



e-ΧΝΗΛΑΤΗΣ

εξυπνη ΧωροχροΝική ολοκλήρωση περιβαλλοντικών πολιτιστικών και
τουριστικών διαδρομών περιήγΗΣης
«ΕΡΕΥΝΩ – ΔΗΜΙΟΥΡΓΩ – ΚΑΙΝΟΤΟΜΩ»

Π2.1.1

Τεχνικές αυτόματης αναζήτησης και εξόρυξης γνώσης από κοινωνικά δίκτυα και τον ιστό

Επίπεδο διάδοσης:	Δημόσιο
Ημερομηνία παράδοσης σύμφωνα με το έργο:	Μήνας 21, 5/6/2020
Πραγματική ημερομηνία παράδοσης:	Μήνας 21, 5/6/2020
Ενότητα εργασίας:	ΕΕ2 Ανάπτυξη τεχνολογιών χωρο-χρονικής διασύνδεσης πολλαπλών χώρων και γεγονότων περιβαλλοντικού, πολιτιστικού και τουριστικού ενδιαφέροντος
Υποενότητα:	ΥΕΕ2.1 Αυτόματη αναζήτηση και γνώση από κοινωνικά δίκτυα και το διαδίκτυο σε πραγματικό χρόνο
Είδος:	Έκθεση + Λογισμικό
Κατάσταση έγκρισης:	Εγκεκριμένο
Έκδοση:	1.0
Αριθμός σελίδων:	60
Όνομα αρχείου:	Π2.1.1_final.docx

Περίληψη

Αυτό το παραδοτέο περιγράφει τη σχετική βιβλιογραφία σχετικά με την ανίχνευση και ανάκτηση πληροφορίας με τις διαδικασίες των crawling και scrapping από το διαδίκτυο καθώς και την προσέγγιση που ακολουθήθηκε και υλοποιήθηκε. Επίσης, περιγράφονται τα είδη των διαδικτυακών πηγών και οι στοχευμένες πηγές από τις οποίες αντλήθηκε περιεχόμενο στα πλαίσια του έργου. Επιπλέον περιγράφονται εκτενώς τα μοντέλα δεδομένων τα οποία υιοθετήθηκαν/διαμορφώθηκαν για τις ανάγκες του e-ΧΝΗΛΑΤΗ καθώς και τα σύνολα δεδομένων τα οποία δημιουργήθηκαν. Το παραδοτέο κλείνει με τα συμπεράσματα που εκμαιεύθηκαν ως τώρα και τα επόμενα εκτιμώμενα βήματα και ακολουθούν παραρτήματα με τις αναφορές, κάποια παραδείγματα πηγαίου κώδικα και αποκρίσεις APIs από χρήσιμες πλατφόρμες.



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο
Περιφερειακής Ανάπτυξης



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ
ΟΙΚΟΝΟΜΙΑΣ & ΑΝΑΠΤΥΞΗΣ
ΕΙΔΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΕΠΠΑ & ΤΣ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ ΕΠΑΝΕΚ



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ Παιδείας,
Ερευνας και Θρησκευμάτων



ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ
ΕΡΕΥΝΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ



ΕΥΜΕ
ΕΤΟΚ



ΕΠΑΝΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ
ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ
ΚΑΙΝΟΤΟΜΙΑ



ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Ιστορικό

Έκδοση	Ημερομηνία	Αιτιολογία	Αναθεωρήθηκε από Εγκρίθηκε από
V0.1	30/4/2020	Πίνακας περιεχομένων	Ευάγγελος Σταθόπουλος Αλέξανδρος Κοκκάλας
V0.3	5/5/2020	Ανάλυση τεχνολογιών αιχμής	Ευάγγελος Σταθόπουλος
V0.8	25/5/2020	Κυκλοφορία πρώτης έκδοσης για σχόλια	Ευάγγελος Σταθόπουλος
V0.9	30/5/2020	Εσωτερική αναθεώρηση	Αλέξανδρος Κοκκάλας
V0.99	2/6/2020	Ενσωμάτωση σχολίων	Ευάγγελος Σταθόπουλος
V1.0	4/6/2020	Ετοιμασία τελικής έκδοσης	Ευάγγελος Σταθόπουλος

Λίστα συγγραφέων

Οργανισμός	Όνομα	Στοιχεία επικοινωνίας
ΕΚΕΤΑ	Ευάγγελος Σταθόπουλος	estathop@iti.gr
ΕΚΕΤΑ	Αλέξανδρος Κοκκάλας	akokkalas@iti.gr

Περίληψη κυριότερων σημείων

Το παρόν παραδοτέο εμπεριέχει και αναλύει κυρίως τις εργασίες που πραγματοποιήθηκαν στο πλαίσιο της υποενότητας εργασίας ΥΕ2.1, τις σχετικές με την ανάπτυξη εργαλείων για τη διεκπεραίωση των διαδικασιών της ανάκτησης πληροφορίας από το διαδίκτυο και τη διαμόρφωση συνόλων δεδομένων προς εξυπηρέτηση των σκοπών του e-ΧΝΗΛΑΤΗ.

Πιο συγκεκριμένα, παρουσιάζεται μια αναλυτική αναφορά των τεχνολογιών αιχμής στον κλάδο της ανάκτησης πληροφορίας, η οποία έθεσε τις βάσεις για τη διαμόρφωση και ανάπτυξη των εργαλείων τα οποία και χρησιμοποιήθηκαν εκτενώς στο έργο. Το εργαλείο επιτρέπει στους χρήστες του να εξάγουν δεδομένα σχετικά με σημεία ενδιαφέροντος από ετερογενείς πηγές ιστού με ημι-αυτόματο τρόπο, ορίζοντας μονάχα ένα αρχείο διαμόρφωσης ανά πηγή δεδομένων. Περιγράφεται αναλυτικά το εν λόγω αρχείο διαμόρφωσης καθώς και η δυνατότητα υποστήριξης πλοήγησης στην εκάστοτε ιστοσελίδα καθώς και λοιπές τεχνικές λεπτομέρειες που κρίνονται ως χρήσιμες για τον αναγνώστη του παραδοτέου.

Στη συνέχεια, γίνεται αναφορά στον τύπο των διαδικτυακών πηγών από όπου αντλήθηκαν πληροφορίες, οι οποίες είναι είτε ιστοσελίδες με επικυρωμένο περιεχόμενο από επίσημους φορείς, είτε ελεύθερα προσβάσιμες γεωβάσεις με έγκυρο γεωχωρικό περιεχόμενο, είτε ελεύθερα προσβάσιμα APIs από μεγάλους οργανισμούς.

Έπειτα, παρουσιάζονται τα μοντέλα δεδομένων τα οποία υιοθετήθηκαν για να ενθυλακώσουν την εξόρυξη γνώσης σε σύνολα δεδομένων από το διαδίκτυο είτε δομικά σε πλούσιες και χρήσιμες μορφές (GEOSON), είτε σημασιολογικά για καλύτερη αναπαράσταση γνώσης και περαιτέρω εφαρμογή κανόνων συλλογιστικής για εξόρυξη υποβόσκουσας γνώσης (Οντολογίες OWL/RDF). Τα σύνολα δεδομένων που ακολουθούν είναι παραδείγματα άντλησης και ενσωμάτωσης πληροφοριών από 3 μεγάλες ιστοσελίδες συναφείς με το αντικείμενο του e-ΧΝΗΛΑΤΗ.

Εν κατακλείδι, αναφέρονται ρητά τα συμπεράσματα που εξήχθησαν κατά τις εργασίες της υποενότητας ΥΕ2.1 καθώς και μια μελέτη για τα μελλοντικά βήματα. Το παραδοτέο κλείνει με τις αναφορές και τα παραρτήματα που απεικονίζουν παραδείγματα πηγαίου κώδικα και αποκρίσεις των εξωγενών APIs. Πιο προηγμένες τεχνικές αυτόματης αναζήτησης και εξόρυξης γνώσης θα αντιμετωπιστούν σε μελλοντικές εκδόσεις του πλαισίου και θα αναφερθούν σε μελλοντικό παραδοτέο (Π.2.1.2, M46).

Συντομογραφίες και ακρωνύμια

API	Application Programming Interface
CSS	Cascading Style Sheets
DOM	Document Object Model
HTML	Hyper Text Markup Language
OWL	Web Ontology Language
REGEX	Regular Expressions
REST	Representational State Transfer
SHACL	Shapes Constraint Language
SPARQL	SPARQL Protocol and RDF Query Language
UML	Unified Modelling Language
URL	Uniform Resource Locator
AMEA	Άτομα με ειδικές ανάγκες

Πίνακας περιεχομένων

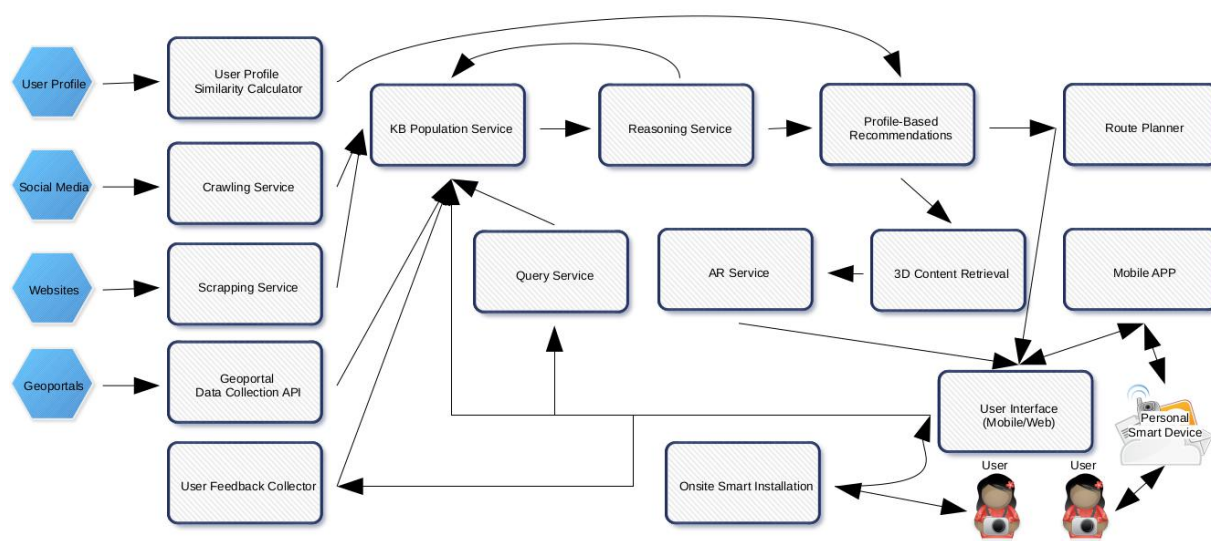
Εισαγωγή	8
Σχετικές τεχνικές απαιτήσεις	10
γενικό πλαίσιο	11
Ανακτηση πληροφορίας	13
Ανίχνευση & Ανάκτηση με CRAWLING	13
Σχετικές εργασίες	13
Ανάκτηση πληροφορίας με SCRAPING	14
Σχετικές εργασίες	15
Προσέγγιση & Υλοποίηση	15
Περιγραφή αρχείου διαμόρφωσης	16
Υποστήριξη πλοήγησης	21
Τεχνικές λεπτομέρειες	22
Αναζήτηση σε WEB & κοινωνικές πλατφόρμες	26
Ιστοσελίδες	26
Γεωβάσεις	30
APIs	31
TICKETMASTER API	31
Περιγραφή Πόρων	32
API & TOKENS	32
EVENT SEARCH API	32
Παράμετροι ερωτήματος αναζήτησης	32
OPEN WEATHER API	36
Περιγραφή Πόρων	37
API & TOKENS	37
ONE CALL API	37
Παράμετροι ερωτήματος αναζήτησης	37
Μοντέλο δεδομένων	38

ΓΕΟJSON	39
Οντολογία	41
Σύνολα δεδομένων	42
Σύνολα δεδομένων από ΟΔΥΣΣΕΥΣ	43
Σύνολα δεδομένων από RELIGIOUS GREECE	43
Σύνολα δεδομένων από VERY MACEDONIA	45
Συμπεράσματα & επομενα βήματα	47
αναφορές	48
Παράρτημα	49

1 Εισαγωγή

Η τεράστια ποσότητα δεδομένων, δομημένη και μη, όπου άτομα, επιχειρήσεις και κυβερνήσεις συνεχίζουν να δημιουργούν με άνευ προηγουμένου ρυθμό και τα προβλήματα χρηστικότητας που προκύπτουν από την προσπάθεια αποθήκευσης και διαχείρισης αυτών των δεδομένων, καθιστούν την ανάλυση και την ευρετηρίαση δεδομένων που προέρχονται από πλατφόρμες κοινωνικής δικτύωσης και ιστότοπους μια πολύ δύσκολη εργασία. Η πληθώρα πληροφοριών και τα αυξημένα επίπεδα θορύβου σε αυτά απαιτούν ευφυείς αλγόριθμους ανίχνευσης και υλοποιήσεις αντίστοιχα. Με συνεχώς επεκτεινόμενα δεδομένα χωρισμένα σε διαφορετικές μορφές, πολλαπλούς κωδικούς και γλώσσες και διάφορες κατηγορίες, που διασυνδέονται χωρίς ιδιαίτερη τάξη, ευφυή προγράμματα ανίχνευσης ιστού προσαρμοσμένα στις ανάγκες του έργου και των χρηστών είναι απαραίτητα για την αντιμετώπιση των προαναφερθέντων προκλήσεων.

