EACEA Action Grants

Project reference No: 575734-EPP-1-2016-1-NL-EPPKA2-KA

International Security Management Knowledge Alliance (ISM-KA)
Addressing Security Challenges in an Interconnected World

WP1 Horizon Scanning:

Description Functionality Horizon Scanning Tool

Deliverable Form	
Project Reference No.	575734-EPP-1-2016-1-NL-EPPKA2-KA
Document Title	Description Functionality Horizon Scanning Tool Public Version
Relevant Work package:	WP1 – Horizon Scanning
Nature:	Report
Dissemination Level:	PU
Document version:	Final
Date:	11.12.2019
Authors:	Dutch Police Academy
Document description:	This document describes the delivered functionality of the horizon scanning tool for the ISM-KA project.

Co-funded by the
Erasmus+ Programme
of the European Union

Table of Contents

LIST OF FIGURES **2**

INTRODUCTION **3**

1. AIM OF THE DOCUMENT **3**

2. FUNCTIONAL MODULES **3**

2.1. Real time search 3
2.2. Discovery and Content Creation 4
2.3. Content Management (Process and System module) 5

LIST OF FIGURES

Figure 1 HST Home Screen 3
Figure 2 HST Search Screen 4
Figure 3 Document Funnel - Trend Module 4
Figure 4 (semi) automated scanning by Pan Oramix module 6
Figure 5 Content Management System 6
Figure 6 Body of Knowledge 7

Co-funded by the
Erasmus+ Programme
of the European Union

Introduction

1. Aim of the document

In this report we describe the functionality the Horizon Scanning Tool (HST) as developed and technologically implemented on a test server. This report has been annotated in order for it to be publicly accessible. We describe the functionality of our developed Horizon Scanning Tool in order to inspire other interested parties.

The Horizon Scanning Tool, in its first implementation, will not be accessible to the public but will be used as part of the delivery of the Executive Master on International Security Management. If and when the Horizon Scanning Tool becomes publicly available in the future, we will communicate this on our website www.ism-ka.eu

2. Functional Modules

In this section the four functional modules are described: real time search, discovery, content creation, content management.

The Pelorus entry point – the Home tab - is a WordPress website that also functions as a CMS. Through plug-in the WordPress site also gathers content from the internet and converts the captured information into a pleasantly readable format. The CMS supports text, image, video and sound media. User management with levels for readers, posters and administrators is handled by WordPress.

