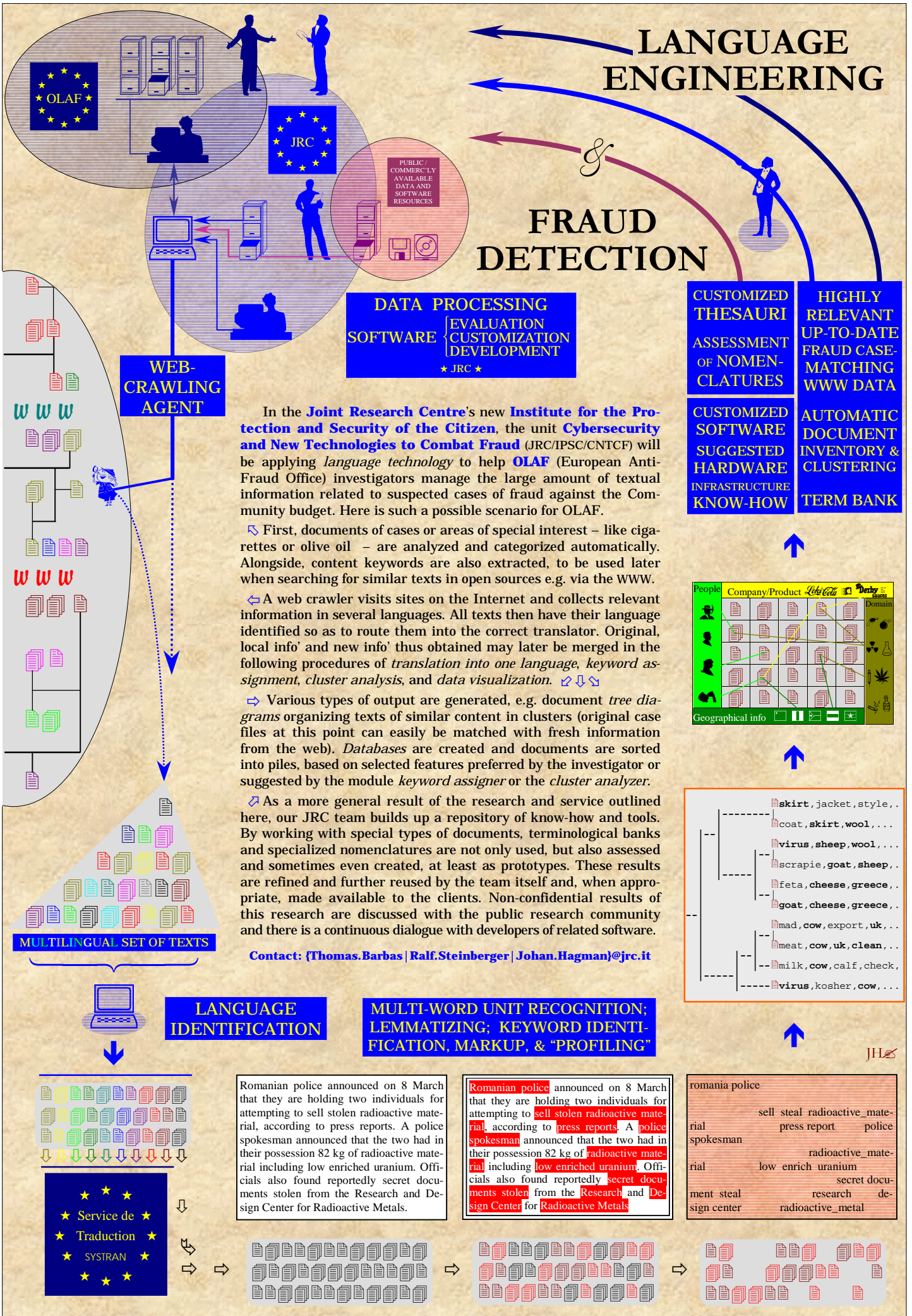


# LANGUAGE ENGINEERING

# FRAUD DETECTION



**DATA PROCESSING SOFTWARE**  
 EVALUATION  
 CUSTOMIZATION  
 DEVELOPMENT  
 ★ JRC ★

In the **Joint Research Centre's** new **Institute for the Protection and Security of the Citizen**, the unit **Cybersecurity and New Technologies to Combat Fraud** (JRC/IPSC/CNTCF) will be applying *language technology* to help **OLAF** (European Anti-Fraud Office) investigators manage the large amount of textual information related to suspected cases of fraud against the Community budget. Here is such a possible scenario for OLAF.