Για να αντιμετωπίσουμε αυτό το ζήτημα, στον e-ΧΝΗΛΑΤΗ αναπτύσσουμε crawlers που χρησιμοποιούν φίλτρα μόνο για τη συλλογή υποσυνόλων ενδιαφέροντος των σχετικών πληροφοριών. Τέτοια φίλτρα σχετίζονται με συγκεκριμένα μεταδεδομένα για οντότητες που μας ενδιαφέρουν, όπως για παράδειγμα γεωσυντεταγμένες, ωράρια λειτουργίας, περιγραφές και άλλα. Η Εικόνα 1 απεικονίζει την αρχιτεκτονική του e-ΧΝΗΛΑΤΗ. Αυτό το έγγραφο αναφέρεται στο αριστερό μέρος της συλλογής δεδομένων, και πιο συγκεκριμένα τα κάτω 3 μπλε πολύγωνα και τις αμέσως επόμενες υπηρεσίες που τροφοδοτούν.



Εικόνα 1: Αρχιτεκτονική του e-ΧΝΗΛΑΤΗ

Τρεις τύποι δεδομένων ανιχνεύονται και ανακτώνται στα πλαίσια του e-ΧΝΗΛΑΤΗ: προσφερόμενα ελεύθερα APIs από κοινωνικές πλατφόρμες που σερβίρουν χρήσιμα δεδομένα και μεταδεδομένα (όπως για παράδειγμα πλατφόρμες με εκδηλώσεις ή πρόγνωση καιρικών φαινομένων), επίσημες στοχευμένες ιστοσελίδες από έγκυρους φορείς με πλούσιο περιεχόμενο, όπως επίσης και γεωπύλες από αντίστοιχα έγκυρους φορείς.

Όλα τα εργαλεία που αναπτύχθηκαν και χρησιμοποιήθηκαν για την εξυπηρέτηση των αναγκών του e-ΧΝΗΛΑΤΗ εξάγουν ελεύθερα διαθέσιμο περιεχόμενο κειμένου από ανοιχτές διαδικτυακές πηγές. Τα εργαλεία αυτά βασίζονται σε προϋπάρχουσες τεχνολογίες crawlers και scrappers ανοιχτού κώδικα για τη συλλογή περιεχομένου βάσει ημι-αυτόματης σύνταξης κατάλληλων ερωτημάτων. Για τις ιστοσελίδες συγκεκριμένα, η διαδικασία του scraping εκτελείται για λήψη μόνο κειμενικού περιεχομένου με νόημα από την αφθονία πληροφοριών που μπορεί να ανιχνευθεί στο διαδίκτυο. Είναι σημαντικό να διευκρινιστεί ότι για scrapping ιστοσελίδων δεν διενεργείται αυτόματα πρώτα κάποια ιχνηλάτηση ή αναζήτηση πόρων αλλά εξάγεται απευθείας η πληροφορία με τη σύνταξη των κατάλληλων αρχείων διαμόρφωσης.

Το συλλεγόμενο υλικό αποθηκεύεται και χρησιμοποιείται από τις υπόλοιπες λειτουργικές μονάδες του έργου. Πιο συγκεκριμένα, εξυπηρετούνται: η χωρική αναπαράσταση περιεχομένου και ο σημασιολογικός συσχετισμός για τη δημιουργία διαδρομών (ΥΕ 2.3) και οι τεχνικές προσωποποίησης εμπειρίας χρήστη (ΥΕ 2.4).

2 Σχετικές τεχνικές απαιτήσεις

Αυτή η ενότητα παρουσιάζει την πλατφόρμα εφαρμογής του έργου, αναλύοντας τις σχετικές τεχνικές απαιτήσεις που καθοδηγούν την ανάπτυξη της ανάκτησης πληροφορίας από διάφορες πηγές στο διαδίκτυο. Τέλος, οι παρακάτω πίνακες παρουσιάζουν τις σχετικές συναρτήσεις παρέχοντας ένα επιθυμητό επίπεδο λεπτομέρειας.

Συνάρτηση	Εισερχόμενα δεδομένα	Εξερχόμενα Δεδομένα
Crawling Service	API καιρού, Events API	μεταδεδομένα καιρού, μεταδεδομένα events
Scrapping Service	CSS Selectors από websites	Κείμενο ιστοσελίδων με μεταδεδομένα
Geoportal Data Collection API	Geodatabases exports	geojson με μεταδεδομένα

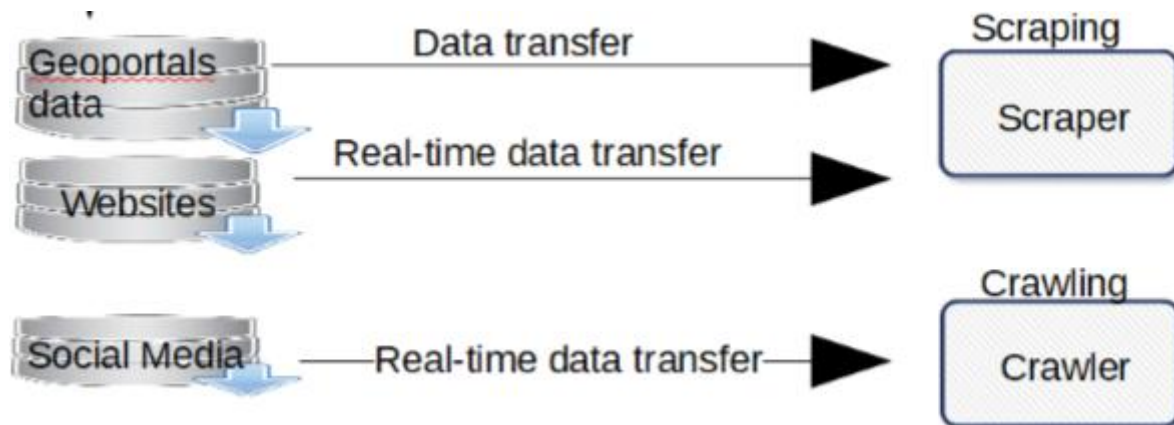
Πίνακας 1: Συναρτήσεις Ανάκτησης Πληροφορίας

Ενότητα Εργασίας	Περιγραφή	Εξαρτήσεις
ΥΕ 2.1	Συλλογή από κοινωνικές πλατφόρμες και αποθήκευση σε βάση	Τροφοδοτεί την ΥΕ 2.2
ΥΕ 2.1	Συλλογή από επίσημες ιστοσελίδες και αποθήκευση σε βάση	Τροφοδοτεί την ΥΕ 2.2
ΥΕ 2.6	Συλλογή δεδομένων από επίσημες γεωπύλες και αποθήκευση σε βάση	Τροφοδοτεί την ΥΕ 2.2

Πίνακας 2: Συσχέτιση με ενότητες εργασίας

3 γενικό πλαίσιο

Ο σκοπός και οι στόχοι των υποσυστημάτων ανάκτησης πληροφορίας που αναπτύχθηκαν για το έργο είναι η συλλογή χρήσιμων πληροφοριών και μεταδεδομένων για την άμεση ή έμμεση εξυπηρέτηση των μόντουλων των ενοτήτων εργασίας ΕΕ2 και ΕΕ3. Πρακτικά πρόκειται για το πρώτο βήμα προ-επεξεργασίας μιας σειράς διαδικασιών που ενισχύει και υποβοηθάει το σύστημα προς ευφύτεστερες παραγωγές προτεινόμενων διαδρομών στην πλατφόρμα. Στην εικόνα 2 απεικονίζονται επακριβώς τα μόντουλα που εκτελούνται σε αυτό το γενικό πλαίσιο.



Εικόνα 2: Υποσυστήματα Ανάκτησης Πληροφορίας

Βασιστήκαμε σε υπάρχων περιεχόμενο στο διαδίκτυο, σε κοινωνικές πλατφόρμες και γεωβάσεις. Οι ιστοσελίδες είναι ειδικά επιλεγμένες διευθύνσεις URL από domains δικτύου επίσημων φορέων όπου το περιεχόμενο τους δύναται να εξυπηρετήσει τις ανάγκες του έργου. Οι κοινωνικές πλατφόρμες επί του παρόντος αφορούν κυρίως ανοιχτές υπηρεσίες όπου μπορούν να παρέχουν δεδομένα και μεταδεδομένα στο σύστημα του e-ΧΝΗΛΑΤΗ υπό την μορφή δωρεάν ελεύθερων κλήσεων των διεπαφών προγραμματισμού εφαρμογών (Application Programming Interface calls). Τέλος, οι γεωβάσεις αφορούν ελεύθερα προσβάσιμα σύνολα δεδομένων (datasets), τα οποία παρέχουν χωρίς χρέωση χωρικές πληροφορίες μαζί με ενδεχομένως επιπρόσθετες πληροφορίες που εξυπηρετούν τις ανάγκες του έργου.

Στην περίπτωση των websites όπως και των κοινωνικών μέσων, είναι προγραμματισμένη η μεταφορά δεδομένων να γίνεται σε πραγματικό χρόνο και καθ' απαίτηση, δηλαδή όταν εκτελούνται οι εφαρμογές από το σύστημα του e-ΧΝΗΛΑΤΗ. Αυτό συμβαίνει με αυτόματο τρόπο στην κατανάλωση δεδομένων από κοινωνικές πλατφόρμες και με ημι-αυτόματο τρόπο στην περίπτωση των ιστοσελίδων, κι αυτό γιατί κάθε ιστότοπος έχει διαφορετική δόμηση που εξακολουθώς απαιτεί και χειροκίνητη ad hoc κατασκευή των αρχείων διαχείρισης που απαιτούνται για την εκτέλεση των scrappers. Όσον αφορά τη μεταφορά δεδομένων από τις γεωβάσεις προς τις βάσεις δεδομένων του e-ΧΝΗΛΑΤΗ αυτό λαμβάνει χώρα μία φορά ανά γεωβάση υπό τη διαδικασία εξαγωγής των δεδομένων σε αρχεία τύπου geojson τα οποία εν συνεχεία εισάγονται χειροκίνητα αρχικώς στη γεωχωρική βάση δεδομένων του έργου, όπως επίσης και στη γνωσιακή βάση η οποία επικοινωνεί με τη γεωχωρική βάση και βρίσκονται διαρκώς σε συγχρονισμό.

Η ανάπτυξη των crawlers του e-ΧΝΗΛΑΤΗ περιλαμβάνει ευθέως την υποενότητα εργασίας :

ΥΕ 2.1: Αυτόματη αναζήτηση και γνώση από κοινωνικά δίκτυα και το διαδίκτυο σε πραγματικό χρόνο. Στο γενικό πλαίσιο αυτής της υποενότητας εργασίας, έχουν εντοπιστεί ορισμένες επίσημες ιστοσελίδες με κατάλληλο περιεχόμενο, έχει διεκπεραιωθεί η διαδικασία του scrapping σε αρκετά από αυτά και καθώς γράφεται το παρόν παραδοτέο έχουν εισαχθεί τα δεδομένα τουλάχιστον ενός εξ αυτών. Επίσης, έχει γίνει εντοπισμός και επιλογή ανοιχτά προσπελάσιμων γεωβάσεων από τις οποίες μπορούν να αντληθούν χρήσιμες γεωχωρικές πληροφορίες, έχουν εξαχθεί τα κατάλληλα αρχεία κι εκκρεμεί η εισαγωγή των δεδομένων στις βάσεις του έργου. Επιπλέον, έχει γίνει η διερεύνηση και επιλογή των κοινωνικών πλατφόρμων που θα αντλήσουμε δεδομένα, τόσο για προγραμματισμένα γεγονότα όσο και για καιρικές προγνώσεις.

Τέλος, το παρόν παραδοτέο περιγράφει την αρχική υλοποίηση του προαναφερθέντος γενικού πλαισίου. Το τελικό γενικό πλαίσιο θα περιγραφεί εκτενέστερα στο Π2.1.2 κατά τον 46ο μήνα του έργου.

4 Ανακτηση πληροφοριας

Το διαδίκτυο εμπεριέχει τεράστιες ποσότητες δεδομένων που περιλαμβάνουν γεωχωρικές πληροφορίες για σημεία ενδιαφέροντος με μεταδεδομένα, γεγονότα και ζωντανές καιρικές προγνώσεις που μπορούν να χρησιμοποιηθούν στο γενικό πλαίσιο του έργου. Αυτά τα δεδομένα μπορούν να βρεθούν σε διαφορετικών τύπων διαδικτυακούς πόρους όπως ιστοσελίδες, κοινωνικές πλατφόρμες και λοιποί ιστότοποι που σερβίρουν πληροφορίες με APIs. Τα προαναφερθέντα αντικείμενα είναι υψηλού ενδιαφέροντος μιας και χρησιμοποιούνται στην ανακάλυψη σημείων ενδιαφέροντος στο χάρτη όπως επίσης και στους αλγορίθμους δρομολόγησης, και παρέχουν μεταδεδομένα που συμπεριλαμβάνονται στους αλγορίθμους προσωποποίησης.

4.1 Ανίχνευση & Ανάκτηση με Crawling

Στον e-ΧΝΗΛΑΤΗ, παρουσιάζεται η ανάγκη για ανίχνευση διαδικτυακών δεδομένων που σχετίζονται με τα σενάρια χρήσης με σκοπό τη εναπόθεση αυτών στις βάσεις δεδομένων του έργου. Αναπτύχθηκαν τα κατάλληλα εργαλεία για την παραλαβή δεδομένων από τέτοιου είδους κοινωνικές πλατφόρμες, που ουσιαστικά υπό την κατάλληλη μορφή ερωτημάτων, λαμβάνεται η επιθυμητή στοχευμένη απόκριση.

4.1.1 Σχετικές εργασίες

Οι διαδικτυακοί crawlers, γνωστοί και ως “spiders”, “bots” ή “wanderers”, είναι προγράμματα λογισμικού τα οποία αυτόματα επισκέπτονται το διαδίκτυο με έναν μεθοδικό, αυτόματο τρόπο και ανακαλύπτουν διαδικτυακούς πόρους.

