

Administrator

The magazine for professional system and network administration

Special Edition for Nogacom

Under Test

Nogacom NogaLogic 3.8.0.6

With order and structure 



Under Test: **Nogacom NogaLogic 3.8.0.6**

With order and structure

by **Thomas Bär**



Source: stkyo30 - 123RF

The enormous quantity of unstructured data in companies – from Word documents, through CAD drawings to business plans for spreadsheet calculations – is and remains a problem. Pieces of information are rapidly produced, distributed, copied and processed by many different people. Security aspects, compliance or best practice ought to be the least of our concerns when a tender urgently needs to be sent to a customer. Such files potentially contain highly-sensitive information, which – in the wrong hands – could have extremely negative implications for the company.

Taming unstructured data

Bringing a company's unstructured information into line with the structured data is the task to which Nogacom is committed. NogaLogic has been available as a product in Germany since March 2010 and its entry into the market coincided with winning the Initiative Mittelstand award for innovation at CeBIT 2010 in the "Content Management" category. Of course, Nogacom is not alone in the automatic data classification field and there are various competitors. But before we look at the technical aspects, let us first consider how the soft-

ware works. For effective storage management and the migration of data from one storage area to another, the data itself must first be classified. Using copy and move actions, the software enables its users to automatically copy rarely used documents into less redundant storage systems, while leaving behind a link to the new storage location, or to automatically archive legacy data. However, this is just the final stage in the data analysis process.


Effective document organization through data analysis

The data analysis process consists of identifying the content of the software and the relevance of the data for the company, and marking the files accordingly. However, this tagging process does not modify the file itself – NogaLogic only requires write permissions when copy and move actions are requested. The information obtained, the context and additional meta information are stored by the software in the separate SQL database. The business context is the key to organizing documents more effectively. This context decides, for example, which access and distribution authorizations a document must be subjected

to, and how, where and for how long it must be kept. In the classification of documents, the user first considers properties such as "confidential" or "public", which can be appended via the operating system. If the company is newly established and uses these property fields consistently from day one on, the on-board facilities in the operating systems would probably be sufficient. However, reality looks quite different. Keywords are suitable for classification of documents in principle; the software however always requires the exact wording. Yet names often have different variations – therefore these are keywords for a single term. For example, in some

Two quad-core CPUs with 2.5 GHz, 16 Gigabyte RAM and storage depending on data volume. The configuration in a RAID system with multiple disks is recommended. The software requirements are for a Windows Server 2008 R2 Standard or Enterprise, 64-bit, in English, Microsoft Office 2003 Professional or Office 2007, Microsoft SQL Server 2008 SP1 Standard or Enterprise, Microsoft Internet Explorer 7.0 or higher and the Microsoft Silverlight 4 browser plug-in are also required.

