

THE OVERVIEW

InfoMAX is a data-warehouse building tool targeted for specific need of the current industry. The tool is developed with a modular approach. The various modules are Extraction, Cleaning, Parsing, Intelligent Search and Update, Selection, De-duplication, Normalization and Retrieval.

The main attractions of this tool are the Flexible Rule Builder and the Matching Engine used for De-duplication. The Flexible Rule Builder allows the user to create their own rules for Cleaning, Intelligent Search and Update and the Selection Processes. The user can define a set of rules to be acted upon the specified database. InfoMAX also have its own set of pre-defined rules, which is enough to clean and parse the database.

Our tool is dynamic and has a revolutionary artificial intelligence that parses data even more accurately as time goes on. The more accurate the parsing the more accurate the de-duplication process will be so the longer you use the product the more accurate future parsing will be.

InfoMAX incorporates the very latest fuzzy logic matching algorithms and will allow you to find these duplicated records. InfoMAX has the highest retention of error matching than any other product on the market due to its fuzzy logic advanced algorithms.

The Matching Engine is very much different from the existing matching engines available in the market. The user can specify which fields to take part in the matching process and can also specify the weight of each field. Also it's possible to match the same set of records with different settings. This process is very much dynamic and can be used to match any type of data.

Something very unique to InfoMAX is the market expansion for other countries. If you require country specific settings we can provide that country and add it to your InfoMAX with our install new market feature or what we call the (Market Discovery). Unique (Market Discovery) installs dependant upon the language can take 1 - 2 months to create. This is far faster than implementing a new software package otherwise required with any other competitor.

BUSINESS OBJECTIVE

InfoMAX can be used for the following business events.

- Cleaning and parsing of records.
- De-duplication of records in a huge database.
- Compliance with PATRIOT ACTS of USA and any other EUROPEAN Countries.
- Can be used in case of Company Mergers to clean and de-duplicate huge databases and build one database.
- The Client needs to get more customers but they can't identify proper prospects.
- InfoMAX can help to know who are the profitable customers.
- InfoMAX can gather all customer data in your organisation and help to analyze the data that can help to get new customers and retain existing profitable customers.
- InfoMAX can assure you that you are not contacting the same customer again and again.
- With the InfoMAX Fraud detection you can see from which customers you are losing money.

TECHNICAL OVERVIEW OF MODULES.

Extraction & Identification

This Process is used to extract data from databases. Other than databases InfoMAX can work with text files bought from market. Each extraction process will build a common pool of data from different databases. These databases can be heterogeneous. This common pool can be user specific, may be a text file or a database table. If the data extracted is in text files and not properly defined then it needs identification i.e. Mapping with InfoMAX supplied fields. If the data extracted is in a database and is well defined then also its needed to be mapped with InfoMAX fields. But these well-defined records may not need Cleaning and Parsing.

The user can specify Criteria for extracting from a database, like 'Where the age is > 21'. This type of proactive data selection can help the user to avoid delay in processing for all the later processes that needs to be run in the data.

In this Extraction process the system can identify a broader distinction of the data.

e.g. The process can identify NAME and ADDRESS but may not identify the FirstName, LastName, or StreetNo, or StreetName. This Finite Field Analysis is done in the parsing process.

The most challenging job that InfoMAX can do is to clean and parse data from text files faster and accurately.

Cleaning

This process is used to remove junk characters from the data. The user can specify which field is to be cleaned and the user can also specify what characters should be cleaned. There are various dictionary tables provided by InfoMAX and also can be defined by the users. A process can use this dictionary tables and take specific action according to the data in those tables.

Other than removing junks the user can also use this process to remove records with slang in any field. The users in a dictionary can also specify the slang.

Making a field or a record in UPPER CASE or LOWER CASE is also possible in this cleaning process.

Example:

The record below contains junks like ".", "#" and ","

MARY KING 423, AYLESFORD PL APT#4, LEXINGTON, KY, 40508 PHONE: 428-4791

After cleaning it becomes.