Οι διαδικτυακοί crawlers βλέπουν το διαδίκτυο ως έναν κατευθυνόμενο γράφο, όπου οι κόμβοι αναπαριστούν μοναδικές ιστοσελίδες (βασισμένες στο URL) και κάθε κατευθυνόμενη ακμή αναπαριστά έναν μοναδικό υπερσύνδεσμο ανάμεσα σε δύο σελίδες. Οι crawlers είναι ικανοί να διασχίσουν όλο το διαδικτυακό γράφο με διάφορους τρόπους, όπως με breadth-first ή depth-first τρόπο, δοσμένης της σειράς των υπερσυνδέσμων μέσα στις ιστοσελίδες ή με best-first τρόπο βασισμένο σε κάποια κατάταξη, όπως ο αριθμός των εισερχομένων συνδέσμων (Olston & Najork, 2010).

Οι διαδικτυακοί crawlers εκκινούν με μια προκαθορισμένη λίστα από seed URLs, φέρνουν και αναλύουν το καθένα από αυτά, εξάγουν υπερσυνδέσμους που αυτές οι σελίδες εμπεριέχουν, τοποθετούν τους εξαγμένους υπερσυνδέσμους σε μια ουρά και, συστηματικά, φέρνουν και αναλύουν URLs συσχετισμένα με αυτούς τους υπερσυνδέσμους από την ουρά και ούτω καθεξής. Αυτή η διαδικασία διαδοχικά επαναλαμβάνεται σε αυθαίρετο βάθος, εξαρτώμενο από το στόχο ο οποίος μπορεί να είναι μέχρι ένας ικανοποιητικός αριθμός σελίδων ανιχνεύεται.

Η συμπεριφορά ενός διαδικτυακού crawler δεσμεύεται από έναν συνδυασμό πολιτικών:

- Η πολιτική επιλογής που καθορίζει τις σελίδες,
- η πολιτική επανεπίσκεψης που καθορίζει πόσο συχνά θα έπρεπε μια σελίδα να ελέγχεται για αλλαγές στο περιεχόμενο της,

- η πολιτική διπλότυπων που εξετάζει τις ομοιότητες στο περιεχόμενο ιστοσελίδων,
- η πολιτική ευγένειας που παρέχει κανόνες για την αποφυγή υπερφόρτωσης μιας ιστοσελίδας ή μιας ιστοβάσης,
- η πολιτική παραλληλοποίησης όπου συγκεκριμενοποιεί τον τρόπο που θα συντονιστούν καταναεμημένοι διαδικτυακοί crawlers,
- η πολιτική απορρήτου της κάθε ιστοσελίδας που καταδεικνύει ποιες περιοχές δεν θα έπρεπε να αναλυθούν ή να ανακτηθούν.

Σήμερα, υπάρχουν μερικοί καταξιωμένοι διαδικτυακοί crawlers. Ανάμεσα τους ξεχωρίζουν οι Heritrix, Apache Nutch και crawler4j ως τα πιο ευρέως χρησιμοποιούμενα ανοιχτού κώδικα λογισμικά. Το Heritrix είναι ένα ανοιχτού κώδικα, επεκτάσιμο, διαδικτυακής κλίμακας έργο που υποστηρίζεται από μια μεγάλη κοινότητα χρηστών. Ο καλά τεκμηριωμένος κώδικας του και η εύκολη χρήση της διεπαφής χρήστη το καθιστούν περισσότερο ως μια ελαφριά λύση για γενική χρήση. Το Apache Nutch είναι ένας πιο εύρωστος, μεγάλης κλίμακας και συγκρίσιμο με το Google-Bot crawler που είναι εξαιρετικά επεκτάσιμο και ανοιχτού κώδικα. Το crawler4j είναι ένα πρόγραμμα ανίχνευσης ιστού ανοιχτού κώδικα γραμμένο σε Java, το οποίο παρέχει μια απλή διεπαφή για ανίχνευση ιστού. Χρησιμοποιώντας το, μπορεί κανείς να ρυθμίσει γρήγορα ένα πρόγραμμα ανίχνευσης ιστού με πολλά παράλληλα νήματα.

Προσφέρονται και ως εναλλακτικές επιλογές crawling διαδικτυακές εφαρμογές όπως τα Octoparse¹ και import.io², όπου προσφέρουν το πλεονέκτημα ότι δεν απαιτούν προγραμματιστικές δεξιότητες από τον χρήστη μιας και οι λειτουργίες τους υποστηρίζονται μέσω διαδικτυακής διεπαφής χρήστη. Ωστόσο, για να μην βασιστεί ο e-ΧΝΗΛΑΤΗΣ σε υπηρεσίες τρίτων, αναπτύχθηκε ένα πλαίσιο που χρησιμοποιεί βιβλιοθήκες προγραμματισμού ανοιχτού κώδικα.

4.2 Ανάκτηση πληροφορίας με Scrapping

Το scrapping ως όρος αναφέρεται στη διαδικασία αυτόματης εξαγωγής πληροφοριών από πάσα ανθρωπίνως αναγνώσιμη πηγή. Παραδείγματα τέτοιων πηγών αποτελεί κείμενο από ιστοσελίδες. Σε αυτό το έργο ενδιαφερόμαστε κυρίως για scrapping από ιστοσελίδες, το οποίο είναι μια εξειδικευμένη περίπτωση scraping που εφαρμόζεται μόνο σε διαδικτυακούς ιστοτόπους με στόχο να φιλτράρει περιττά στοιχεία όπως μπάρες πλοήγησης και διαφημίσεις. Το διαδικτυακό scrapping σχετίζεται επίσης με την ευρετηρίαση, εστιάζοντας στη μετατροπή μη δομημένων δεδομένων στον ιστό, συνήθως σε HTML μορφή, σε δομημένα δεδομένα που μπορούν να αποθηκευτούν και να αναλυθούν σε μια κεντρική τοπική βάση δεδομένων. Συνήθως οι τεχνικές scrapping εξαρτώνται από τον τύπο του ιστοτόπου (δηλαδή, στατικό ή δυναμικό). Επιπλέον, τέτοιες τεχνικές, μπορεί να εξαρτώνται από τον τρόπο με τον οποίο δομούνται ή παρουσιάζονται οι πληροφορίες και οι scrappers πρέπει να διαμορφωθούν βασισμένα σε αυτά. Για παράδειγμα, ένας scraper ιστού

¹ <https://www.octoparse.com/>

² <https://www.import.io/>

που εξάγει περιεχόμενο από ένα φόρουμ πρέπει να ρυθμιστεί διαφορετικά από έναν scraper που εξάγει περιεχόμενο από άρθρα. Επίσης ο ίδιος scraper ενδέχεται να χρειάζεται διαφορετική διαμόρφωση κατά την εξαγωγή 2 διαφορετικών τομέων από τον ίδιο τύπο ιστοτόπου. (για παράδειγμα 2 διαφορετικά φόρουμς). Στον e-ΧΝΗΛΑΤΗ υπάρχει ανάγκη για εξόρυξη και συγκέντρωση περιεχομένου από τέτοιες ετερογενείς πηγές σε ένα κεντρικό αποθετήριο χρησιμοποιώντας μια κοινή μορφή.

4.2.1 Σχετικές εργασίες

Στα scrapping frameworks η επιλογή στοιχείων σε μια ιστοσελίδα ορίζεται ως λειτουργία επιλογής, ενώ η αποδοχή ή απόρριψη ενός αποτελέσματος επιλογής ορίζεται ως λειτουργία επικύρωσης. Η λειτουργία επιλογής είναι υπεύθυνη για την επιλογή ενός κομματιού πληροφοριών (δηλαδή ενός στοιχείου HTML, μιας λίστας στοιχείων HTML, πλειάδες κειμένων και ούτω καθεξής) από μια δοσμένη ιστοσελίδας και η παράδοση των αποτελεσμάτων επιλογής σε κατάλληλη μορφή. Η λειτουργία επικύρωσης χρησιμοποιεί διαστάσεις κειμένου των αποτελεσμάτων επιλογής μαζί με την αρχική ιστοσελίδα. Ωστόσο, όταν τα στοιχεία που θα επιλεγούν αλλάζουν δυναμικά, άλλες διαστάσεις ενδέχεται να είναι πιο αποτελεσματικές.

Πλέον υπάρχει μια πληθώρα από scrapers οι οποίοι υποστηρίζουν ένα ευρύ φάσμα λειτουργιών και χαρακτηριστικών, όπως το JSoup³, το HTMLUnit⁴, το Import.io⁵, το DEIXTo⁶, το Kimono Labs⁷, το ScraperWiki⁸, το easIE⁹ και το boilerpipe library¹⁰.

Προσέγγιση & Υλοποίηση

Το γενικό πλαίσιο που στηρίχθηκαν οι τεχνολογίες του έργου βασίστηκε σε ανοιχτού κώδικα βιβλιοθήκες και αναπτύχθηκε ανάλογα με τις απαιτήσεις που όφειλαν να καλυφθούν. Πιο συγκεκριμένα, επιτρέπει στους διαχειρίζοντες χρήστες να εξάγουν δεδομένα σχετικά με σημεία ενδιαφέροντος από ετερογενείς πηγές ιστού με ημι-αυτόματο τρόπο, ορίζοντας μονάχα ένα αρχείο διαμόρφωσης ανά πηγή δεδομένων. Μια επισκόπηση του γενικού πλαισίου απεικονίζεται στην εικόνα 3. Ένα αρχείο διαμόρφωσης παρέχεται ως είσοδος, αναλύεται και επικυρώνεται από το πρόγραμμα ανάλυσης διαμόρφωσης. Στο δεύτερο βήμα, ο Web Wrapper Generator δημιουργεί ένα προσαρμοσμένο wrapper για την web πηγή προορισμού, το οποίο στη συνέχεια εφαρμόζεται στη σελίδα από τον wrapper executor. Ως τελικό βήμα, ο διαχειριστής δεδομένων εκτελεί ενοποίηση των δεδομένων συνδυάζοντας εγγραφές που αναφέρονται στο ίδιο σημείο ενδιαφέροντος.

³ <https://jsoup.org>

⁴ <http://htmlunit.sourceforge.net/>

⁵ <https://import.io/advanced--data--platform>

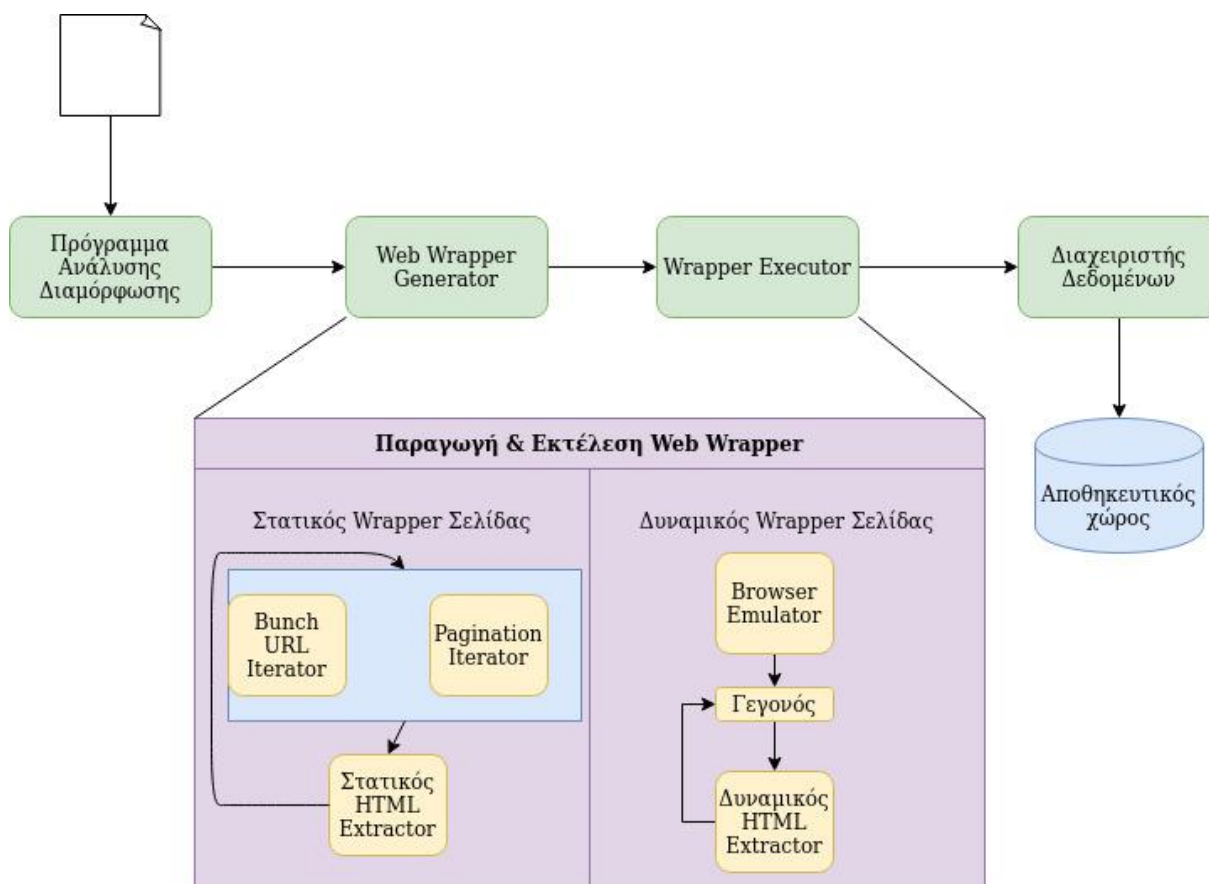
⁶ <http://deixto.com/>

⁷ <https://www.kimonolabs.com>

⁸ <https://scraperwiki.com>

⁹ <https://github.com/MKLab--ITI/easIE>

¹⁰ <https://code.google.com/archive/p/boilerpipe/>



Εικόνα 3: Γενικό Πλαίσιο

Συγκεκριμένα, δημιουργούνται δύο τύποι από wrappers ανάλογα με το περιεχόμενο HTML, αν είναι στατικό ή δυναμικό. Καθένα από αυτά εκτελεί την ακόλουθη ροή επεξεργασίας:

- ανάκτηση σελίδας,
- εξαγωγή δεδομένων,
- μεταεπεξεργασία δεδομένων,
- επανεκκίνηση, δηλαδή προσδιορισμό της επόμενης σελίδας για ανάκτηση από την τρέχουσα ιστοσελίδα.

