

# Text & Data Mining for the National Library of Greece in consideration of GDPR

Dr. Marinos Papadopoulos  
Attorney-at-Law

Mr. Konstantinos Vavousis, M.Sc.  
IT Security expert

Mrs. Dimitra Hioti, M.A.  
Information Scientist, Librarian



# NLG & TDM in Greece

- Law 4452/2017 art.4(4)(b)
- Regulation 2016/679/EU (GDPR)
- Directive 2019/790/EU (Directive on Copyright in the DSM)
- NATIONAL LIBRARY OF GREECE (NLG) deployed TDM since FEB.2017
- Harvesting the .GR, .EDU, & .COM of the Web
- 18 TB of information

## Working Stages Of Web Archiving In Greece By NLG

<b>Stage I</b>	Economic and technical study on the needs and content of the Greek web harvest. Study of international experience	1 <sup>st</sup> web harvest: broad crawl – national level: text data only
<b>Stage II</b>	Definition of "Greek" sites to be mined	
<b>Stage III</b>	Data Analysis of 1st web harvest to create a National Web Archiving System	
<b>Stage IV</b>	Installing and checking the operation of tools for all phases of national web archiving: extraction, archiving / classification and finally, user search and access): Heritrix for harvesting, Solr for indexing and Open Wayback for web site reconstitution. Netarchive Suite using.	2 <sup>nd</sup> web harvest: broad - national level: text only) thematic (text and images)
<b>Stage V</b>	Developing a National Archiving System of Greek Web ("ΕΣΑΕΙ"): the Greek user/librarian interface	

# Harvesting

- Bulk harvesting of works in the Greek Web.
- Selective harvesting leveraging on criteria:
  1. Subject/Topic
  2. Creator/Provenance
  3. Type/Format
- The NLG Curator tool allows for either domain or selective harvests according to 6 predefined schedules.

# Preservation & Access

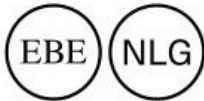
- Copies of harvested works are preserved in different servers in different buildings of NLG.
- Users can access works leveraging on the NLG Curator tool.

# ΕΣΑΕΙ WEB ARCHIVE

Greek | English



Definitions  
Harvest status  
Harvest Channels  
Bitpreservation  
Systemstate



Welcome to **National Archiving System of Greek Web (NASGW)**

NASGW is the national archive for the data of the greek web and contains web pages under the .gr domain and web pages from other domains (e.g. .com, .edu etc) that were written by using the greek language. With NASGW you can search for a web page by URL, category or by using keywords and then you can choose between the available snapshots of the available crawled versions.

For the time being the search process is limited due to legal restrictions into a small number of computers that belong to the National Library of Greece. The content is governed by applicable copyright and personal data protection laws and therefore reproducing, downloading or broadcasting the content is forbidden without written permission of the content creator.

NASGW is a pilot program and the available collection contains crawls of three topic categories: news reports, local government and education. The crawling procedure started at December 2017. The massive web crawling of the greek web (the whole .gr domain) and the crawling of these topic categories requires significant amounts of computer resources.

The main target is to gradually crawl the greek web in depth (including multimedia content). The System was developed by the Data Mining research team of the Athens University of Economics and Business under the supervision of professor Michalis Vazirgiannis with lead developer Mr. Polykarpos Meladianos. The research and the development of the System started at September 2016 until December 2017.

The project was funded by the Stavros Niarchos Foundation in the context of the Resettlement Plan of the National Library of Greece into the cultural center of Stavros Niarchos Foundation 2015-2017, during its 'Action 2: "Digital Services NLG"'.

The development of the System is based on well-known technologies and open source software, like Heritrix for crawling, Solr for indexing and Wayback Machine for searching and displaying the web pages.

For more information you can contact with [webarchive@nlg.gr](mailto:webarchive@nlg.gr)

Εθνικό Σύστημα Αρχειοθέτησης Ελληνικού Ιστού - Εθνική Βιβλιοθήκη της Ελλάδος





# ΕΣΑΕΙ WEB ARCHIVE

Greek | English



## Definitions

- Selective Harvests
- Snapshot Harvests
- Schedules
- Find Domain(s)
- Create Domain
- Domain Statistics
- Edit Harvest Templates
- Global Crawler Traps
- Harvest status
- Harvest Channels
- Bitpreservation
- Systemstate

## Find Domain(s)

Enter domain query

Search domains by

Ημερομηνία Συλλογής

Από:  (format: DD/MM YYYY hh:mm)

Έως:  (format: DD/MM YYYY hh:mm)

Εθνικό Σύστημα Αρχειοθέτησης Ελληνικού Ιστού - Εθνική Βιβλιοθήκη της Ελλάδος



# ΕΣΑΕΙ WEB ARCHIVE

Greek | English



## Definitions

- Selective Harvests
- Snapshot Harvests
- Schedules
- Find Domain(s)
- Create Domain
- Domain Statistics
- Edit Harvest Templates
- Global Crawler Traps
- Harvest status
- Harvest Channels
- Bitpreservation
- Systemstate

## Find Domain(s)

Enter domain query

Search domains by

Ημερομηνία Συλλογής

Από:  (format: DD/MM YYYY hh:mm)

Έως:  (format: DD/MM YYYY hh:mm)

December, 2018						
?	< <	Today	> >	>>	>>>	×
wk	Mon	Tue	Wed	Thu	Fri	Sat Sun
48						1 2
49	3	4	5	6	7	8 9
50	10	11	12	13	14	15 16
51	17	18	19	20	21	22 23
52	24	25	26	27	28	29 30
1	31					
Time: 08 : 09						
Select date						

