# EMM OSINT Suite – Improved Entity Recognition

## Description

The EMM OSINT Suite is a desktop software package which consists of various tools based on JRC's research in open source text analysis and mining. The latest version features an improved entity recognition (extraction) module.

The software consists of the following core modules:

### Data Acquisition

- Search – a component to extract search results from online search engines
- Crawler – a HTTP crawler module to harvest data from targeted web sites ("crawling")
- Grabber – a HTTP client module to download text based or binary documents from web sites for further processing

### Data Processing

- Text Extraction – extracts texts from different text based and binary formats (XML, TXT, PDF, MS Word, MS Excel, MS PowerPoint, Open Office)
- Entity Extraction – a set of modules to extract named entities from raw text. Entity types are people, organisations, locations, address information, VAT numbers and user defined custom types
- Category Matching – categorises text according to key word based category definitions

### Data Analysis

- Reporting – a component to create reports for end users of for further external processing of extraction results
- Local Search – a local search index to provide full text search of downloaded artefacts
- Entity Browser – an analysis component to aggregate found entity data and allows browsing through the results.

The different tools are made available with a graphical user interface based on the Eclipse Rich Client Platform which is an open source toolkit for desktop applications.

## Release Notes

The 2015 release contains the following improvements and bug fixes:

### Improvements

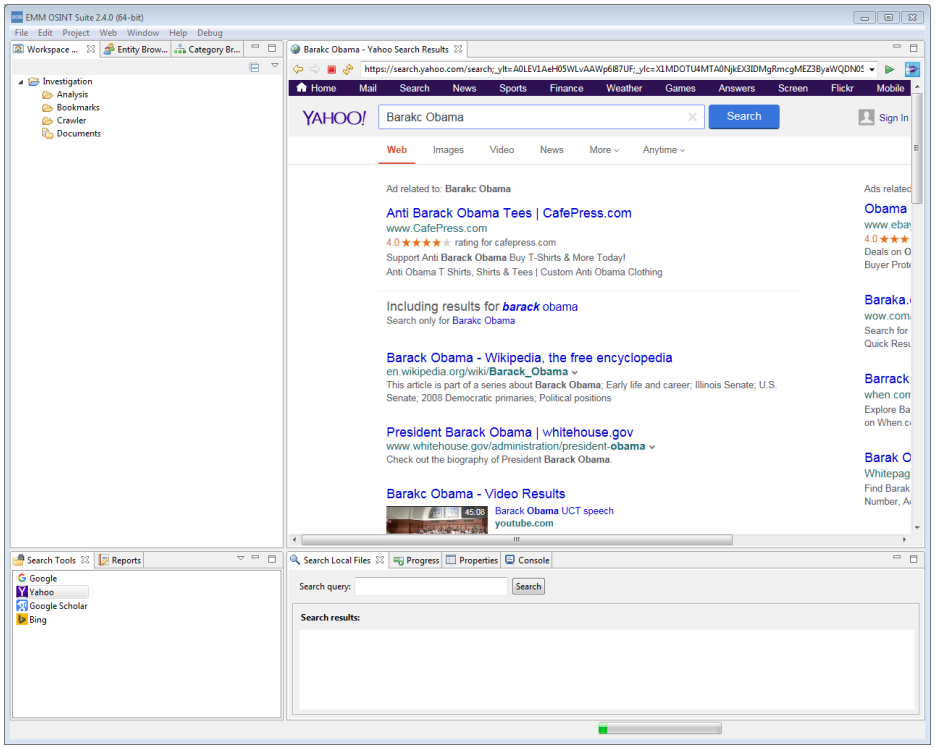
- **Improved Entity Recognition (Extraction)**
  - **Latest language resources and guessing patterns**
  - **New graphical view to test regular expression patterns (bot for BRICS and JAVA dialect)**
  - **New default custom pattern (for example Dutch zip code)**
- Improved Category Matching module
  - Find documents which match a specific keyword pattern
  - Updated domain specific language
- Internal analysis data model
  - Complete revision with substantial performance improvements and much less memory consumption
- Document handling
  - Improved document repository to enable on-demand translation service
  - Experimental import of complex document formats (for example MS Outlook PST)
- Automatic Software Update
  - In-place update from EMM server repository

### Fixed issues

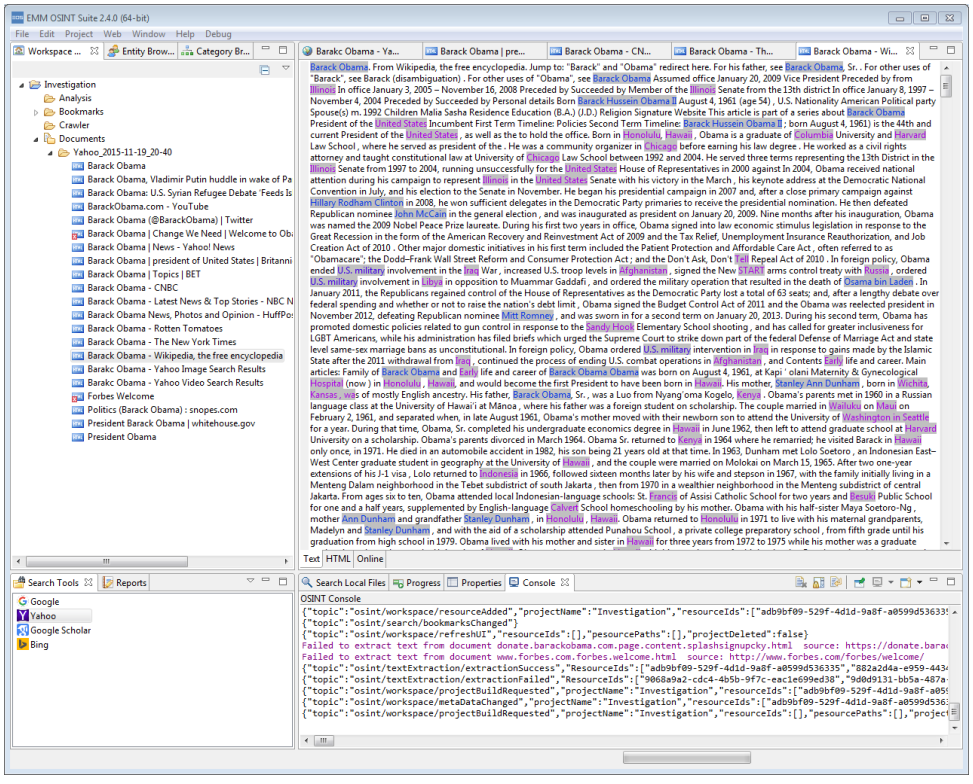
- Corrected graphical representation of entity relationships
- Updated Java Runtime
- Performance Improvements
- Completely revised build process based on Apache Maven and Tycho
- Fixed search result extraction patterns

# Screenshots

## Data Acquisition



## Entity Extraction



# Custom Entity Extraction