Σε περίπτωση στατικών σελίδων, ο executor εφαρμόζει τον wrapper σε μια ακολουθία ιστοσελίδων που ορίζεται είτε μέσω μιας λειτουργίας σελιδοποίησης (Pagination Iterator), είτε μέσω μιας ομάδας διευθύνσεων URL (Bunch URL Iterator). Σε περίπτωση δυναμικών σελίδων, ο executor εκκινεί έναν εξομοιωτή προγράμματος περιήγησης (browser emulator) και εκτελεί ένα σύνολο συμβάντων χρήσης (όπως καθορίζονται στο αρχείο διαμόρφωσης), δηλαδή συμβάντα click and scroll, προκειμένου να ληφθούν και να εξαχθούν τα στοχευμένα δεδομένα.

Περιγραφή αρχείου διαμόρφωσης

Το αρχείο διαμόρφωσης ορίζεται σε δομή JSON και αποτελείται από ένα σύνολο από στοιχεία όπως περιγράφονται στον πίνακα 3. Για τον ορθό καθορισμό του

αρχείου διαμόρφωσης, είναι ζωτικής σημασίας οι διαχειριστές να είναι εξοικειωμένοι με τη σύνταξη των CSS selectors και ερωτημάτων (queries). Πιο συγκεκριμένα, οι κανόνες εξαγωγής πρέπει να διαμορφωθούν σε μορφή CSS προκειμένου να προσδιοριστούν τα στοχευμένα στοιχεία του εγγράφου εισαγωγής HTML. Τα μοντέρνα προγράμματα περιήγησης διαθέτουν ενσωματωμένα εργαλεία (DOM Inspectors) που διευκολύνουν την εξαγωγή ενός CCS ή XPath Selector που απευθύνεται σε ένα συγκεκριμένο στοιχείο HTML. Στο αρχείο διαμόρφωσης, τα δηλωμένα στοιχεία καθορίζουν

α) τον τύπο των δεδομένων που πρόκειται να εξαχθούν,

β) τις λεπτομέρειες των βημάτων που περιλαμβάνουν τη ροή επεξεργασίας του wrapper (ανάκτηση, εξαγωγή δεδομένων, μεταεπεξεργασία, επανεκκίνηση).

Πεδίο	Τύπος	Περιγραφή
url	string	URL της στοχευμένης σελίδας
group_of_urls	array of strings	ομάδα από παρόμοια δομημένα URLs προς εξαγωγή δεδομένων
source_name	string	όνομα προέλευσης επιλεγμένο από χρήστη
table_selector	string	CSS Selector που δείχνει στο DOM υποδέντρο όπου βρίσκονται τα δεδομένα ενδιαφέροντος
info	array	πίνακας από JSON στοιχεία που αναλύονται στον πίνακα 2
metrics	array	πίνακας από JSON στοιχεία που αναλύονται στον πίνακα 2
crawl	object	ένα εμφωλευμένο JSON αρχείο διαμόρφωσης που ανιχνεύει μέσα στην ιστοσελίδα
store	json	προδιαγραφές αποθήκευσης
dynamic_page	boolean	προσδιορίζει εάν η σελίδα

		είναι στατική ή δυναμική
--	--	--------------------------

Πίνακας 3: Κύρια στοιχεία του αρχείου διαμόρφωσης

Το σημείο εισόδου για έναν wrapper είναι το URL που καθορίζει την αρχική ιστοσελίδα που θα ανακτήσει πρώτα ο wrapper για να ξεκινήσει η διαδικασία συλλογής δεδομένων για την πηγή με το αντίστοιχο “source_name” (δηλαδή μια ετικέτα που αναφέρεται στη συγκεκριμένη πηγή). Εν συνεχεία, το αρχείο διαμόρφωσης θα πρέπει να καθορίζει τα χαρακτηριστικά

α) που μπορούν να εξαχθούν για κάθε σημείο ενδιαφέροντος (π.χ. όνομα σημείου, ιστότοπος, συντεταγμένες κ.λπ.) μέσω των στοιχείων “info”,

β) των metrics που θα εξαχθούν και θα συσχετιστούν με κάθε σημείο ενδιαφέροντος.

Ο πίνακας 4 περιγράφει τα πρότυπα των πεδίων “info” και “metrics”. Ένα άλλο βασικό στοιχείο διαμόρφωσης είναι το boolean flag του “dynamic_page” που δείχνει εάν η στοχευμένη σελίδα πρέπει να χειρίζεται από τον wrapper ως στατική ή δυναμική. Στην περίπτωση των στατικών σελίδων, τα δεδομένα προορισμού είναι διαθέσιμα κατά τη φόρτωση της σελίδας. Αντί αυτού, για τις δυναμικές σελίδες το αρχείο διαμόρφωσης θα πρέπει επίσης να καθορίζει το σύνολο των συμβάντων χρήστη (user events, click ή scroll) που απαιτούνται για τη φόρτωση των απαραίτητων αποσπασμάτων HTML στον εξομοιωτή του προγράμματος περιήγησης. Ο αλγόριθμός 1 περιγράφει τη ροή επεξεργασίας του wrapper στατικών ιστοσελίδων και ο αλγόριθμος 2 τις δυναμικές ιστοσελίδες με βάση το αρχείο διαμόρφωσης εισόδου.

Δεδομένα: διαμόρφωση

Αποτέλεσμα: extracted_info, extracted_metrics

1 fetch_page (configuration.url);

2 extract_info(page, configuration.info);

(I) εξήγαγε περιεχόμενο από τα ορισμένα στοιχεία

(II) μεταεπεξεργασία του εξαγόμενου περιεχομένου αν το πεδίο αντικατάστασης είναι ορισμένο

3 extract_metrics(page, configuration.metrics);

4 go-to step 1 if:

(I) configuration.group_of_urls είναι ορισμένο

configuration.url = configuration.group_of_urls.next

(II) configuration.next_page_selector είναι ορισμένο

configuration.url = next_page

(III) configuration.crawl είναι ορισμένο

configuration = configuration.crawl

5. επέστρεψε τα extracted_info, extracted_metrics



Αλγόριθμος 1: Wrapper στατικής ιστοσελίδας

Η πραγματική διαδικασία εξαγωγής δεδομένων ξεκινά δείχνοντας στον wrapper το υποδέντρο DOM που καθορίζεται από το πεδίο “table_selector”. Πρόκειται για έναν CSS selector που δείχνει στον wrapper το σύνολο στοιχείων HTML που περιέχει τις στοχευμένες πληροφορίες σχετικά με τα χαρακτηριστικά και τα “metrics” του σημείου ενδιαφέροντος. Έπειτα, ο wrapper επισκέπτεται επαναληπτικά όλα τα στοιχεία του επιλεγμένου υποδέντρου και εφαρμόζει τους κανόνες εξαγωγής δεδομένων που καθορίζονται από τα “metrics” και το “info” (βλέπε πίνακα 4) που αντιστοιχούν τμήματα αυτών των στοιχείων στα αντίστοιχα δομημένα πεδία.

label	μπορεί να είναι προσαρμοσμένο (ορισμένο από το χρήστη όνομα μετρικής ή πληροφοριών) ή μπορεί να εξαχθεί (τότε το πεδίο ορίζεται όπως στο πίνακα 6)
value	μπορεί να είναι προσαρμοσμένο (ορισμένο από το χρήστη όνομα μετρικής ή πληροφοριών) ή μπορεί να εξαχθεί (τότε το πεδίο ορίζεται όπως στο πίνακα 6)

Πίνακας 4: Προσδιορισμός των metrics και info

Για κάθε πληροφορία που αναφέρεται σε “info” ή “metric”, οι χρήστες πρέπει να καθορίσουν ένα label και ένα value. Αυτά μπορεί να είναι προσαρμοσμένα (καθορίζονται από τους χρήστες) ή μπορούν να εξαχθούν από τη σελίδα HTML. Σε περίπτωση εξαγωγής τους από μια σελίδα HTML, το αντίστοιχο πεδίο θα πρέπει να οριστεί όπως περιγράφεται στον πίνακα 6. Σε περίπτωση εξαγωγής μιας τιμής από τη σελίδα εισαγωγής HTML, οι χρήστες πρέπει να ορίσουν έναν CSS selector που εκφράζει τον κανόνα εξαγωγής που δείχνει το στοιχείο ενδιαφέροντος που περιέχει το επιθυμητό κομμάτι πληροφοριών. Εκτός από τον CSS selector, πρέπει επίσης να καθοριστεί ο τύπος των εξαγόμενων δεδομένων. Οι επιτρεπτοί τύποι είναι:

- 1) κείμενο
- 2) σύνδεσμος

- 3) εικόνα
- 4) αριθμός
- 5) οποιαδήποτε συμβολοσειρά (σε αυτήν την περίπτωση εξάγεται η τιμή του χαρακτηριστικού με το καθορισμένο όνομα).

Τέλος, μπορεί να εφαρμοστεί προαιρετικά ένα στάδιο για την επεξεργασία των εξαγόμενων δεδομένων δηλώνοντας το πεδίο αντικατάστασης όπως φαίνεται στον πίνακα 5.

Δεδομένα: διαμόρφωση
 Αποτέλεσμα: extracted_info, extracted_metrics
 1 start_browser_emulator(configuration.url)
 2 execute_events(configuration.events)
 3 fetch_page(configuration.url);
 4 extract_info(page, configuration.info);
 (I) εξήγαγε περιεχόμενο από τα ορισμένα στοιχεία
 (II) μεταεπεξεργασία του εξαγόμενου περιεχομένου εάν το πεδίο αντικατάστασης είναι ορισμένο
 5 extract_metrics(page, configuration.metrics);
 6 go-to step 1 if:
 (I) AFTER_EACH_EVENT τύποι συμβάντων λαμβάνουν χώρα
 (II) configuration.crawl είναι ορισμένο
 configuration = configuration.crawl
 7 επέστρεψε extracted_info, extracted_metrics

Αλγόριθμος 2: Wrapper δυναμικής ιστοσελίδας

Σε πολλές περιπτώσεις, μία πηγή ενδιαφέροντος περιέχει πολλές σελίδες με δεδομένα. Σε τέτοιες περιπτώσεις, αντί να ορίζει διαφορετικά αρχεία διαμόρφωσης και να αλλάζει μόνο το πεδίο URL, ο χρήστης πρέπει να καθορίσει έναν πίνακα διευθύνσεων URL (group_of_urls) που θα συλλέγονται από τον ίδιο wrapper. Ένα άλλο σενάριο συλλογής δεδομένων πολλών σελίδων αφορά σε σελιδοποιημένα δεδομένα. Για την δυνατότητα σελιδοποιημένες πρόσβασης σε τέτοια δεδομένα, πρέπει να καθοριστεί το πεδίο “next_page_selector” που δείχνει το στοιχείο που βρίσκεται το κουμπί “επόμενο” στη σελίδα. Η πλοήγηση στη σελίδα είναι ενεργοποιημένη εάν έχει οριστεί το πεδίο ανίχνευσης. Ο χρήστης πρέπει να δηλώσει ένα εμφωλευμένο αρχείο διαμόρφωσης που λαμβάνει ως εισαγωγή το URL ή τα URLs που λαμβάνονται από τη μέτρηση που ονομάζεται “crawl_to”.

selector	CSS selector που εκφράζει τον κανόνα εξαγωγής για την εξαγωγή της επιθυμητής πληροφορίας
type	τύπος εξαγόμενων δεδομένων (κείμενο, εικόνα κ.λπ.)
replace	ένα ζευγάρι πεδίων regex και with που ορίζουν ένα βήμα μεταεπεξεργασίας για τα εξαγόμενα δεδομένα

	regex	πίνακας από εκφράσεις regex που προσδιορίζουν τα μοτίβα που θα αντικατασταθούν
	with	πίνακας από συμβολοσειρές προς αντικατάσταση των μοτίβων που εντοπίστηκαν από τα regex

Πίνακας 5: Στοιχεία που πρέπει να καθοριστούν ώστε να εξαχθεί επιθυμητό περιεχόμενο από ιστοσελίδα.

Μια τελική επιλογή σύνθετης επεξεργασίας σελίδων αφορά τη λήψη στοιχείων HTML από δυναμικές ιστοσελίδες. Σε τέτοιες περιπτώσεις, η ροή εργασίας ανάκτησης και επεξεργασίας καθορίζεται από το πεδίο συμβάντων (Πίνακας 6). Υπάρχουν δύο τύποι συμβάντων που μπορούν να εκτελεστούν από τον εξομοιωτή του προγράμματος περιήγησης:

- (i) click,
- (ii) scroll down.

type	τύπος συμβάντος (click ή scroll_down)
selector	CSS selector του στοιχείου όπου το συμβάν θα εκτελεστεί
times_to_repeat	φορές επανάληψης του συγκεκριμένου συμβάντος
extraction_type	προσδιορίζει πότε να καλέσει την εξαγωγή δεδομένων, μία μετά από κάθε συμβάν (AFTER_EACH_EVENT) ή μία κατά την ολοκλήρωση (AFTER_ALL_EVENTS)

Πίνακας 6: Προσδιορισμός των συμβάντων για δυναμικές σελίδες

Υποστήριξη πλοήγησης

Μέσω της διαδικασίας συλλογής δεδομένων, έπρεπε να αντιμετωπιστούν πιο περίπλοκες πηγές που απαιτούσαν σημαντικές επεκτάσεις. Πολλές πηγές περιέχουν εξειδικευμένες σελίδες με περισσότερες πληροφορίες σχετικά με σημεία ενδιαφέροντος. Ορισμένες πηγές απαιτούν ένα βήμα πλοήγησης αρά σε ορισμένες περιπτώσεις απαιτούνται παραπάνω από ένα. Ως αποτέλεσμα αυτής της ανάγκης για την υποστήριξη πλοήγησης σε μια ιστοσελίδα αναπτύχθηκε ένα νέο πεδίο στο

αρχείο διαμόρφωσης που ονομάζεται “crawl” (όπως φαίνεται στον πίνακα 3) που καθιστά δυνατή την πλοήγηση μέσα στη σελίδα. Ο χειριστής χρειάζεται μόνο να δηλώσει τα εμφωλευμένα αρχεία διαμόρφωσης που λαμβάνουν ως είσοδο τη διεύθυνση URL ή τις διευθύνσεις URLs που λαμβάνονται από το πεδίο με την ετικέτα “crawl_to”.

