

GathererPlus

Automated document spidering and rich metadata extraction

Managing the ever-growing quantities of electronic resources amassed by any organisation is a recurring headache for knowledge workers. Web pages, in-house documents, reports received in a variety of formats, images, emails: all add up to the vital resources that support an enterprise's day to day activity and decision making. Is there a way to automatically capture the metadata describing each resource and incorporate it into an intelligent, searchable entity?

Our answer is "Yes – with **GathererPlus**"!

Lexalytics has created GathererPlus to automate the task of collecting information about this wide range of resources. Not only will it tell you what resources are available, it will also extract useful details from the resources it finds. The results of a GathererPlus "run" or "job" are returned in a format that can be incorporated into a searchable database, providing a one-stop front end to all of the resources.

GathererPlus Features at a Glance

- Works with a range of electronic assets
- Extracts standard and customised metadata
- Can be scheduled to run during periods when the network is not busy
- Gathers only from new or updated assets
- Produces industry standard XML

GathererPlus incorporates two software "engines" which combine to locate resources and extract the details that will enhance the search capabilities of whatever front end is selected to provide access to the resource collection.

An **Acquisition Engine** performs the Gathering task. You select the type of resources to be gathered, and "point" the Engine to their location. Types that can be gathered include:

- File system contents: Microsoft Office documents, Adobe PDF
- FTP sites
- Web or intranet sites
- RSS Feeds
- Lotus Notes and ODBC compliant databases

Type	Name	Scheduled	Running	Log
	Maxus Web Site	No	Running	View
	IT feeds	No	Stopped	View
	Local FTP folders	No	Running	View
	Marketing Dept Emails	No	Stopped	View
	Internal File Servers	No	Stopped	View

Acquisition Engine home page, showing a list of GathererPlus jobs

In the background the **Salience Engine** extracts metadata from each resource. Out of the box the Salience Engine provides extraction of document properties (filename, size, date created/edited, author), as well as entities: people, places, companies, full text and summaries. The performance of the entity extraction can be improved, or trained, by the addition of inclusion and exclusion lists. You also have the power to set up your own tailored lists that will provide metadata relevant to your own industry or area of interest: pharmaceutical names, medical terms, software – the choice is yours!

Home->View Log

Maxus Web Site

Level	Date-time	Message
INFO	05/07/2007 14:05:04	Populating Collection
INFO	05/07/2007 14:05:04	Gather Started
INFO	04/26/2007 14:47:28	Gather Complete
INFO	04/26/2007 14:47:28	Finalizing Publishers
INFO	04/26/2007 14:47:28	No new Assets found
INFO	04/26/2007 14:47:27	Publishing Assets with file
INFO	04/26/2007 14:47:25	Outputting Assets with lxa
INFO	04/26/2007 14:47:25	Enhancing Assets with checksum
INFO	04/26/2007 14:47:01	Extracting Text and Native Properties
INFO	04/26/2007 14:47:00	Restricted 0 of 112 assets
INFO	04/26/2007 14:47:00	Restricting Files
INFO	04/26/2007 14:47:00	Expanding Composite Files
INFO	04/26/2007 14:26:11	Populating Collection
INFO	04/26/2007 14:26:11	Gather Started

Operations
Purge Log

Documentation
Overview
Job Details
Output
Filter
Schedule
Configuration File

GathererPlus log page for a typical web site spidering job

What's more, you have the capability of scheduling the GathererPlus to run at times when your system is less busy and at intervals to suit the volume of new resources being added; and the output from GathererPlus is in industry standard XML format. So you can begin to understand the power of this technology. Simply feed the XML data into your favourite search engine to complete the collection and publishing cycle.

Users of Inmagic® *WebPublisher PRO* software can take advantage of GathererPlus to automate the whole process: install and schedule GathererPlus to collect information as described above, then schedule the Inmagic Importer to load the XML data automatically into a suitable textbase with a web search screen that resides on your intranet.

GathererPlus: available from Maxus Australia and the network of local Maxus representatives.

About Lexalytics®
Lexalytics is a software and services company specialising in text analytics. Lexalytics technology has been adopted by some of the largest companies in the world, including Cisco Systems, and provides Cisco and other customers with the ability to understand and act on the important items in the sea of information that all companies now deal with.

About Maxus
Maxus Australia provides information management software and consulting services to a wide range of organisations across Australia and in the South East Asian region. Our consultants are the leaders in their field, with extensive information management and software experience, and specialise in programming, scripting, design, database set up and training.