System requirements



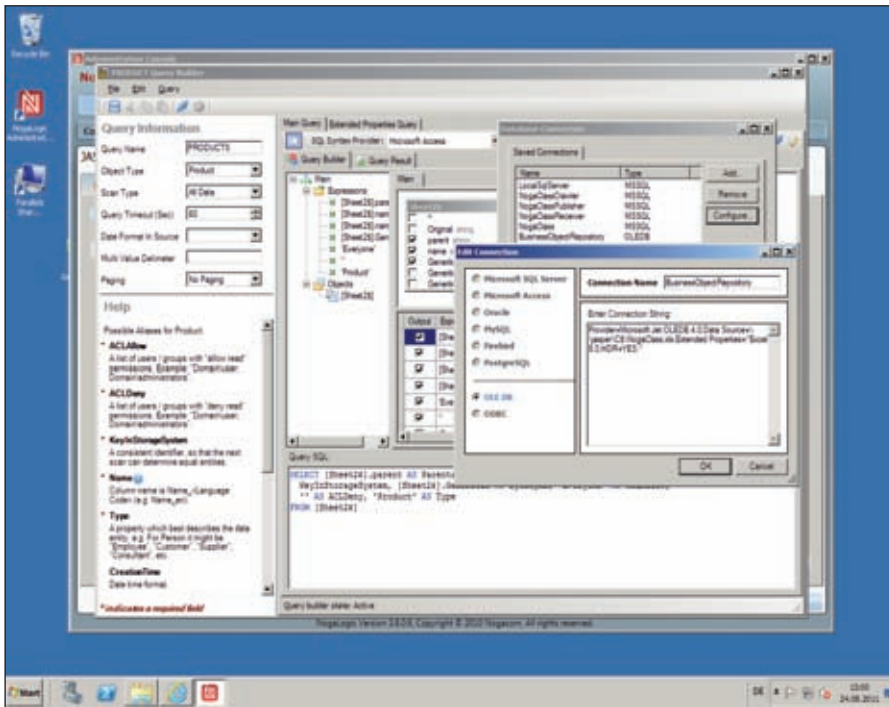


Figure 1: NogaCom's NogaLogic imports structured data from files or databases or connects directly with the data. This information is brought into line with the unstructured data. SQL knowledge is of advantage here.

business matters only part of the information may be used: acronyms, or an e-mail address, user name or identification number instead of a person's name.

Context-sensitive data classification through NLP

To ensure accuracy and completeness of tagging, data classification technologies must have an understanding of complex texts written by humans for humans. This is clearly explained in the manufacturer's product documentation using the word "nice" as an example. Humans can easily discern the meaning of the word when it is used in sentences such as "It was a nice day today" or "We met with Mr. David Nice". Pure keyword-based queries cannot detect such differences.

NogaLogic uses different morphological, lexical and semantic technologies for Natural Language Processing (NLP) during the data classification process. Using NLP, different business divisions to which reference is made in a document can be better identified. These are recognized not only by their actual names, but also on the basis of pseudonyms, synonyms and other features, such as – for example – e-mail addresses, telephone numbers or other unique or partially unique names.

NLP operates with the main European languages, including English, French, German, Dutch and Italian.

Installation in hours, classification in days

The installation of NogaLogic is generally straightforward but requires a little time for preparation. As recommended by the manufacturer, we installed NogaLogic on an English-language Windows Server 2008 R2 in the x64 version and on a Microsoft SQL Server 2008, also in English. According to information by the manufacturer, even though it would be possible to run the software on a German-language version of Windows, NogaCom recommends using an English-language system to circumvent any problems with settings for text recognition. The server was installed with Parallels Desktop 6 on an iMac with Intel Core i5-CPU, with four 2.5-GHz cores and 8 Gigabytes of RAM allocated – the maximum that Parallels can assign to a Windows client.

Depending on the anticipated data volume, in practice it is recommended to install the software on a Windows server with a minimum of 16 Gigabytes of main memory and a very fast hard-disk array. With correspondingly powerful hardware, according to the manufacturer, it is pos-

sible to classify a Terabyte of data in six days. In our example, with almost 12,000 documents and some SharePoint posts – totaling almost 550 Megabytes of data – the test machine needed a few hours before all the data was processed.

The installation instructions consist of a multipage list in English and describe every mouse click and every input, step by step. If the instructions are followed rigidly, the system – including database server and the first basic configuration – can be up and running in around 120 minutes. The ease of installation should not belie the fact that a workshop on using the software is necessary in all cases – otherwise the user will have difficulties in defining the relationships of the various "entities".

Static attributes create added value

Entities, otherwise known as static attributes, are crucial for describing the specific knowledge of a company. How otherwise is a computer to know that the number sequence "911" is a product of a car manufacturer, for example? Without storing entities, NogaLogic would be nothing more than an oversized search function. However, a large proportion of entities are produced without intervention by the administrator. The access rights, the definition of users in the active directory and the properties, as well as the e-mail addresses from the directory service, are also used by NogaLogic as attributes for linking. After the entities were defined the programs used as Windows services for data analysis, known as crawlers, began to evaluate the data. In the test the definition of entities was done by importing a list of generic drugs from Excel as products and the colleagues listed in the Active Directory as employees. The scanning frequency can be set by the administrator via the local admin console. All further steps such as analysis, job definitions or reporting are carried out by administrators and users via the Web browser.

Classification in the practice test

In our test the software was confronted with almost 12,000 complex Word do-



uments, which we wanted to look at in different logical combinations. In addition to File Services, NogaLogic had to gather the information from a Microsoft SharePoint Services 3.0 installation as data material. For this purpose we installed a small plug-in on the SharePoint server and defined the user account under which NogaLogic operates as the secondary website administrator in SharePoint. This process took no more than three minutes.