Όπως απεικονίζεται στην Εικόνα 4, τα στοχευμένα δεδομένα από την Ιερά Αρχιεπισκοπή Αθηνών όσον αφορά θρησκευτικούς τουριστικούς προορισμούς στην ευρύτερη περιοχή της Θεσσαλονίκης¹¹ είναι καταναμεμένα σε πολλές ιστοσελίδες και για κάθε σημείο ενδιαφέροντος υπάρχει μια εξειδικευμένη ιστοσελίδα που περιέχει περισσότερες πληροφορίες σχετικά με το εκάστοτε σημείο. Για αυτόν τον λόγο, η συγκεκριμένη πηγή απαιτεί παραπάνω από ένα βήματα πλοήγησης για τη λήψη των επιθυμητών δεδομένων. Το αντίστοιχο αρχείο διαμόρφωσης παρέχεται στην ενότητα των παραρτημάτων.

The screenshot displays a web application interface. On the left, there is a grid of nine small images, each representing a different religious site in Thessaloniki. Each image is accompanied by a caption in Greek. On the right, there is a detailed text description of the 'I. N. Αγίων Αποστόλων' (Church of the Holy Apostles). The text describes the church's history, its location, and its architectural features. It mentions that the church was built in the 13th century and is one of the oldest in Thessaloniki. It also describes the church's interior, including its frescoes and mosaics. The text is written in a formal, academic style.

Εικόνα 4: Στα αριστερά, η ιστοσελίδα με τα σημεία ενδιαφέροντος στην Θεσσαλονίκη και στα δεξιά μία εξειδικευμένη ιστοσελίδα σημείου.

Τεχνικές Λεπτομέρειες

Το εν λόγω πλαίσιο αναπτύχθηκε με βάση τις βιβλιοθήκες Java της JSoup¹² και της Selenium¹³ για τη δημιουργία wrappers ικανών να εξάγουν δεδομένα από στατικές και δυναμικές σελίδες HTML αντίστοιχα. Τα πιο σημαντικά στοιχεία είναι τα wrapper generators, οι extractors και οι fetchers. Εν συνεχεία αναλύονται λεπτομερώς τα προαναφερθέντα στοιχεία.

Οι fetchers είναι υπεύθυνοι για τη λήψη του εγγράφου HTML μιας καθορισμένης ιστοσελίδας. Όλοι οι fetchers επεκτείνουν την αφηρημένη κλάση fetcher για να

¹¹

<http://www.religiousgreece.gr/west-central-macedonia/->

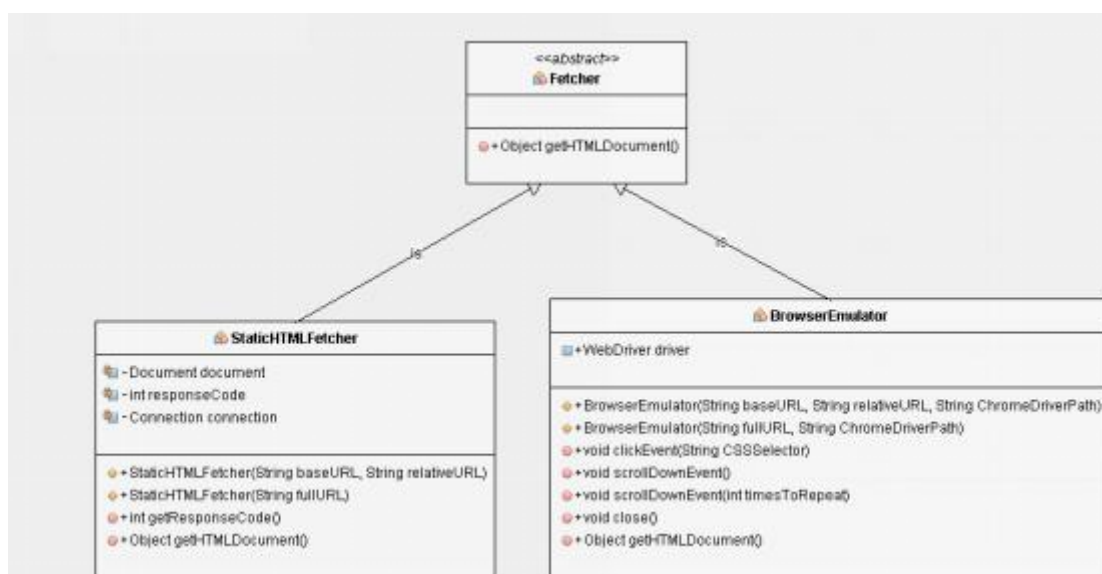
[/asset_publisher/3qlw8hzBmYgG/content/thessalonike](#)

¹² <http://jsoup.org/>

¹³ <http://www.seleniumhq.org/>

υλοποιήσουν τη μέθοδο `getHTMLDocument`. Ειδικότερα, υπάρχουν δύο κλάσεις που επεκτείνουν την αφηρημένη κλάση `fetcher`:

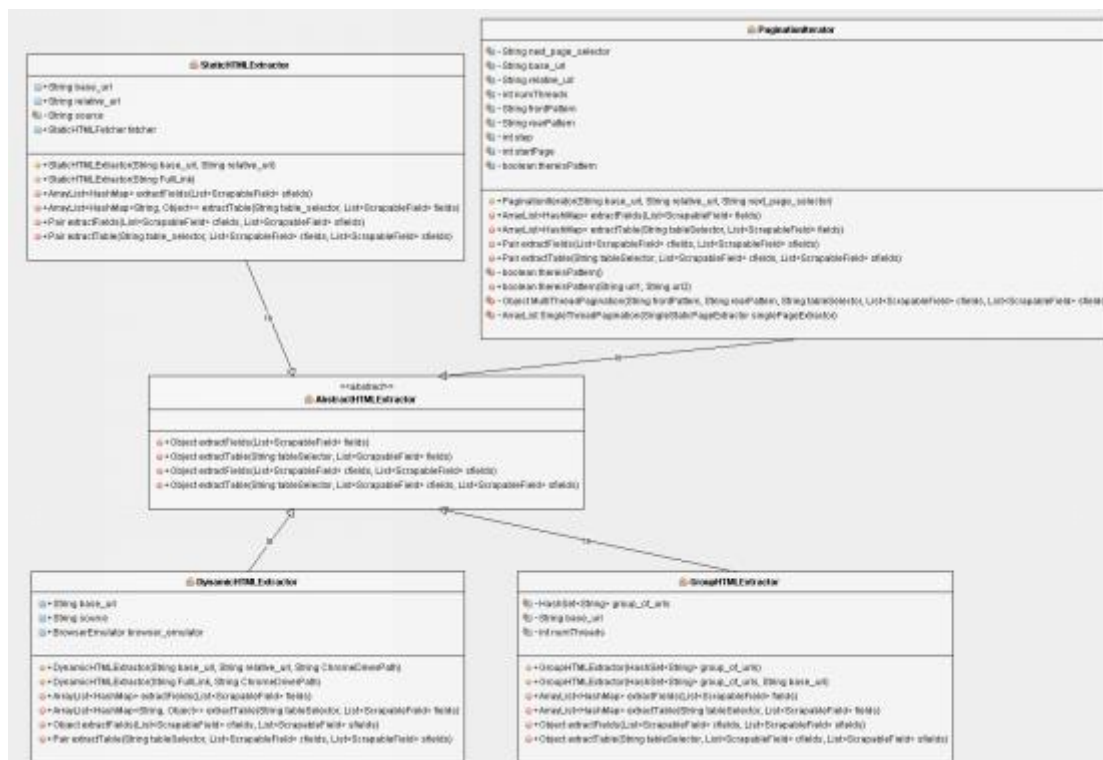
- `StaticHTMLFetcher`: είναι υπεύθυνος για την απόκτηση ενός εγγράφου HTML από μια στατική ιστοσελίδα HTML.
- `BrowserEmulator`: είναι υπεύθυνος για τη μίμηση ενός προγράμματος περιήγησης για τη λήψη του εγγράφου HTML από μια δυναμική ιστοσελίδα. Ο εξομοιωτής υποστηρίζει την εκτέλεση συμβάντων `click` και `scroll down`.



Εικόνα 5: UML αναπαράσταση της υλοποίησης των fetchers.

Όλοι οι `extractors` επεκτείνουν την `AbstractHTMLExtractor` για να υλοποιήσουν τις μεθόδους `extractFeilds` και `extractTable` που είναι υποχρεωτικές λειτουργίες για κάθε τύπο `extractor`. Οι `extractors` χρησιμοποιούν `fetchers` για να αποκτήσουν τα έγγραφα HTML για την εξαγωγή δεδομένων από καθορισμένες ιστοσελίδες HTML. Υπάρχουν τέσσερις κλάσεις που επεκτείνουν την κλάση `AbstractHTMLExtractor`:

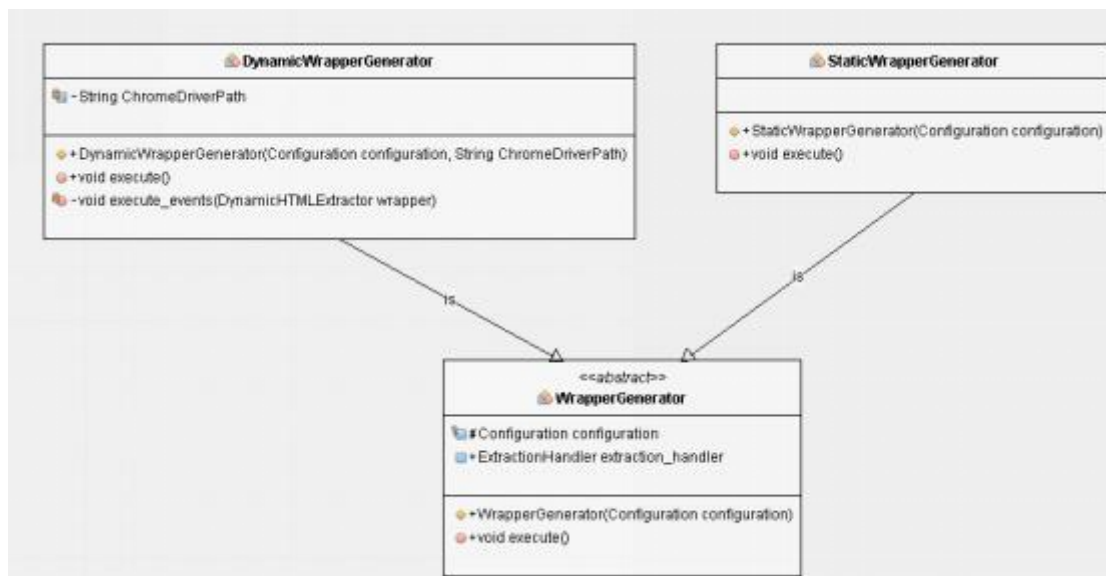
- `StaticHTMLExtractor`: είναι υπεύθυνος για την εξαγωγή δεδομένων από μια στατική ιστοσελίδα HTML.
- `DynamicHTMLExtractor`: είναι υπεύθυνος για την εξαγωγή δεδομένων από μια δυναμική ιστοσελίδα HTML. Το `DynamicHTMLExtractor` χρησιμοποιεί ένα `BrowserEmulator` για να εξαγάγει δεδομένα από δυναμικές ιστοσελίδες.
- `GroupHTMLExtractorQ` είναι υπεύθυνος για την εξαγωγή δεδομένων από μια ομάδα στατικών σελίδων HTML με παρεμφερή δομή.
- `PaginationHTMLExtractor`: είναι υπεύθυνος για την εξαγωγή δεδομένων από όλες τις σελίδες στις οποίες κατανέμονται τα δεδομένα. Στο αρχείο διαμόρφωσης το πεδίο `"next_page_selector"` δείχνει τη διεύθυνση URL της επόμενης σελίδας και ο `extractor` αναζητά ένα μοτίβο στις επακόλουθες διευθύνσεις URL, προκειμένου να αυτοματοποιήσει τη διαδικασία σελιδοποίησης και να εξαγάγει όλα τα στοχευμένα δεδομένα.