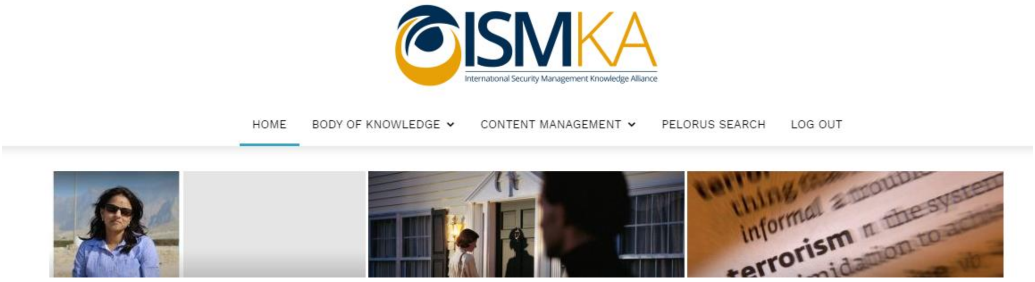


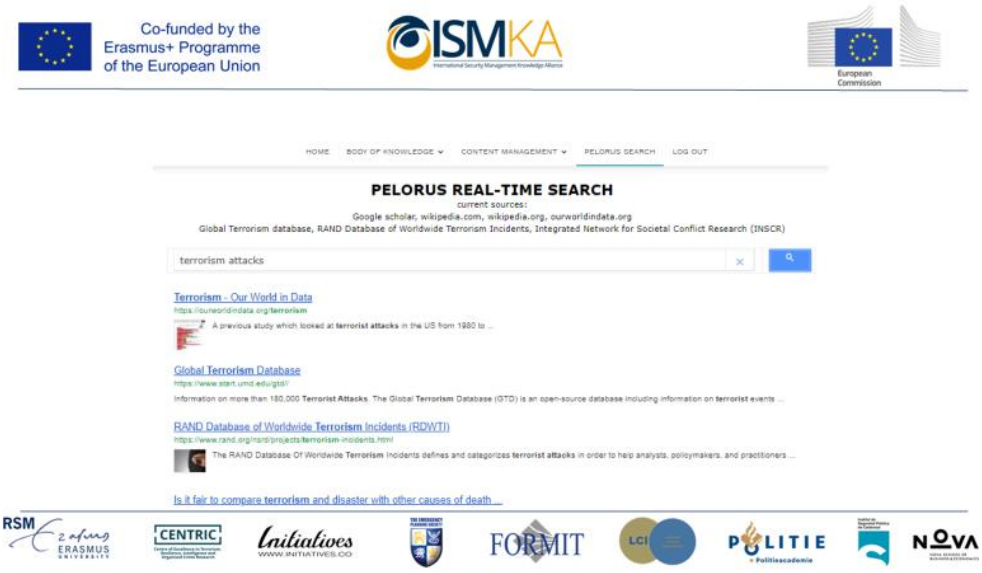
Figure 1 HST Home Screen

2.1. *Real time search*

The key to a successful analysis is not how much information the user gets, but how valuable the information is to him or her. The normal activity of users by using key words as search terms is supported and added by the semantic retrieval engine of Pelorus. It can extract themes, analyse language and synthesize a document.

Pelorus gathers its search/scan results from two types of digital information: CMS postings of relevant content and public Internet sources. Internet results are amalgamated with relevant internal information. Meta information is extracted from documents and is utilized in the semantic analysis of text. In this sense it is a self-learning system that in time – that means after a period of use (search), content validation and posting – improves the quality of the results given. (see paragraph on Discovery and Content Creation).

Co-funded by the
Erasmus+ Programme
of the European Union



PELORUS REAL-TIME SEARCH
current sources:
Google scholar, wikipedia.com, wikipedia.org, ourworldindata.org
Global Terrorism database, RAND Database of Worldwide Terrorism Incidents, Integrated Network for Societal Conflict Research (INSCR)

terrorism attacks

Terrorism - Our World in Data
https://ourworldindata.org/terrorism
A previous study which looked at terrorist attacks in the US from 1980 to ...

Global Terrorism Database
https://www.start.umd.edu/gtd/
Information on more than 180,000 Terrorist Attacks. The Global Terrorism Database (GTD) is an open-source database including information on terrorist events ...

RAND Database of Worldwide Terrorism Incidents (RDWTI)
https://www.rand.org/nsrd/projects/terrorism-incidents.html
The RAND Database Of Worldwide Terrorism Incidents defines and categorizes terrorist attacks in order to help analysts, policymakers, and practitioners ...

Is it fair to compare terrorism and disaster with other causes of death ...

Figure 2 HST Search Screen

2.2. Discovery and Content Creation

After a new post is entered in the CMS the indexer and analyser process the post and creates a keyword index and a semantic graph of the document content. From the graph a list of top document features is extracted and stored as additional meta data.

This analysis is later used by the trend analyser together with additional meta data and different selection criteria.

Periodically the document features are grouped per category and ranked by age and frequency of user access. The resulting features are then submitted to a public internet search engine. The results of the internet search are compared with existing posts for similarity whereby outliers are purged. Only the top-ranking results are then added as new posts to the CMS.

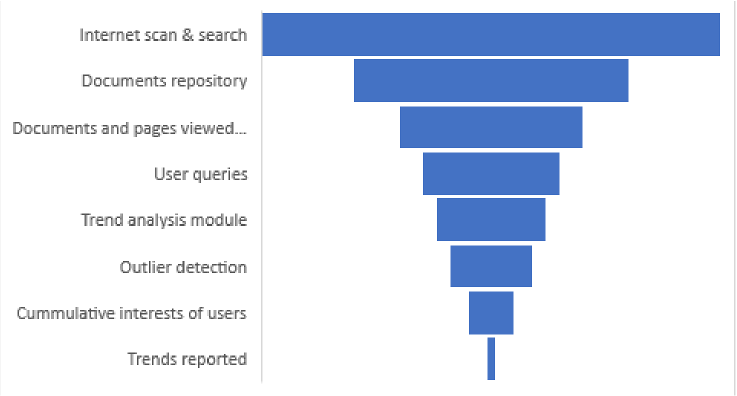


Figure 3 Document Funnel - Trend Module

The analysed content and meta data of posts, as extracted by the semantic network and content discovery module, are used to determine if any trends are emerging and can be distilled from the data. If trends are detected they are show on a per category basis to users in a sidebar on the user interface.

Categories are a first discriminator as these are orthogonal. The information published is analysed with a news detection approach. The ISM-KA data poses a challenge to any available 'first story' algorithm as these are tuned towards large volumes of short articles.

The factors involved in determining the ‘newsness’ of a post are:

- The age of the post.
- The number of users viewing the post over a fixed period.
- The number of trending posts in category.
- The popularity of the top document features compared to the category average.
- The popularity of the category compared to other categories over the last 100 postings.
- Whether the source of the features are a human posting or a discovery auto posting.