MARY KING 423 AYLESFORD PL APT4 LEXINGTON KY 40508 PHONE 428-4791

Parsing

The Parsing process is divided into 2 parts. One is **Inter-Data Type-Parsing** and the other is **Intra-Data Type-Parsing**.

At the time of Extraction the fields may not be identified finitely. Like the system can identify the NAME data type but may not know what is the FirstName and what is the LastName. The Parsing process allows this finite field analysis.

Some time this can also happen that some LastName of some fields are remaining in the ADDRESS data type. In this case we use to extract that data element and place it in its proper field. This process is called **Inter-Data Type-Parsing**.

And the **Intra-Data Type-Parsing** used to find out data elements from the Data type. This process Is used to break up the ADDRESS Data Type into StreetNo, StreetName, StreetType, City, State and PIN.

To do this parsing accurately the InfoMAX takes help of the Market Specific Dictionary tables that is provided by the system or can be defined by the users.

Example:

The record below contains junks like ".", "#" and ","

MARY KING 423 APT4 LEXINGTON KY 40508 PHONE 428-4791

After Parsing it becomes.

MARY	KING	423	AYLESFORD PL	APT 4	LEXINGTO N	KY	40508	4284791
------	------	-----	-----------------	----------	---------------	----	-------	---------

Intelligent Search & Update.

In a large database or text files bought from market some information may be missing. Like the Gender field or the Age field or there may be some requirement for standardizing some fields like Date of Birth or a Salary Field. Or there may be a need for updating the Gender field after looking at the FirstName.

Some fields such as dates, phone numbers SSN numbers, passport numbers and many others can be updated to the same format. This will help in the matching while de-duplication.

Example:

The record below contains junks like ".", "#" and ","

MARY KING 423 APT4 LEXINGTON KY 40508 PHONE 428-4791

After Updating with Intelligent Search it becomes.

MARY	KING	423	AYLESFORD PLACE	4	LEXINGTO N	KY	40508	4284791	F
------	------	-----	--------------------	---	---------------	----	-------	---------	---

In the last column an F is inserted to specify that this person is a FEMALE. And the APT is deleted from the APARTMENT field. The PLACE is substituted for the Abbreviation PL.

Selection.

There are some records that can be rejected and not needed in the database. If there is a record without NAME then its meaning less to keep it as a member of the database. Lets say that there is a record without a ADDRESS or say without the Street No in the ADDRESS. These records can be rejected through the dynamic rules.

InfoMAX provides some preset rules for the rejections and the user can also build their own rules with the help of the Flexible Rule Builder.

The Selection process selects all the records that don't meet the requirements defined by the rejection rules.

Loading.

The loading is the process that loads the data into the database or a separate text file after the serious of processing like Data Cleaning, Data Parsing, Intelligent Search and Update and Data Selection.

The total processing doesn't require all the above processing. Like, if the data is already clean then the user can avoid the cleaning process, if the user don't require the rejections the its not needed to set rules for the selection process. A selection process with no rule will select all the records.

Only the Identification and Loading process is must.

The Cleaning, Intelligent Search & Update and the Selection rules can be set with the Flexible Rule Builder Front End.

De-Duplication

The previous processes were used to prepare the data for the De-duplication process. If the clients data is already prepared and doesn't needs cleaning, parsing and Standardizations of data then the user can run this De-duplication after an Identification and Loading is done.

In this De-duplication process all the duplicate records in the database are assigned group numbers. The user can control the process with a flexible settings control. Users can tell what fields are going to take participation in the matching process and what will be each fields sequence and can also assign a weight for each field.

It's also possible to match two records where the fields are exchanged in two records.

e.g. Lets say there are two records as follows

Rec. No.	FNAME	MNAME	LNAME	ADDR
1	ARTHUR		ACOSTA	616 NORTH FULLER AVE LOSANGELES CA 90036 US
2	ACOSTA		ARTHUR	616 N. F. AVE LOSANGELES CA 90036 US

In this example the FNAME and LNAME field values are interchanged. And InfoMAX can match these values with the help of the Field Interchange Matching Capability.