Εικόνα 6: UML Αναπαράσταση της υλοποίησης των extractors

Για την εξαγωγή δεδομένων από μια καθορισμένη διαδικτυακή πηγή, απαιτείται η δημιουργία ενός διαδικτυακού wrapper. Ο διαδικτυακός wrapper είναι μια διαδικασία που αναζητά ένα σύνολο δεδομένων ενδιαφέροντος και το εξάγει με αυτοματοποιημένο τρόπο. Τα WrapperGenerators είναι υπεύθυνα για την αυτοματοποίηση της διαδικασίας εξαγωγής δεδομένων και χρησιμοποιούν κατάλληλα extractors βάσει καθορισμένων ρυθμίσεων διαμόρφωσης. Υπάρχουν δύο κλάσεις που επεκτείνουν την αφηρημένη κλάση WrapperGenerator:

- **StaticWrapperGenerator**: είναι υπεύθυνος για την αυτόματη δημιουργία ενός ή και παραπάνω wrappers για μια ή και περισσότερες στατικές ιστοσελίδες HTML με βάση τις ρυθμίσεις διαμόρφωσης που καθορίζονται από το χρήστη.
- **DynamicWrapperGenerator**: είναι υπεύθυνος για την αυτόματη δημιουργία ενός ή περισσότερων wrapper για μία ή περισσότερες δυναμικές ιστοσελίδες HTML με βάση τις ρυθμίσεις διαμόρφωσης που καθορίζονται από το χρήστη.



Εικόνα 7: UML Αναπαράσταση της υλοποίησης των wrapper generators.

Τέλος, το WrapperExecutor είναι υπεύθυνο για την εκτέλεση των Wrappers που έχουν δημιουργηθεί από τους wrapper generators και αποθηκεύει τα εξαγόμενα δεδομένα σε μια βάση δεδομένων ή σε ένα αρχείο όπως δηλώνεται στο αρχείο διαμόρφωσης.

Οι εικόνες 5, 6 και 7 απεικονίζουν με τη μορφή διαγραμμάτων UML την υλοποίηση των fetcheres, των extractors και των wrapper generators αντίστοιχα. Στα παραρτήματα παρέχεται ο πηγαίος κώδικας των StaticWrapperGenerator και DynamicWrapperGenerator.

5 Αναζήτηση σε WEB & κοινωνικές πλατφόρμες

Στην προηγούμενη ενότητα, ο στόχος είναι η συλλογή πληροφοριών με βάση την υπόθεση ότι οποιαδήποτε ιστοσελίδα που ανήκει σε έναν καθορισμένο ιστότοπο περιλαμβάνει δεδομένα που μπορούν να χρησιμοποιηθούν για τους σκοπούς του e-ΧΝΗΛΑΤΗ. Ωστόσο, μπορεί εύκολα να παρατηρηθεί ότι δεν ανήκουν όλοι οι ιστότοποι σε αυτήν την κατηγορία. Υπάρχουν ιστότοποι όπου ένα μικρό ποσοστό των ιστοσελίδων τους, που δεν περιλαμβάνονται απαραίτητα σε έναν μοναδικό υποτομέα, μπορούν να αξιοποιηθούν. Επιπλέον, υπάρχουν περιπτώσεις όπου δεν δύναται ο καθορισμός συγκεκριμένων ιστοτόπων, αλλά υπάρχει ενδιαφέρον για τη συλλογή περιεχομένου βάσει ορισμένων κριτηρίων αναζήτησης. Σε άλλα σενάρια υπάρχει η θέληση για συλλογή πληροφοριών από κοινωνικά μέσα ή γεωβάσεις αντί για ιστοτόπους. Σε αυτήν την ενότητα παρουσιάζονται οι τρόποι και οι λόγοι με τους οποίους επιλέχθηκαν οι στοχευμένες πηγές αλλά και οι ίδιες οι πηγές που περιέχουν ετερογενές περιεχόμενο, καθώς και οι διαφορετικοί τρόποι πρόσβασης στα δεδομένα τους.

5.1 Ιστοσελίδες

Για τη συλλογή πληροφοριών και δεδομένων που θα χρησιμοποιηθούν από τον e-ΧΝΗΛΑΤΗ, μια βασική πηγή πληροφορίας θα είναι οι ιστοσελίδες. Οι ιστοσελίδες θα αποτελούν και το και την κύρια πηγή πληροφορίας που θα μπορεί να προσθέτει ο διαχειριστής. Οι ιστοσελίδες που δύναται να χρησιμοποιηθούν πρέπει να ακολουθούν κάποια πρότυπα:

- Η δομή της ιστοσελίδας να έχει κάποια επαναληψιμότητα για να μπορούν να συλλεχθούν τα δεδομένα (information retrieval) με αυτόματο τρόπο
- Η πληροφορία που παρέχεται να είναι αξιόπιστη και εντός του αντικειμένου του e-ΧΝΗΛΑΤΗ
- Παρουσία γραπτών πληροφοριών και πολυμεσικού υλικού ανα σημείο ενδιαφέροντος
- Να περιέχεται με κάποιο τρόπο η χωρική πληροφορία του σημείου ενδιαφέροντος για την ανάκτηση του και την αναπαράσταση του ως σημείο στο χάρτη

Για την ανάκτηση των πληροφοριών από τις ιστοσελίδες χρησιμοποιείται η μέθοδος scraping όπως αναφέρθηκε σε προηγούμενο κεφάλαιο και πιο ειδικά το πρόγραμμα easIE¹⁴. Το easIE για τη λειτουργία του δέχεται ένα αρχείο το οποίο πρέπει να τροποποιηθεί κατάλληλα για την κάθε ιστοσελίδα. Το αρχείο αυτό υποδικνίζει τον τρόπο και τα σημεία από τα οποία που θα συλλεχθεί η πληροφορία. Ένα παραδειγμα τέτοιου αρχείου μπορεί φαίνεται στο παράρτημα (Παράρτημα Α).

Αρχείο που επιστρέφει το easIE:

```
[  
  {  
    "company": [{"name": "Ι. Ν. Παναγίας Κοσμοσώτρας"}],
```

¹⁴ <https://github.com/MKLab-ITI/easIE>

```

"metrics": [[
  {
    "name": "full_descr",
    "citeyear": 2020,
    "source": "http://www.religiousgreece.gr/east-macedonia-thrace/-
/asset_publisher/mHqF2LEbi6sS/content/i-n-panagias-kosmosoteras",
    "type": "other",
    "value": "<p style=\"text-align: justify;\">Στο κέντρο της κωμόπολης των
φερών, στη δυτική όχθη του ποταμού Έβρου, βρίσκεται ένα από τα πιο συμβολικά μνημεία της
Θράκης, ο<strong> ιερός ναός της Παναγίας, άλλοτε καθολικό της Μονής Κοσμοσώτειρας ή
της Βήρας</strong>, όπως λεγόταν η τοποθεσία των φερών στη βυζαντινή εποχή.</p> \n<p
style=\"text-align: justify;\">Η μονή ιδρύθηκε το 1152 από τον
σεβαστοκράτορα<em> Ισαάκιο Κομνηνό</em>, γιο του<em> Αλεξίου Α΄
Κομνηνού</em> και της<em> Ειρήνης Δούκαινας</em>, ο οποίος έχει ταφεί μέσα στο
καθολικό, κάτω από τον τρουλίσκο της βορειοδυτικής γωνίας. Ο ίδιος μετέφερε εδώ τα γλυπτά
του τάφου του, που είχε ετοιμάσει αρχικά στο καθολικό της Μονής της Χώρας, στην
Κωνσταντινούπολη. Γύρω από το μοναστήρι σταδιακά αναπτύχθηκε μικρός οικισμός, που
οχυρώθηκε με τείχος, καθώς βρισκόταν σε σημαντική θέση, στον δρόμο προς την Αδριανούπολη
και την Κωνσταντινούπολη. Μετά το 1371 η μονή κατακτήθηκε από τον<em> Λαλά
Σαχίν</em> και τον<em> Εβρενός</em> και το καθολικό της μετατράπηκε σε
τέμενος του<em> Σουλειμάν πασά</em> και του<em> σουλτάνου Ορχάν</em>.
Το 1940 έγιναν εργασίες αποκατάστασης του ναού και ξεκίνησε πάλι η λειτουργία του.</p>
\n<p style=\"text-align: justify;\">Από το αρχικό μοναστηριακό συγκρότημα διατηρούνται
μόνο τμήματα του οχυρωματικού περιβόλου και το καθολικό.<br>
Το<strong> καθολικό</strong>, που απηχεί την τέχνη της Κωνσταντινούπολης του 12ου
αιώνα, ανήκει στον αρχιτεκτονικό τύπο του<em> πεντάτρουλου</em>, απλού δικιόνιου
σταυροειδούς εγγεγραμμένου ναού</em> με νάρθηκα στα δυτικά. Στην περίοδο της
Τουρκοκρατίας αλλοιώθηκαν τα παράθυρα των τρούλων και των αψίδων και κατεδαφίστηκε ο
νάρθηκός του. Οι τέσσερις αντηρίδες που υπάρχουν εξωτερικά είναι μεταγενέστερες.<br>
Στο<strong> εσωτερικό</strong> του ναού, στη νοτιοανατολική γωνία, σώζεται
εντοιχισμένο κεραμικό κόσμημα, στο οποίο απεικονίζεται ο μονοκέφαλος αετός, σύμβολο της
δυναστείας των Κομνηνών της Τραπεζούντας. Επίσης, διατηρείται μέρος του τοιχογραφικού
διακόσμου, που αποτελεί σημαντικό δείγμα της κωνσταντινοπολίτικης ζωγραφικής των μέσων του
12ου αιώνα. Μεταξύ άλλων εικονίζονται τέσσερις στρατιωτικές μορφές (ανά ζεύγη ανάμεσα στα
τοξωτά παράθυρα του βόρειου και νότιου κλίτους), για την ταύτιση των οποίων έχει
διατυπωθεί η άποψη ότι πιθανώς πρόκειται για προσωπογραφίες μελών της οικογένειας των
Κομνηνών.</p> \n<p>Ο ναός όχι άδικα αποτελεί ένα από τα σημαντικότερα βυζαντινά μνημεία
της Ελλάδας και έχει καθιερωθεί ως το<strong> Παγκόσμιο Προσκύνημα των απανταχού
Θρακών</strong>.</p>"
  },
  {
    "name": "lat",
    "citeyear": 2020,
    "source": "http://www.religiousgreece.gr/east-macedonia-thrace/-
/asset_publisher/mHqF2LEbi6sS/content/i-n-panagias-kosmosoteras",
    "type": "other",
    "value": "40.893967"
  },
  {
    "name": "lng",
    "citeyear": 2020,
    "source": "http://www.religiousgreece.gr/east-macedonia-thrace/-
/asset_publisher/mHqF2LEbi6sS/content/i-n-panagias-kosmosoteras",
    "type": "other",
    "value": "26.170395"
  },
  {
    "name": "image",
    "citeyear": 2020,
    "source": "http://www.religiousgreece.gr/east-macedonia-thrace/-
/asset_publisher/mHqF2LEbi6sS/content/i-n-panagias-kosmosoteras",
    "type": "link",
    "value": "/image/journal/article?img_id=329520&t=1445618729164&width=940"
  },
  {
    "name": "info",
    "citeyear": 2020,
    "source": "http://www.religiousgreece.gr/east-macedonia-thrace/-
/asset_publisher/mHqF2LEbi6sS/content/i-n-panagias-kosmosoteras",
    "type": "textual",
    "value": "Χρονολόγηση: 1152 Περίοδος: Βυζαντινή Πανηγυρίζει: 15 Αυγούστου
Ιερά Μητρόπολη: Αλεξανδρουπόλεως Φορέας προστασίας: Εφορεία Αρχαιοτήτων Κομοτηνής
Διεύθυνση: Φέρες, Τ.Κ. 68500 Πρόσβαση: Ι.Χ. αυτοκίνητο, Υπεραστικό Λεωφορείο ΚΤΕΛ,

```

```

Αεροπορικώς έως Αλεξανδρούπολη Πρόσβαση ΑμεΑ: Ναι Στάθμευση: Ελεύθερη Εξωτερικός
σύνδεσμος"
    },
    {
      "name": "url",
      "citeyear": 2020,
      "source": "http://www.religiousgreece.gr/east-macedonia-thrace/-
/asset_publisher/mHqF2LEbi6sS/content/i-n-panagias-kosmosoteras",
      "type": "link",
      "value": "http://www.imalex.gr/DB735F3B.el.aspx"
    }
  ]
},
...
]

```

Το τελικό αρχείο που παράγεται από το easIE είναι της μορφής json array το οποίο πρέπει να μετατραπεί κατάλληλα στη ενιαία μορφή geoJSON που χρησιμοποιείται στην πλατφόρμα. Ακόμη κάποιες ιστοσελίδες μπορεί να περιέχουν πληροφορίες που θα χρειαστούν περαιτέρω επεξεργασία για να μπορέσουν να αξιοποιηθούν από την πλατφόρμα του e-ΧΝΗΛΑΤΗ.