First, documents of cases or areas of special interest – like cigarettes or olive oil – are analyzed and categorized automatically. Alongside, content keywords are also extracted, to be used later when searching for similar texts in open sources e.g. via the WWW.

A web crawler visits sites on the Internet and collects relevant information in several languages. All texts then have their language identified so as to route them into the correct translator. Original, local info' and new info' thus obtained may later be merged in the following procedures of *translation into one language, keyword assignment, cluster analysis, and data visualization*.

Various types of output are generated, e.g. document *tree diagrams* organizing texts of similar content in clusters (original case files at this point can easily be matched with fresh information from the web). *Databases* are created and documents are sorted into piles, based on selected features preferred by the investigator or suggested by the module *keyword assigner* or the *cluster analyzer*.

As a more general result of the research and service outlined here, our JRC team builds up a repository of know-how and tools. By working with special types of documents, terminological banks and specialized nomenclatures are not only used, but also assessed and sometimes even created, at least as prototypes. These results are refined and further reused by the team itself and, when appropriate, made available to the clients. Non-confidential results of this research are discussed with the public research community and there is a continuous dialogue with developers of related software.

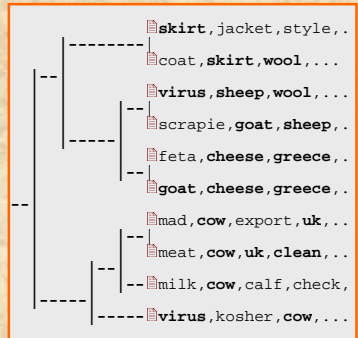
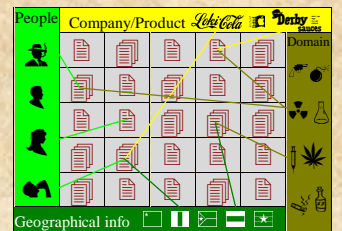
Contact: {Thomas.Barbas | Ralf.Steinberger | Johan.Hagman}@jrc.it

**CUSTOMIZED THESAURI**  
 ASSESSMENT OF NOMENCLATURES

**HIGHLY RELEVANT UP-TO-DATE FRAUD CASE-MATCHING WWW DATA**

**CUSTOMIZED SOFTWARE SUGGESTED HARDWARE INFRASTRUCTURE KNOW-HOW**

**AUTOMATIC DOCUMENT INVENTORY & CLUSTERING TERM BANK**



**LANGUAGE IDENTIFICATION**

**MULTI-WORD UNIT RECOGNITION; LEMMATIZING; KEYWORD IDENTIFICATION, MARKUP, & "PROFILING"**

Romanian police announced on 8 March that they are holding two individuals for attempting to sell stolen radioactive material, according to press reports. A police spokesman announced that the two had in their possession 82 kg of radioactive material including low enriched uranium. Officials also found reportedly secret documents stolen from the Research and Design Center for Radioactive Metals.

Romanian police announced on 8 March that they are holding two individuals for attempting to sell stolen radioactive material, according to press reports. A police spokesman announced that the two had in their possession 82 kg of radioactive material including low enriched uranium. Officials also found reportedly secret documents stolen from the Research and Design Center for Radioactive Metals.

romania police  
 sell steal radioactive\_material  
 press report police  
 radioactive\_material  
 low enrich uranium  
 secret document  
 steal research design\_center  
 radioactive\_metal



JH