Although we expected that interest in older postings would decay quickly and new posting started gathering interest this pattern was less clear than expected. The windows of analysis had therefore to be widened to get acceptable results. This may change when Pelorus is more heavily used and contains an order of magnitude more postings.

A factor not considered in the trend module is the influence of the poster. This (meta)data is not available in the CMS and could therefore not be used. This has to be established by setting up a process of evaluation of retrieved and posted content.

The trend algorithm implementation executes these steps to form an ensemble classifier:

- Pre-processing reduces lexical variations and reduces the vocabulary’s size. This is a rule-based process to replace for instance ‘coz with because. Hyperlinks are extracted and added a meta data for later analysis.
- The text is split into words and represented into vector space. It is then compared with the vectors from same category to determine the similarity quotient, using Locality Sensitive Hashing. This reduces the dimensionality and therefore reduces the number of cosine comparisons the vector will need to find the N nearest neighbours. Having computed the Euclidian distance with all near neighbour candidates, the post with the closest distance is assigned as the nearest. If the distance is above a certain threshold, the new post is considered a first story.
- The post is compared to a limited number of most recently posts. The neighbour with the shortest distance (the highest cosine score) is assigned as the nearest and the distance represents the score of the post. If a post has score above a certain threshold, it is identified as a candidate trending topic.

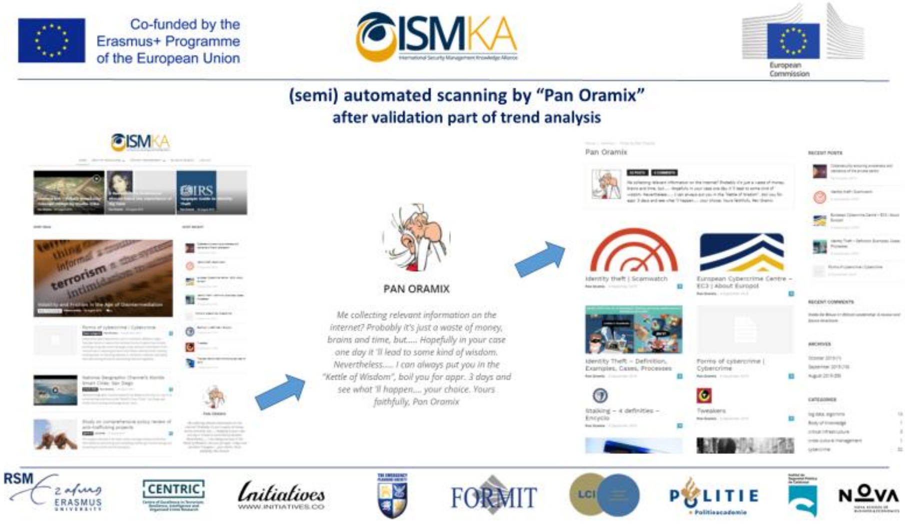
2.3. Content Management (Process and System module)

The content management process is human manual activity and consists of the following steps: selection of manual posted content, evaluating the automatic generated content, uploading and categorizing (using meta data and the taxonomy key concepts), adding extra information when desired, making it findable and publishable.

Content creation (see paragraph before) and management is building and maintaining a body of knowledge.

In the horizon scanning tool part of the content is ‘created’ by a module called Pan Oramix. This is an automated search and scan functionality that is fed with relevant key words by the administrator.

Co-funded by the
Erasmus+ Programme
of the European Union



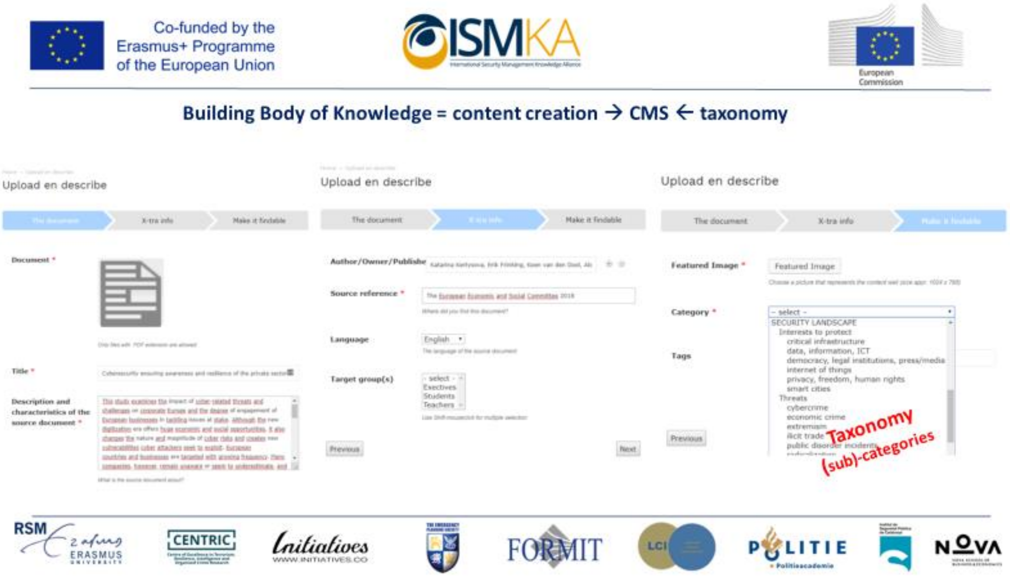
(semi) automated scanning by "Pan Oramix"
after validation part of trend analysis