Για τη μετατροπή στη επιθυμητή μορφή geoJSON, πρέπει να υλοποιηθεί ένα κατάλληλο script γραμμένο σε γλώσσα javascript. Ο κώδικας που ακολουθεί παρουσιάζει την περίπτωση της ιστοσελίδας www.religiousgreece.gr

```

function create_geoJSON(easie_result){
  $.each( easie_result, function( key, value ) {
    var new_json = JSON.parse(JSON.stringify(json_template));

    new_json.properties.source_cat = "http://www.religiousgreece.gr";
    new_json.properties.category.push("02");

    new_json.properties.name = value.company[0].name;

    new_json.geometry.coordinates.push(parseFloat(value.metrics[0][2].value)); //lat
    new_json.geometry.coordinates.push(parseFloat(value.metrics[0][1].value)); //lon

    new_json.properties.full_descr = value.metrics[0][0].value;
    //new_json.properties.website = value.metrics[0][0].source;

    new_json.properties.image = "http://www.religiousgreece.gr" +
    value.metrics[0][3].value; //image

    new_json.properties.website = value.metrics[0].length > 5 ?
    value.metrics[0][5].value : null; //url

    info_str = value.metrics[0][4].value; //info string
    info_str = info_str.replace("Εξωτερικός σύνδεσμος", "");
    info_json = create_info_json(info_str); //info json

    new_json.properties.address = info_json["Διεύθυνση"];
    new_json.properties.telephone = info_json["Τηλέφωνο"];
    debugger
    new_json.properties.work_hours = info_json["Ωράριο"];

    geojson_template.features.push(new_json);
    return geojson_template;
    //new_json.full_descr = value.metrics[0][0].value;

  });
}

```

```
function create_info_json (v_string){
    var info_array = v_string.split(" ");
    var info_json = {};
    var json_key = "";
    var json_value = "";
    //debugger;
    var arrayLength = info_array.length;
    for (var i = 0; i < arrayLength; i++) {
        //debugger;
        //console.log(i);
        var element = info_array[i];
        //console.log(element);
        if (element.endsWith(":")){ //found key
            if (json_value != ""){ //populate json
                info_json[json_key] = json_value;
                json_key = element.slice(0,-1); //remove ":"
                json_value = "";
            }
            else{
                json_key = element.slice(0,-1);
                json_value = "";
            }
        }
        else{
            json_value = json_value + " " + element;
        }
        if (i == arrayLength - 1){ //last element
            if (json_key != ""){
                info_json[json_key] = json_value;
            }
        }
    }
    //console.log(info_json);
    return info_json;
}

var geojson_template = {
    "type":"FeatureCollection",
    "name":"points_of_interest",
    "features":[
    ]
};

var json_template = {
    "id":null,
    "type":"Feature",
    "geometry":{
        "type":"Point",
        "coordinates":[]
    },
    "properties":{
        "wc":null,
        "food":null,
        "name":null,
        "dimos":null,
        "email":null,
        "image":null,
        "is_ar":null,
        "awards":null,
        "refuge":null,
        "routes":null,
        "unesco":null,
        "address":null,
        "parking":null,
        "s_descr":null,
        "website":null,
        "category":[],
        "cultural":null,
    }
}
```

```
"name_eng":null,  
"shopping":null,  
"periferia":null,  
"protected":null,  
"published":null,  
"telephone":null,  
"exhibition":null,  
"full_descr":null,  
"restaurant":null,  
"source_cat":null,  
"time_spent":null,  
"amea_access":null,  
"per_enotita":null,  
"s_descr_eng":null,  
"subcategory":[],  
"ticket_cost":null,  
"interchanges":null,  
"accommodation":null,  
"certification":null,  
"info_material":null,  
"observatories":null,  
"difficulty_lvl":null,  
"full_descr_eng":null,  
"interchange_id":[],  
"kallikratis_id":null,  
"historic_period":null,  
"outdoor_company":null,  
"nearest_neighbour":null,  
"outdoor_activities":null,  
"recreational_areas":null,  
"work_hours":null  
}  
}
```

5.2 Γεωβάσεις

Μια άλλη πηγή ανάκτησης πληροφοριών και δεδομένων που αφορούν την πλατφόρμα του e-ΧΝΗΛΑΤΗ, είναι διάφορες ανοιχτές γεωβάσεις. Οι γεωβάσεις αποτελούν βάσεις δεδομένων που περιέχουν χωρική πληροφορία.

Η πληροφορία που θα μπορούσε να χρησιμοποιηθεί από γεωβάσεις μπορεί να αποτελείται από συλλεγμένα και ομαδοποιημένα δεδομένα με γεωχωρική πληροφορία (όπως οικισμοί, όρια νομών, περιοχές natura κ.α.).

Τέτοιου είδους δεδομένα που δύναται να χρησιμοποιηθούν είναι της Ελληνικής Στατιστικής Αρχής¹⁵ (ΕΛΣΤΑΤ), του geodata.gov.gr¹⁶ και του openStreetMap¹⁷ τα οποία περιέχουν πολλαπλές πληροφορίες που μπορούν να αξιοποιηθούν από την πλατφόρμα του e-ΧΝΗΛΑΤΗ.

Η κάθε γεωβάση έχει δική της δομή οπότε στις περιπτώσεις που τα δεδομένα θα χρησιμοποιηθούν για την εισαγωγή σε κάποιον υπάρχων πίνακα της γεωβάσης του e-ΧΝΗΛΑΤΗ θα χρειαστούν την κατάλληλη επεξεργασία για την ενσωμάτωσή τους.

¹⁵ <https://www.statistics.gr/el/digital-cartographical-data>

¹⁶ geodata.gov.gr

¹⁷ <https://www.openstreetmap.org/export>

	OBJECTID	KALCODE	LEKTIKO	SHAPE_Leng	SHAPE_Area
1	72	1703	ΔΗΜΟΣ ΠΡΕΣΠΩΝ	101023,67717200000	518686745,26499998569
2	71	1702	ΔΗΜΟΣ ΑΜΥΝΤΑΙΟΥ	149551,64860099999	590178248,60000002384
3	66	1502	ΔΗΜΟΣ ΔΕΣΚΑΤΗΣ	146299,95102099999	433312847,87900000811
4	65	1501	ΔΗΜΟΣ ΓΡΕΒΕΝΩΝ	284113,50485299999	1862400117,38000011444
5	68	1602	ΔΗΜΟΣ ΝΕΣΤΟΡΙΟΥ	172204,74818400000	620167716,79900002480
6	67	1601	ΔΗΜΟΣ ΚΑΣΤΟΡΙΑΣ	181843,28366399999	763890110,64699995518
7	78	1806	ΔΗΜΟΣ ΚΟΝΙΤΣΑΣ	191149,12985500001	953579447,28299999237
8	77	1805	ΔΗΜΟΣ ΖΙΤΣΑΣ	147065,15700800001	568213320,25800001621
9	80	1808	ΔΗΜΟΣ ΠΩΓΩΝΙΟΥ	144871,16898100000	701996732,15499997139
10	79	1807	ΔΗΜΟΣ ΜΕΤΣΟΒΟΥ	112568,74969000000	363334198,60199999809
11	74	1802	ΔΗΜΟΣ ΒΟΡΕΙΩΝ ΤΖΟΥΜΕΡΚΩΝ	119433,08374000000	357696497,56599998474
12	73	1801	ΔΗΜΟΣ ΙΩΑΝΝΙΝΩΝ	134165,74145800000	403044778,09600001574
13	76	1804	ΔΗΜΟΣ ΖΑΓΟΡΙΟΥ	160727,39053400001	992553183,89800000191
14	75	1803	ΔΗΜΟΣ ΔΩΔΩΝΗΣ	171880,91982700001	659796619,72500002384
15	86	2002	ΔΗΜΟΣ ΣΟΥΛΙΟΥ	140363,88825200000	502176405,86799997091
16	85	2001	ΔΗΜΟΣ ΗΓΟΥΜΕΝΙΤΣΑΣ	230195,88985800001	430453123,22299998999

Εικόνα 8: Παράδειγμα πίνακα καλλικρατικών δήμων από ΕΛΣΤΑΤ

5.3 APIs

Σε προηγούμενα κεφάλαια παρουσιάστηκαν οι τρόποι με τους οποίους λαμβάνουν χώρα οι διαδικασίες του crawling και του scrapping, τρόποι οι οποίοι προϋποθέτουν τον εντοπισμό κατάλληλων διαδικτυακών πηγών βάσει σχετικού περιεχομένου και ανάλογη υλοποίηση κι επέκταση εργαλείων για αυτόματη/ημι-αυτόματη εξαγωγή δεδομένων. Σε ορισμένες περιπτώσεις διαδικτυακών πηγών προσφέρεται, είτε ανοιχτά και δωρεάν είτε με πληρωμή και ελεγχόμενο τρόπο, η πρόσβαση σε APIs που αποκρίνονται με τα ζητούμενα δεδομένα με καθορισμένη δομημένη μορφή, όπως η μορφή JSON. Η απόκριση αυτή συνήθως εξαρτάται από το ίδιο το ερώτημα που κάνει κάποιος λαμβάνοντας ως απάντηση τα αντίστοιχα δεδομένα. Στο παρόν έργο προτιμήθηκαν πλατφόρμες που προσφέρουν δωρεάν κάποιες υπηρεσίες που εξυπηρετούν τις ανάγκες του ίδιου έργου για:

- εύρεση κι ανάκτηση χρονικά μεμονωμένων ή περιοδικών γεγονότων γεωχωρικά στο πεδίο δράσης του e-ΧΝΗΛΑΤΗ,
- εύρεση κι ανάκτηση δελτίων πρόγνωσης καιρού ή δελτίων έκτακτων καιρικών φαινομένων γεωχωρικά στο πεδίο δράσης του e-ΧΝΗΛΑΤΗ.

5.3.1 Ticketmaster API

Το ticketmaster API παρέχει αρκετών ειδών μεταδεδομένα που έγκεινται μέσα στους σκοπούς του e-ΧΝΗΛΑΤΗ με εύκολα προσβάσιμο και επεξεργάσιμο τρόπο. Δεδομένου ότι το περιεχόμενο της ticketmaster αφορά και την ελληνική επικράτεια είναι ιδανικό για να καλύψει τις ανάγκες για άντληση γεγονότων μαζί με χρήσιμα μεταδεδομένα όπως γεωχωρικές συντεταγμένες, ημερομηνίες και ώρες διεξαγωγής, παροχές προσβασιμότητας σε ΑΜΕΑ και άλλα. Μπορούν να ανακτηθούν με χρήση καθολικής εύρεσης πόρων, θέτοντας ως χώρα στόχο την Ελλάδα, του ακόλουθου API:

https://app.ticketmaster.com/discovery/v2/events.json?countryCode=GR&apikey=*

Η υπόλοιπη υποενότητα αναλώνεται στην επεξήγηση της πρόσβασης και χρήσης βασικών λειτουργιών του Ticketmaster REST API V2.

Περιγραφή Πόρων

Το Ticketmaster REST API χρησιμοποιείται κυρίως για mobile εφαρμογές που επιθυμούν να κάνουν κάποια κράτηση σε κάποιο γεγονός. Δεν είναι δημόσια υποστηριζόμενο API και ισχύουν κάποιοι κανόνες πρόσβασης από την εταιρεία. Το API αντικατοπτρίζει στενά το διαθέσιμο περιεχόμενο στον αντίστοιχο ιστότοπο παρέχοντας πρόσβαση σε όλη την πληροφορία με τις αποκρίσεις να είναι υπό τη μορφή JSON. Η βασική διεύθυνση URL είναι η : <https://app.ticketmaster.com/> . Διαθέτει πολλούς πόρους που ομαδοποιούν διάφορα τελικά σημεία (endpoints) για τη διευκόλυνση δομημένων ερωτημάτων σε αυτό, όπου ορισμένοι εξ αυτών ενδέχεται να έχουν εμφωλευμένους πόρους. Το σημείο εκκίνησης για όλες τις αναφορές είναι ο πόρος διαμόρφωσης που διατίθεται στη διεύθυνση: <https://app.ticketmaster.com/discovery/v2/> .

API & Tokens

Προς το παρόν, απαιτείται κλειδί developer API και oauth access token με τη δημιουργία λογαριασμού προγραμματιστή στην πλατφόρμα για την υποβολή αιτημάτων στο API. Ωστόσο, οι περιορισμοί που ισχύουν αφορούν 5000 αίτηματα ημερησίως στο API, που είναι υπερακρετά για τις ανάγκες του έργου και 100 authorization αιτήματα εφαρμογής ανά λεπτό.

Event Search API

Το Event Search API επιτρέπει την εκτέλεση της ίδιας αναζήτησης που είναι διαθέσιμη μέσω της σύνθετης αναζήτησης στον αντίστοιχο ιστότοπο. Αυτό σημαίνει ότι μπορούν να χρησιμοποιηθούν διάφορες τροποποιήσεις και ακόμη παραπάνω στα ερωτήματα όπως: id, keyword, attractionId, venueId, postalCode, latlong, radius, unit, source, locale, marketId, startDate, endDate, includeTBA, includeTBD, includeTest, size, page, sort, onsaleStartDate, onsaleEndDate, city, countryCode, stateCode, classificationName, classificationID, dmald, localStartDate, localEndDate, publicVisibilityStartDate, preSaleDate, onsaleOnStartDate, onsaleOnAfterStartDate, collectionId, segmentId, segmentName, includeFamily, promoterId, genreId, subGenreId, typeId, subTypeId, geoPoint, preferredCountry, includeSpellCheck, domain. Ένα παράδειγμα αποκρίσης του API παρουσιάζεται εκτενώς στο παράρτημα.