Εθνικό Σύστημα Αρχειοθέτησης Ελληνικού Ιστού - Εθνική Βιβλιοθήκη της Ελλάδος



# GDPR vs SECURITY?

- Digital environments
- Strong Security Mechanisms

# R.O.S.I.

- Minimize investments
- Heavy fines by GDPR

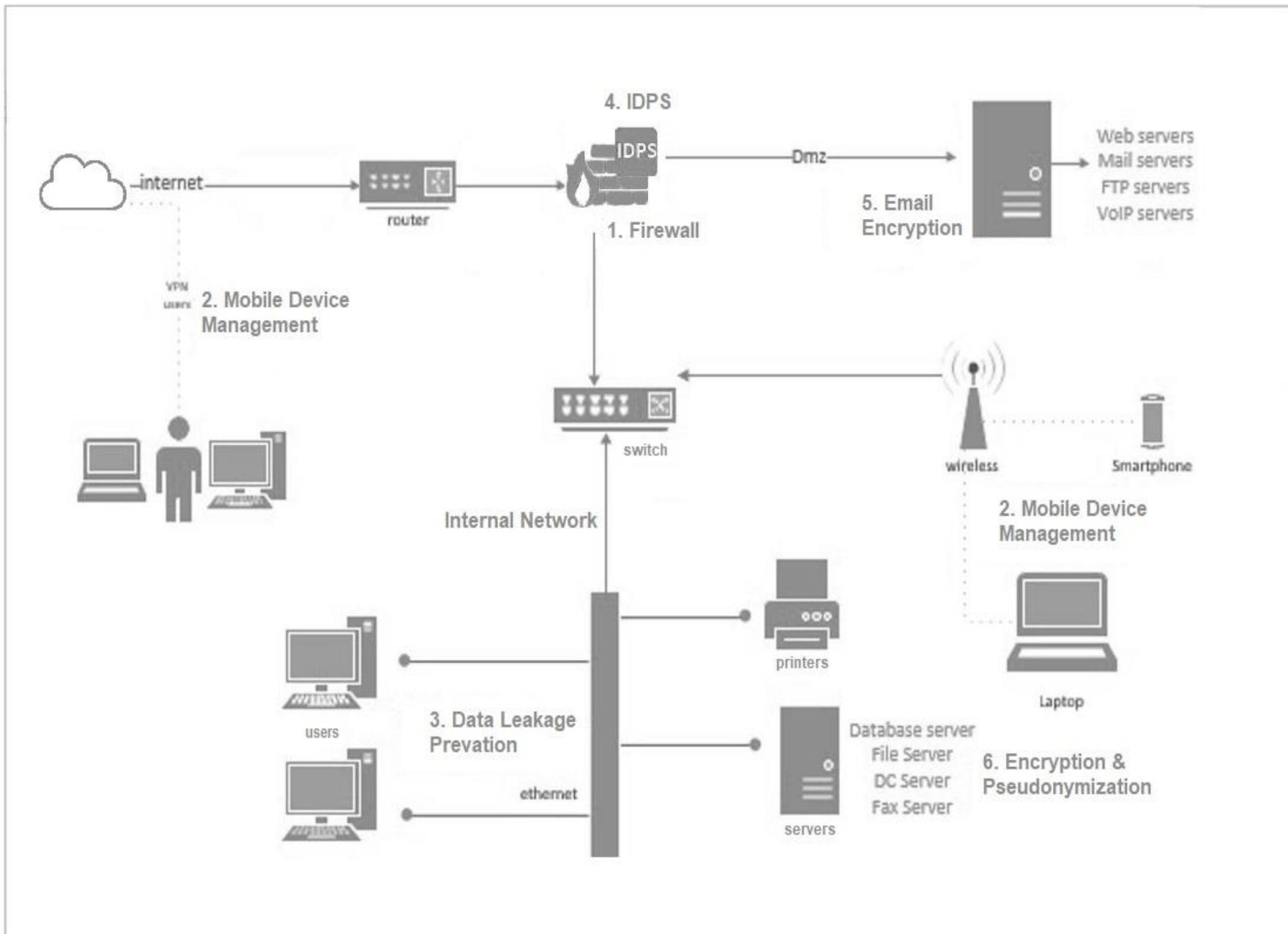
# Strong Security Mechanisms

- Functions (Rec. 78, 108, 119)
- Processes (Rec.63, Art.4, 11, 28)
- Controls (Rec.37)
- Systems (Rec.49, 105, 151)
- Procedures (Rec.71, 88, Art.6, 12, 40, 41, 43, 47, 70)
- Policies (Rec.78, Art.4, 24, 39)

← Personal and critical data protection

# Optimal Infrastructure

- Firewall
- DLP
- MDM
- IDPS
- Email ENCRYPTION
- Database ENCRYPTION & PSEUDONYMIZATION



# Authentication & Authorization

- Strong access control policies



# Epilogue

- CYBER-SECURITY is top priority for GDPR

# Text & Data Mining for the National Library of Greece in consideration of GDPR

c



Dr. Marinos Papadopoulos Attorney-at-Law	Mr. Konstantinos Vavousis, M.Sc. IT Security Expert Trust-IT Ltd.
Dr. Charalampos Bratsas Mathematician Aristotle University of Thessaloniki	Mrs. Eliza Makridou, M.Phil. Information Scientist, Librarian National Library of Greece
Dr. Michalis Gerolimos Information Scientist, Librarian National Library of Greece	Mrs. Dimitra Hioti, M.A. Information Scientist, Librarian National Library of Greece



Thank  
you