PAN ORAMIX

Me collecting relevant information on the internet? Probably it's just a waste of money, brains and time, but..... Hopefully in your case one day it 'll lead to some kind of wisdom. Nevertheless..... I can always put you in the "Kettle of Wisdom", boil you for appr. 3 days and see what 'll happen.... your choice. Yours faithfully, Pan Oramix

Figure 4 (semi) automated scanning by Pan Oramix module

This (semi) automated search and posting is combined with the manual posting / input of content in de CMS. In this way a content base or body of knowledge is growing over time.



Building Body of Knowledge = content creation → CMS ← taxonomy

Upload en describe

The document | X-tra info | Make it findable

Document *

Title *

Description and characteristics of the source document *

Upload en describe

The document | X-tra info | Make it findable

Author/Owner/Publishe

Source reference *

Language: English

Target group(s): - select - Exectives Students Teachers

Previous | Next

Upload en describe

The document | X-tra info | Make it findable

Featured Image *: Featured Image

Category *: - select -

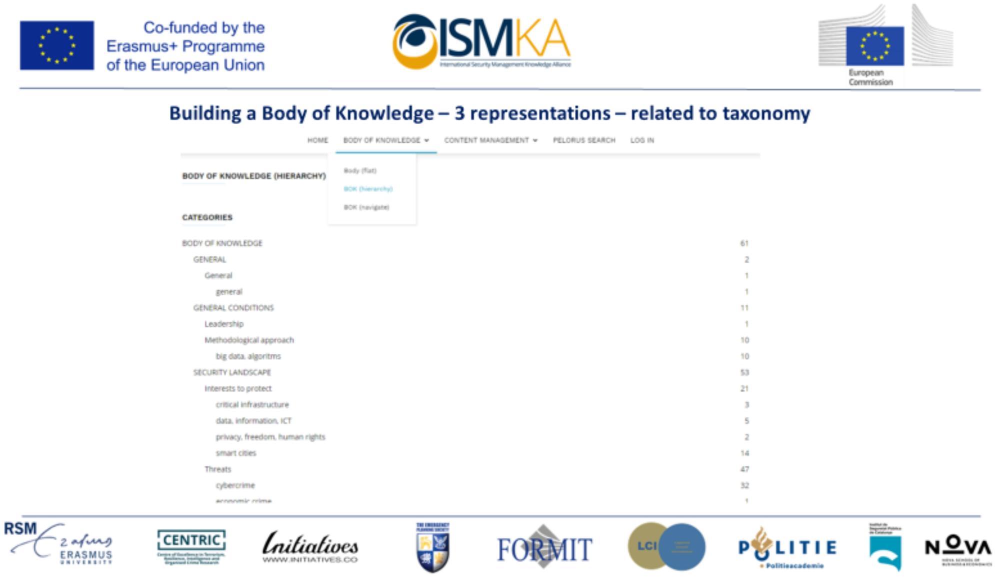
SECURITY LANDSCAPE
  Interests to protect
    critical infrastructure
    data, information, ICT
    democracy, legal institutions, press/media
    internet of things
    privacy, freedom, human rights
    smart cities
  Threats
    cybercrime
    economic crime
    extremism
    illicit trade
    public disorder incidents

Tags

Previous

Figure 5 Content Management System

After selection, evaluation and categorizing the content / knowledge is presented via the Body of Knowledge tab in the tool. It can present it in three ways: flat, as a hierarchy related to the categories and taxonomy and as navigation screen.



Building a Body of Knowledge – 3 representations – related to taxonomy

BODY OF KNOWLEDGE (HIERARCHY)

CATEGORIES

BODY OF KNOWLEDGE	61
GENERAL	2
General	1
general	1
GENERAL CONDITIONS	11
Leadership	1
Methodological approach	10
big data, algoritms	10
SECURITY LANDSCAPE	53
Interests to protect	21
critical infrastructure	3
data, information, ICT	5
privacy, freedom, human rights	2
smart cities	14
Threats	47
cybercrime	32
economic crime	1

Figure 6 Body of Knowledge