Connectors reveal associations

NogaLogic offers a multitude of connectors to various systems such as Exchange, Lotus Notes and various database systems. If there is no other option a data connection is possible via ODBC and SQL command. For example, if unstructured data such as invoices in PDF format or tenders in Word format are to be associated with the structured data from a database, this is possible either via a direct database connection or via a regular importation of data files. In the test, we imported the aforementioned file containing the names of generic medications. For other environments this might have been a customer list, a spare parts catalogue or a list of credit card numbers.

The associations between documents were detected by the software during the test

without difficulty. Even though they are stored in different directories, the software identified (for example) the medical history of a patient along with the medical reports relating to the hospital stay for the physicians providing further treatment. NogaLogic is therefore able to detect that documents containing similar concepts and from the same period belong together. The qualitative assessment of the association is expressed using the value “Association Rate”.

This value also makes it very easy to identify a copy of a file stored in a different location. If the value is “1,000”, this means that it is an exact copy of the file – even if the file name has changed. NogaLogic detects this value by content comparison and not by means of a simple hash value calculation. If NogaLogic is linked via connectors with mail systems such as Exchange or Notes, the user is able to see in the “Distribution” tab, for each document, when, where and by whom the document was sent.

Gaps in the scanning of exotic file formats

Classic file types such as Word and Excel files or PDF files can be searched for straight away with NogaLogic. According to the manufacturer, the software detects around 300 different data formats. If a for-

mat is not recognized, the software can be extended by the manufacturer if it is notified of the necessary specifications. However, the additional installation of any iFilter – as was required, for example, for the index service in the Microsoft environment – is not done with NogaLogic. For example, the relatively unknown format *.mm from the free, Java-based mind-management solution “Freemind”, was detected by the crawler so that the contents were indexed, yet later, in the search functions, we were unable to find the contained word “acetylsalicylic acid”.

Search results adapted to permissions

Simple search queries are built up in successive stages in the software. If the user begins by entering a search term or a previously defined entity and receives an excessively large number of hits, he then limits the search through further stages until the required result is shown. The results of a search always correspond to the access rights of the logged-on user. Thus using the tool does not enable the user to access information for which he is not authorized.

By clicking on “Advanced Search”, the user is also able to construct the search using the Boolean functions “AND” and “OR”. There are however sometimes cases that cannot be reproduced. If a name such as “Prof.Dr.Mayer” is erroneously written together as one word in a document, it is not possible to structure a search in the form of “Show me all the documents by Dr. Schmidt, which do not contain Prof. Dr. Mayer” since there is no “does not contain the word” option in the software. This is an aspect that the NogaLogic programmers really ought to rework.

Fortunately there is a preview function that allows the documents located to be viewed. If the user saves this search result, it can be called up in future under “My Views”. Users with administrative rights can make their views accessible to other users – the main search algorithms for the company are therefore provided centrally. These search results in “MyViews” are an essential basis for data migration and data management with NogaLogic.

