or even semantic matching solutions products.

It is a product that is in a class of its own offering revolutionary technology for extracting information from “flat” documents.

What is it used for?

This digitization and transformation of content - which is often either done by hand by outsourcing companies, or in many cases not even carried out at all due to the prohibitive costs involved - is exactly what the software automates.

Applications

Applications transforming and processing your digital content include:

- Converting your legacy content into XML
- Generating the metadata which describes and defines such content
- Providing indexed and searchable content – giving the end user a much richer experience from the data
- Producing structured and styled information which can be output as PDF (with optional XMP metadata stream)
- Generating fully styled HTML (i.e. CSS based) to deliver content via the web

- Providing XML output for metadata harvesters
- Re-purposing legacy XML that needs to be parsed/validated against a new or standardized DTD/XML schema

Formats

Input formats

Khemeia is one solution for multiple input formats of document which include:

- PDF, Word, RTF, HTML, Excel, text, QuarkXPress, Adobe InDesign, OCR (optical character recognition) formats, XML, SGML, paper

Multiple output formats

Typical output formats include:

- XML, PDF, HTML, JPEG, XMP, XBRL, iXBRL, S1000D, NewsML, NITF, EPUB and other e-book formats, customer-specific DTDs, etc.
- SQL Server, MySQL, Oracle, CMS systems (e.g. Documentum), etc.

How does Khemeia work?

The software works on the principle of detecting structure and patterns that the eye sees when viewing a page - humans create documents primarily using visual logic (even if they do not use styles/style sheets). Khemeia interprets this visual logic.

For each publication type, rules are configured according to the precise requirements of the data. During processing automated segmentation of the input file takes place (for example into articles for a magazine or newspaper), plus further segmentation according to the metadata requirement specified by the customer (i.e. an article is then further divided).

Typical items identified by Khemeia include:

- Page numbers, section numbers, titles, headings, notes, footnotes, end notes, “see” references, bullet points, number lists, images, captions, etc.
- Styles on the basis of their:
 - o Position
 - o Fonts and their formatting
 - o Co-ordinates on the page
- Tables
- Specific categories of term (from which metadata is generated)
- Metadata for defined elements

.... with the output transformed into, for example, XML, PDF or HTML.

Features of Khemeia include:

- Comprehensive metadata generation without any limitation – the user interface allows the creation of as many tags/element types as required
- Hierarchical and interlinked metadata - not simply flat structured XML
- Outputs: valid XML parsed against customer DTDs/XML schema, HTML, any database or search format
- Benchmarked processing capacity at 2 to 10 million characters per hour (subject to page complexity, images, etc)

Example - Magazines and Newspapers

A magazine or newspaper consisting of multiple PDF pages; Khemeia segments the data by article, recognizing the beginning and end of each, and then within the article segments further, for example, into title, sub-title, author, date, etc. – this segmentation even handles images, captions, photo credits and text boxes. Content can be output as valid XML parsed

against a specific DTD/schema including NITF and NewsML, or in a specific e-book format.

Example - Books

A book consisting of multiple chapters; Khemeia segments the data by chapter, recognizing the beginning and end of each, and then within the chapter segments further into title, paragraphs, page numbers, footnotes, images, captions, etc.

Example - Legal Documents

Legal documents consisting of for example a PDF with a hundred case documents; Khemeia segments the data case by case, and then within each case segments further into name of the court, case number, date, jurisdiction, presiding judge, plaintiff/appellant, defendant/respondent, judgment, etc.

Example – Scientific Journal Abstracts

A scientific journal consisting of a number of articles; Khemeia segments the data into articles, and then within each article identifies the required information for the abstract, for example, title, author, publisher, date, subject, classification, description, edition, abstract, identifier, format, extent, language, availability, location, source, coverage, rights, terms of use, citation, etc.

Financial Accounts

Khemeia provides an automated batch process for company accounts converting PDFs or Excel files to XBRL or iXBRL; the resulting output is validated against a specific schema, for example, US or UK GAAP. Heuristic dictionary functionality allows the user to match unrecognized labels to the taxonomy.

What is wrong with current workflows?

Where Khemeia is not being utilized, current workflows for enhancing digitized content typically have fundamental drawbacks:

- Such methods are either limited in their scope for automation
- Or rely upon large amounts of manual work

In the case of more challenging content for example certain specialist scientific, technical and medical publications, or for information with complex structural requirements, the transformation of such data can often be viewed as at best extremely labor-intensive, or at worst quite simply as not being viable. Foreign language content can also create an additional overhead.