Another important is the Address Matching. Most of the time there can be multiple components in the Street Name part. In the above example the Street Name is as follows.

Rec. No.	STREET NAME
1	NORTH FULLER AVE
2	N. F. AVENUE

Our tool can match these two strings and say what is the percentage of match. Like N will match with NORTH and F will Match with FULLER and AVE and AVENUE will be 100% match as AVE is the abbreviation of AVENUE. With this special logic of matching we can also match strings if there are spelling mistakes and other abbreviations.

The Matching process is used to group similar records as per user rules. The user can specify the group specification. It can be an address group or it can be an individual group, or may be any other group depending upon user needs.

For example if the user needs to group peoples who lives in same house then she can include only ADDRESS related fields in the Matching rules. And if she needs both house grouping and individual level matching then she can set the rules such, that first InfoMAX will prepare the household group and then the individual group in one go.

There are lots of special capabilities in our matching capabilities that can be shown in a Demo releasing soon or in a Pilot run.

Normalization.

After the Matching all the duplicates are grouped and the user must need to find out the single best record of a group. This Normalization process can be used to do this.

There are two ways to do this.

1. To find out the best record with user driven settings. In this process the user can set rules like "If the NAME field is not null", "The record with the most lengthy Street Name" etc. The different rules can be applied with AND / OR combinations.

If no records can be finding with the rules then the user can also specify what to do. In these cases the first record or the last record can be taken. Or the user can also specify other rules with the Flexible Rule Builder.

2. Creation of a new record with the best field values from the different records in the same group. The user can build a new database or a text file with these records with different file number. The user can set rules for this process. The rules can be like as follows:

"Take the field value from the FNAME field where the length is max"

"Take the field value from the LNAME field where the length is max"

"Take the field value from the MNAME field where the MNAME is not null"

"If MNAME is null in all records the make it NULL.

"Take the STREET NO from the record where the length is max"

"Take the STREET NAME from the record where the length is max"

"Take the PHONE from the record where it is with the STD code or at least 10 digit"

... And many more

There are many such rules provided by InfoMAX but the user can create their own rule according the database record characteristics. The InfoMAX Flexible Rule Builder can be used to set this rules.

After selecting the record with any of the above two types of actions the Normalization can be instructed to build a new database or text file with this new records and with this process the existing database can be purged. The newly created records or the best records can be kept in the existing database and the other records can be marked for deletion, or can be deleted from the database, or a new repository can be created to keep the new records and the existing database will be untouched or the user can keep the new or best records in the existing database and make a history repository and keep the other records in that repository.

Retrieval

After the records are prepared and de-duplicated and the database is normalized, now the user will definitely need a process to identify the correct records and format them in a user – specified format like a Mailing Label.

In this process the user can do the Following actions.

InfoMAX Data-Warehouse Builder - White Paper

1. User can specify which records to Retrieve
2. User can specify whether to keep the records in database or the records will be exported to a text file with special format.
3. User can specify the format in which the records will be formatted in the text file.

e.g. The format can be like this

```
Line 1 SALUTATION<1 SPACE>FNAME <1 SPACE>MNAME<1
SPACE>LNAME
Line 2 STREET NO, <1 SPACE> STREET NAME<1 SPACE> STREET
TYPE
Line 3 CITY, <1 SPACE> STATE
Line 4 COUNTRIES <1 SPACE>--<1 SPACE> PINCODE
```

4. The User can specify the line gaps between two formatted records.

This is the last process in the InfoMAX data-warehouse builder. With this process the user can do two things. Number one is to select specific records from the database and Number two is to create special formatted output for various needs.

For more details and pricing.

CONTACT

Orbitron Technologies Ltd

Tel. 0870 760 7536

Fax. 0870 133 7017

email: infomax@orbitrontech.com

website: www.orbitrontech.com