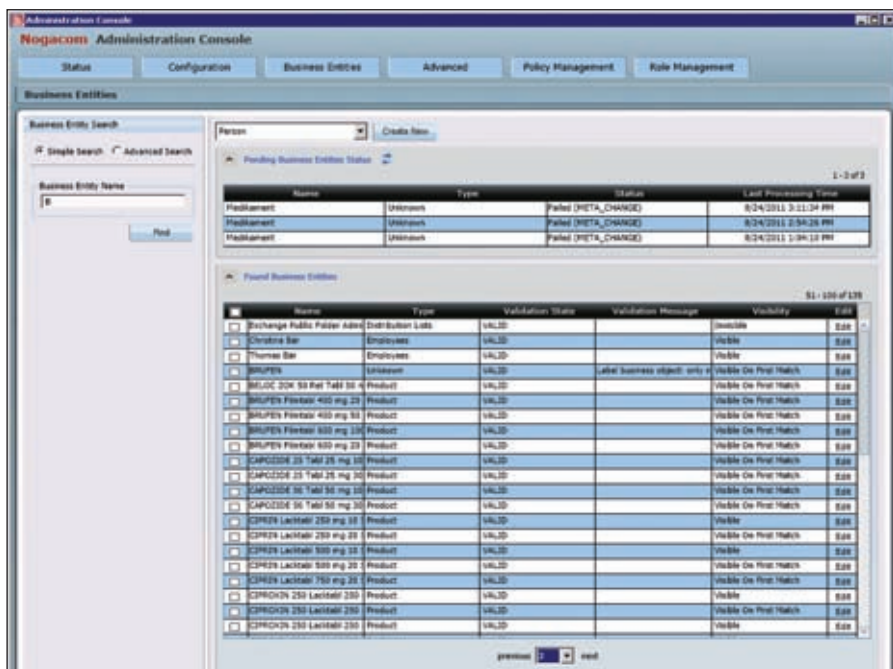


Figure 2: How is the data combined? The user sets up a catalogue of specific entities via the management console. The master data of the Active Directory is immediately available.



From the administrator's perspective, there is one further simple yet exciting function: limiting the search according to timeframes. In the standard product, the options "Past 24 hours" as well as last week and last year can be selected directly. If a user then calls and is missing a file, the administrator is able to find it quickly – wherever it is located. Even better: if the user has access to the tool, he can search for the file himself – his access rights always being taken into account.

Rule-based data tiering

If, for example, a search is structured to include all documents that have not been opened for more than 180 days, this is well suited for an automatic copy job. Keeping unused documents in expensive high-speed storage in the company is extremely costly. Instead of storing these documents – which are rather unimportant as far as the daily work is concerned – in a first tier memory, and possibly saving them several times a day, these documents would probably be better placed in a less important storage system with slower access speeds and a lower backup-rate.


In NogaLogic the administrator sets up such automation jobs in "Policies". A policy is based on a "Scope" – which is actually a search template that the user has defined previously under "My Views". What happens with this data is defined in the second stage. Classic actions are copying it into another shared folder on the network, moving it to a different directory or transferring it to a different technology such as SharePoint, for example. These

policy jobs are executed automatically via a scheduler, thereby relieving the storage systems of the burden of unnecessary data and ensuring that the documents of high importance for the company are automatically held in a different storage location.

Summary

During the test we were particularly impressed by the quality of the classification and the modern design of the graphical interface. It is thanks to Microsoft's Silverlight, as the underlying technology, that pie charts – for example – take shape in an attractive way before the user's eyes. During the test, there were frequent interruptions in the display when Firefox 5 was used on Windows, so that we were finally obliged to work with Microsoft Internet Explorer.

With NogaLogic, IT is able to get a picture of the state of the stored data, set up a new classification system, and maintain compliance. With advanced techniques, such as the classification of confidentiality and the monitoring of outgoing e-mails, it is also possible to protect against the loss of confidential information (Data Leak Prevention).

NogaLogic is presented as a classic tool – the conceptual work and the knowledge around the associations must be contributed by the users. The more detailed the company information is, the more accurate the subsequent results which can be achieved with the software. Using the master data of the software the administrator can make many adjustments to optimize the product for his specific environment. (In) 

Product

Program for data classification and rule-based document handling.

Manufacturer

Nogacom
www.nogacom.com

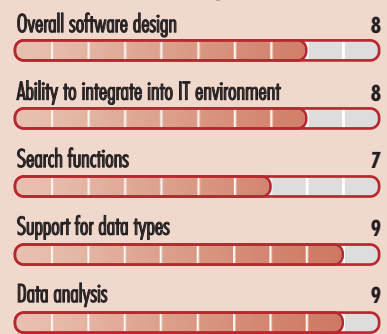
Pricing

The license starts at 35,581 Euros, includes three connectors (Active Directory, File System, ODBC) and is suitable for a million documents (following initial consolidation of legacy file versions and copies). Maintenance in the first year is included, along with 24-hour online support, telephone support on weekdays, and also updates and upgrades. The manufacturer optionally offers further connectors (Exchange, Lotus, SharePoint) and document packages ranging from 500,000 up to a maximum of 20 million documents.

Technical data

www.it-administrator.com/downloads/datenblaetter

IT-Administrator rating (points out of 10)



This product fits

perfectly for companies that need to bring order and overview to large quantities of unstructured data

somehow for companies with well thought-out concepts for putting unstructured data in order

not for companies in which there is only very little digital data.

Nogacom NogaLogic 3.8.0.6