For this reason, such work, where undertaken, is very often outsourced to companies located in countries with lower labor costs.

Who should be using Khemeia?

Any organization wishing to:

- Extract structure and generate metadata from their content
- Reduce the costs of enhancing their digital content

.... including publishing, information management and ICT professionals.

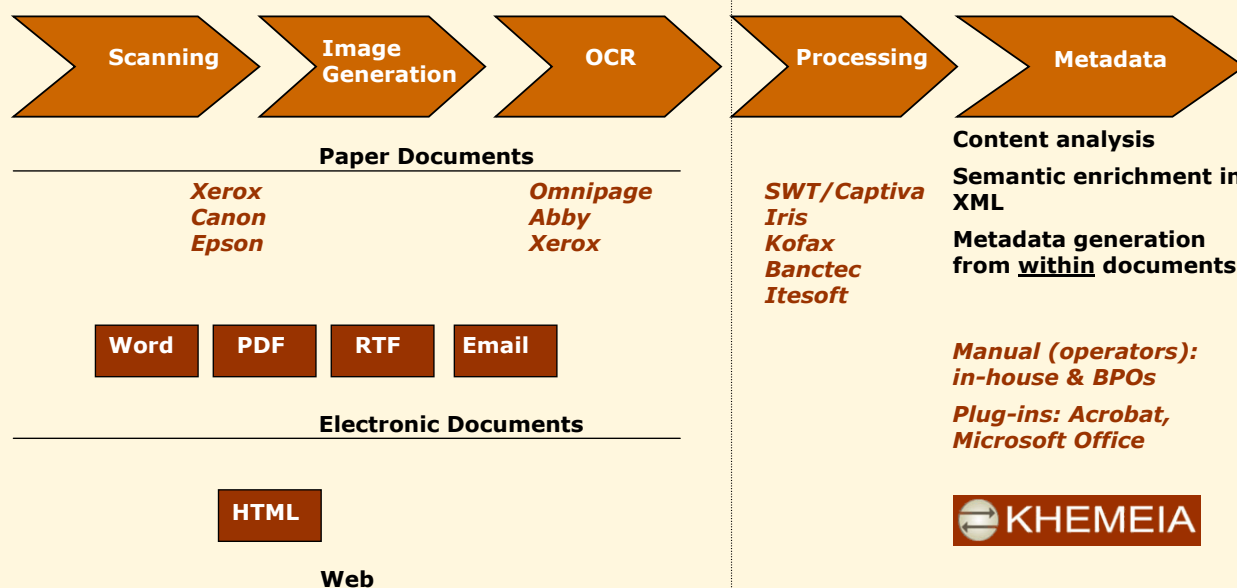
Publishers and information providers

Some of the categories of company that will benefit from using Khemeia include:

- Legal publishers
- Magazine publishers
- Newspaper companies
- Book publishers
- Organizations processing company report or financial information
- Journal publishers (scientific, technical, medical)

Information Management: Landscape

Structured, semi-structured, unstructured information



Enterprise Applications: ECM, EDM, Workflow, Search Engines, Databases, XML Repositories

Documentum, Exalead, Filenet/IBM, Hummingbird, Ever, W4, Oracle, Tamino ...

- Aerospace and defense technical documentation - in the case of legacy content, for example, converting PDF to S1000D data modules
- Publishers of e-learning materials
- Government agencies
- Content aggregators (e.g. portals, and online search engines)
- Etc.

Service organizations

Service organizations including outsourcing companies who wish to fulfil better the needs of their clients:

- BPO (business process outsourcing) companies
- KPO (knowledge process outsourcing) companies
- LPO (legal process outsourcing) companies

Archiving solutions

For any organization archiving data, retrieving the desired information from the stored data can present substantial challenges.

With KHEMEIA intelligence can be added to unstructured electronic files (such as Microsoft Office, Adobe PDF, and others) allowing them to be indexed and searched intelligently by the user when retrieving historic information from the archive.

Interfacing with your current information system

KHEMEIA is designed to feed your existing information system with content, and can, as required, interface with content management systems, XML repositories, editorial systems, web sites, databases (SQL Server, Oracle, MySQL), document management systems, etc.

The software does not impact the functioning of your existing content system, as KHEMEIA works separately on the basis of In and Out "watched" folders. Input files are detected when they arrive within the In Folder, and, once processed, output files are placed in the Out Folder – from here the files are then picked up for integration into your own system - the presence of any newly-arrived output files can be dynamically notified to your system.

Benefits – productivity gains, cost savings and added value

Efficiency gains

KHEMEIA opens up new possibilities in terms of the viability of digitizing and enhancing content, and where the work has up to now been done manually, the process is altered into one of supreme efficiency.

The use of our software leads to substantial productivity gains enabling you to:

- Process greater quantities of information at a higher level of quality
- Automate much of the work normally done by hand
- Lower the amount of manual quality assurance work required – the workflow changes to one where manual input is only needed in the last stage for quality assurance and rectification work
- Add more value to the content (in terms of metadata, indexing and search capabilities for the user)
- Reduce information processing costs - productivity gains and savings in the cost of digitizing content are substantial, with financial savings on a manual based workflow of up to 70%

Outsourcing

Khemeia makes the option of outsourcing your content digitization a more attractive option, with benefits that include:

- More productive conversion and reduced error rates
- Complex content is much easier to handle
- Functionality for visually comparing results - original versus generated XML
- Outsourcing personnel no longer need to check content line by line
- Processing of foreign languages is made easier

Productivity gains can be enormous – for example, one Khemeia customer has decreased page processing times from 2 hours to six minutes per page.

The end users of your content

The all-important end users of the data benefit from the utilization of Khemeia. The outcome is that you the information provider are able to offer your customers a much richer experience with the content that you supply in terms of

presentation, search, indexing and metadata capabilities enabling users to can locate the information that they need.

System requirement

- Windows 2000, 2003 Server or XP Professional
- 1 GB RAM minimum
- Screen resolution: 1024 x 768 or higher
- CD-ROM drive
- Recommended disk configuration:
 - Partition C : installation of Windows program (4 GB)
 - Partition D : installation of Khemeia (150 MB)
 - Partition E (example) : shared folders (1 GB) :
 - o Configuration
 - o Input folders for source documents
 - o Output folders
 - o Progress folders for non-shared work
- Configured for sharing of folders on PC/MAC networks

Frequently asked questions

How do I know if Khemeia is suitable for me?

From a sample of your content, we are able to evaluate how suitable this is for processing by Khemeia - running a test on your data then verifies the resulting output.

What if I am already using another workflow?

Your current workflow is likely to be either very labour intensive, or not offer the features and benefits of Khemeia. The integration of our product, which follows viability tests, is normally painless and will in fact enhance your existing system for processing information.

What if I already have a system for managing content or documents?

In this case Khemeia feeds enhanced content into whatever system you use for managing your content or documents.

Do Stelae Technologies compete with the services offered by my outsourcing supplier?

No, Stelae Technologies provide a software product which is actually utilized by your outsourcing service provider or yourself.

What if I currently outsource this work?

Like yourself, your outsource solution provider is most likely to welcome the productivity gains that the integration of Stelae's Khemeia into their workflow brings.

Should we or our outsourcing company run the software?

Either option can be attractive. Option 1: you yourselves run the software, and then your outsourcing company performs quality control. Option 2: the outsource company themselves take over

the entire workflow, running the software on your behalf, and also performing quality control.

I can understand that the software can deal with inline content, but can Khemeia really process the complex format of a typical magazine or newspaper?

Yes, absolutely. In fact key Stelae personnel have a magazine/newspaper publishing background, and therefore the software was built with this very application in mind.

Where does OCR (optical character recognition) fit in?

This is relevant if your information source is paper or scanned images of text. This does not present a problem, however an additional stage is then required to convert the source documents into digital files using OCR (e.g. output as PDF with background OCR layer). From this point on, Khemeia then takes over. (We can advise you on this workflow.)

We have our own taxonomy which we would like to utilize?

Yes, we do have this functionality. Khemeia allows specified terms to be validated against a taxonomy, dictionary or definition list.

If Khemeia is such a good product, why have I never heard of it before?

Khemeia, which is the result of years of intensive development work, has only recently emerged from research and development. We are expanding rapidly and already have a number of prestige customers using the product.

To learn more

So if you wish to add value to your data and increase the productivity of digitizing your information, Khemeia could be exactly the solution that you need.

To find out whether you should be using Khemeia to process your content, contact us now.

France: Stelae Technologies, 9 rue Jacques Coeur, 75004 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 93114 19 966

E-mail: info01@stelae-technologies.com

Web: www.stelae-technologies.com