Παράμετροι ερωτήματος αναζήτησης

Παράμετρος	Περιγραφή	Τύπος	Προεπιλεγμένη Τιμή	Απαιτείται
id	Filter entities by its id	String		No
keyword	Keyword to search on	String		No
attractionId	Filter by attraction id	String		No

venueId	Filter by venue id	String		No
postalCode	Filter by postal code / zipcode	String		No
latlong	Filter events by latitude and longitude, this filter is deprecated and maybe removed in a future release, please use geoPoint instead	String		No
radius	Radius of the area in which we want to search for events.	String		No
unit	Unit of the radius	String enum:["miles", "km"]	miles	No
source	Filter entities by its primary source name OR publishing source name	String enum:["ticketmaster", "universe", "frontgate", "tmr"]		No
locale	The locale in ISO code format. Multiple comma-separated values can be provided. When omitting the country part of the code (e.g. only 'en' or 'fr') then the first matching locale is used. When using a '*' it matches all locales. '*' can only be used at the end (e.g. 'en-us,en,*')	String	en	No
marketId	Filter by market id	String		No
startDateTime	Filter with a start date after this date	String		No
endDateTime	Filter with a start date before this	String		No

	date			
includeTBA	yes, to include with date to be announce (TBA)	String enum:["yes", "no", "only"]	no if date parameter sent, yes otherwise	No
includeTBD	yes, to include with a date to be defined (TBD)	String enum:["yes", "no", "only"]	no if date parameter sent, yes otherwise	No
includeTest	Yes if you want to have entities flag as test in the response. Only, if you only wanted test entities	String enum:["yes", "no", "only"]	no	No
size	Page size of the response	String	20	No
page	Page number	String	0	No
sort	Sorting order of the search result. Allowable values : 'name,asc', 'name,desc', 'date,asc', 'date,desc', 'relevance,asc', 'relevance,desc', 'distance,asc', 'name,date,asc', 'name,date,desc', 'date,name,asc', 'date,name,desc', 'distance,date,asc', 'onSaleStartDate,asc', 'id,asc', 'venueName,asc', 'venueName,desc', 'random'	String	relevance, desc	No
onsaleStartDateTime	Filter with onsale start date after this date	String		No
onsaleEndDateTime	Filter with onsale end date before this date	String		No
city	Filter by city	Array		No

countryCode	Filter by country code	String		No
stateCode	Filter by state code	String		No
classificationName	Filter by classification name: name of any segment, genre, sub-genre, type, sub-type. Negative filtering is supported by using the following format '-'. Be aware that negative filters may cause decreased performance.	Array		No
classificationId	Filter by classification id: id of any segment, genre, sub-genre, type, sub-type. Negative filtering is supported by using the following format '-'. Be aware that negative filters may cause decreased performance.	Array		No
dmaId	Filter by dma id	String		No
localStartDateTime	Filter with event local start date time within this range	Array		No
localStartEndTime	Filter event where event local start and end date overlap this range	Array		No
startEndTime	Filter event where event start and end date overlap this range	Array		No
publicVisibilityStartTime	Filter with events with public visibility starting	Array		No
preSaleDateTime	Filter events with a presaleFilterTransformer start and end that intersects with this range	Array		No
onsaleOnStartDate	Filter with onsale start date on this date	String		No

onsaleOnAfterStart Date	Filter with onsale range within this date	String		No
collectionId	Filter by collection id	Array		No
segmentId	Filter by segment id	Array		No
segmentName	Filter by segment name	Array		No
includeFamily	Filter by classification that are family-friendly	String enum:["yes", "no", "only"]	yes	No
promoterId	Filter by promoter id	String		No
genreId	Filter by genreId	Array		No
subGenreId	Filter by subGenreId	Array		No
typeId	Filter by typeId	Array		No
subTypeId	Filter by subTypeId	Array		No
geoPoint	filter events by geoHash	String		No
preferredCountry	Popularity boost by country, default is us.	String enum:["us", "ca"]	us	No
includeSpellcheck	yes, to include spell check suggestions in the response.	String enum:["yes", "no"]	no	No
domain	Filter entities based on domains they are available on	Array		No

5.3.2 Open Weather API

Το Open Weather API παρέχει πληροφορίες σχετικά με ζωντανές καιρικές ενδείξεις αλλά και βραχυπρόθεσμα δελτία καιρού ανά περιοχή. Το περιεχόμενο του αφορά όλο τον πλανήτη, και ως εκ τούτου και την ελληνική επικράτεια, μιας και αντλεί δεδομένα από επίσημους μετεωρολογικούς σταθμούς που έγκεινται στην επικράτεια του έργου. Μπορούν να ανακτηθούν με χρήση καθολικής εύρεσης πόρων θέτοντας

αντίστοιχα τα γεωγραφικά πλάτη και μήκη που ενδιαφέρουν με τη χρήση του ακόλουθου API:

<https://api.openweathermap.org/data/2.5/onecall?lat={lat}&lon={lon}&exclude={part}&appid={YOUR API KEY}>

Η υπόλοιπη υποενότητα αναλώνεται στην επεξήγηση της πρόσβασης και χρήσης βασικών λειτουργιών του Open Weather REST API.

Περιγραφή Πόρων

Το Open Weather REST API ανήκει στην OpenWeather Ltd, μια βρετανική εταιρεία, και παρέχει εκτός από δωρεάν υπηρεσίες και πακέτα επί πληρωμή, ως αποτέλεσμα δεν πρόκειται για ένα δημόσια υποστηριζόμενο API και σαφώς ισχύουν κάποιοι κανόνες πρόσβασης από την ίδια την εταιρεία. Το API αντικατοπτρίζει στενά το διαθέσιμο περιεχόμενο στον αντίστοιχο ιστότοπο παρέχοντας πρόσβαση σε όλη την ζητούμενη πληροφορία με τις αποκρίσεις να είναι υπό τη μορφή JSON. Η βασική διεύθυνση URL είναι η : <https://openweathermap.org/> . Διαθέτει πολλαπλούς πόρους που ομαδοποιούν ανα περίπτωση τις πληροφορίες για διάφορα τελικά σημεία (endpoints) για τη διευκόλυνση δομημένων ερωτημάτων σε αυτό, όπου ενδέχεται ορισμένοι εξ αυτών να εμπεριέχουν εμφωλευμένους πόρους. Το σημείο εκκίνησης για όλες τις αναφορές είναι ο πόρος διαμόρφωσης που διατίθεται στη διεύθυνση: <https://api.openweathermap.org/data/2.5/onecall>

API & Tokens

Προς το παρόν απαιτείται κλειδί API το οποίο παρέχεται με τη δημιουργία λογαριασμού στην πλατφόρμα του <https://OpenWeatherMap.org> . Δεν απαιτείται OAuth access κι επιπλέον ισχύουν κάποιοι περιορισμοί στη δωρεάν έκδοση χρήσης του API, 60 κλήσεις ανά λεπτό και 1.000.000 κλήσεις ανά μήνα που είναι ικανοποιητική συχνότητα για τις ανάγκες του e-ΧΝΗΛΑΤΗ.

One Call API

Το One Call API επιτρέπει την εκτέλεση και ανάκτηση δεδομένων με τον ίδιο τρόπο που εφαρμόζεται η αναζήτηση και ανάκτηση στον αντίστοιχο ιστοτόπο. Αυτό σημαίνει ότι δύνανται να χρησιμοποιηθούν διάφορες τροποποιήσεις πάνω στα ερωτήματα, όπως: lat, lon, API key, part (current, minutely, hourly, daily), units και lang. Ένα παράδειγμα απόκρισης του One Call API παρουσιάζεται λεπτομερέστερα στα παραρτήματα.

Παράμετροι ερωτήματος αναζήτησης

Παράμετρος	Περιγραφή
lat	γεωγραφικό πλάτος
lon	γεωγραφικό μήκος
API Key	μοναδικό κλειδί λογαριασμού
part	Αποκλείονται κάποια κομμάτια από τα

	δεδομένα καιρού κατά την απόκριση βασισμένα στον χρόνο, (τώρα, ανα λεπτό, ανά ώρα, ανα μέρα).
units	διατίθενται οι : στάνταρ, μετρικές και αυτοκρατορικές μονάδες μέτρησης.
lang	διατίθεται πολυγλωσσική υποστήριξη όπου επιστρέφονται τα αποτελέσματα στην επιλεγμένη γλώσσα.

6 Μοντελο δεδομενων

Καθώς υπάρχει πληθώρα σχετικών πόρων που αντλούνται από ετερογενείς πηγές, κρίνεται αναγκαία η διερεύνηση μεθόδου ομογενοποίησης των ληφθέντων δεδομένων και μεταδεδομένων από το διαδίκτυο. Έχει προηγηθεί σχετική έρευνα (που παρουσιάζεται εκτενώς στο παραδοτέο Π1.2) όσον αφορά ποιες πληροφορίες ιδανικά θα μπορούσαν να αξιοποιηθούν από το έργο με τον βέλτιστο τρόπο, τόσο όσον αφορά τους αλγορίθμους δρομολόγησης, προσωποποίησης και συλλογιστικής όσο και την αυτούσια τροφοδότηση χρήσιμων πληροφοριών στον τελικό χρήστη της πλατφόρμας. Αρχική προσέγγιση αποτελεί η δόμηση της συνήθους άναρχης πληροφορίας που καταφθάνει από διαδικτυακές πηγές σε μία ενοποιημένη, καθολική, ειδικού τύπου μορφή JSON, το GEOJSON που αναλύεται στην υποενότητα 6.1. Εν συνεχεία, σειριακά ακολουθεί η αντιστοίχιση (mapping) των πεδίων GEOJSON στις οντότητες της ήδη διαμορφωμένης οντολογίας (όπως παρουσιάζεται εκτενώς στο παραδοτέο Π2.2.1) μέσω αυτοματοποιημένων μεθόδων με σκοπό τον πληθυσμό (population) της γνωσιακής βάσης (Knowledge Base), όπως παρουσιάζεται συνοπτικά στην υποενότητα 6.2.

6.1 GeoJSON

Για τη μεταφορά και την απεικόνιση των δεδομένων έχει επιλεγθεί η μορφή του geoJSON. Το geoJSON βασίζεται στο JavaScript Object Notation (JSON) και το έχει επεκτείνει ώστε να αναπαριστά γεωχωρικά δεδομένα μαζί με τα μη χωρικά στοιχεία τους. Μπορεί να αναπαραστήσει σημεία (points), γραμμές (lines), πολύγωνα (polygons) αλλά και τον συνδυασμό των παραπάνω χωρικών τύπων. Όπως και το JSON είναι γλωσσικά ανεξάρτητο και το καθιστά ιδανικό για επικοινωνία διαφορετικών συστημάτων.

Στη παρούσα υλοποίηση που αναπαριστά σημεία ενδιαφέροντος ο χωρικός τύπος που χρησιμοποιήθηκε είναι του σημείου (point) και για τη πλήρη περιγραφή των δεδομένων του σημείου, επιλέχθηκαν τα κατάλληλα πεδία (attributes).

Παρακάτω ακολουθεί ένα παράδειγμα ενός σημείου ενδιαφέροντος σε μορφή geoJSON.

```
{
  "type": "FeatureCollection",
  "name": "points_of_interest",
  "features": [
    {
      "id": null,
      "type": "Feature",
      "geometry": {
        "type": "Point",
        "coordinates": [
          26.170395,
          40.893967
        ]
      },
      "properties": {
        "wc": null,
        "food": null,
        "name": "Ι. Ν. Παναγίας Κοσμοσώτηρας",
        "dimos": null,
        "email": null,

```

```

"image":
"http://www.religiousgreece.gr/image/journal/article?img_id=329520&t=1445618729164&width=940",
"is_ar": null,
"awards": null,
"refuge": null,
"routes": null,
"unesco": null,
"address": " Φέρες, Τ.Κ. 68500",
"parking": null,
"s_descr": null,
"website": "http://www.imalex.gr/DB735F3B.el.aspx",
"category": [
  "02"
],
"cultural": null,
"name_eng": null,
"shopping": null,
"periferia": null,
"protected": null,
"published": null,
"exhibition": null,
"full_descr": "<p style=\"text-align: justify;\">Στο κέντρο της κωμόπολης των Φερών, στη δυτική όχθη του ποταμού Εβρου, βρίσκεται ένα από τα πιο συμβολικά μνημεία της Θράκης, ο<strong>Ιερός ναός της Παναγίας, άλλοτε καθολικό της Μονής Κοσμοσώτειρας ή της Βήρας</strong>, όπως λεγόταν η τοποθεσία των Φερών στη βυζαντινή εποχή.</p> \n<p style=\"text-align: justify;\">Η μονή ιδρύθηκε το 1152 από τον σεβαστοκράτορα<em>Ισαάκιο Κομνηνό</em>, γιο του<em>Αλεξίου Α΄ Κομνηνού</em><em>και της</em><em>Ειρήνης Δούκαινας</em>, ο οποίος έχει ταφεί μέσα στο καθολικό, κάτω από τον τρουλίσκο της βορειοδυτικής γωνίας. Ο ίδιος μετέφερε εδώ τα γλυπτά του τάφου του, που είχε ετοιμάσει αρχικά στο καθολικό της Μονής της Χώρας, στην Κωνσταντινούπολη. Γύρω από το μοναστήρι σταδιακά αναπτύχθηκε μικρός οικισμός, που οχυρώθηκε με τείχος, καθώς βρισκόταν σε σημαντική θέση, στον δρόμο προς την Αδριανούπολη και την Κωνσταντινούπολη. Μετά το 1371 η μονή κατακτήθηκε από τον<em>Λαλά Σαχίν</em><em>και τον</em><em>Εβρενός</em><em>και το καθολικό της μετατράπηκε σε τέμενος του<em>Σουλεϊμάν πασά</em><em>γιου του</em><em>σουλτάνου Ορχάν</em>. Το 1940 έγιναν εργασίες αποκατάστασης του ναού και ξεκίνησε πάλι η λειτουργία του.</p> \n<p style=\"text-align: justify;\">Από το αρχικό μοναστηριακό συγκρότημα διατηρούνται μόνο τμήματα του οχυρωματικού περιβόλου και το καθολικό.<br> Το<strong>καθολικό</strong>, που απηχεί την τέχνη της Κωνσταντινούπολης του 12ου αιώνα, ανήκει στον αρχιτεκτονικό τύπο του<em>πεντάτρον</em>, απλού δικιόνιου σταυροειδούς εγγεγραμμένου ναού<em>με νάρθηκα στα δυτικά. Στην περίοδο της Τουρκοκρατίας αλλοιώθηκαν τα παράθυρα των τρούλων και των αψίδων και κατεδαφίστηκε ο νάρθηκάς του. Οι τέσσερις αντιρίδες που υπάρχουν εξωτερικά είναι μεταγενέστερες.<br> Στο<strong><em>εσωτερικό</em></strong> του ναού, στη νοτιοανατολική γωνία, σώζεται εντοιχισμένο κεραμικό κόσμημα, στο οποίο απεικονίζεται ο μονοκέφαλος αετός, σύμβολο της δυναστείας των Κομνηνών της Τραπεζούντας. Επίσης, διατηρείται μέρος του τοιχογραφικού διακόσμου, που αποτελεί σημαντικό δείγμα της κωνσταντινοπολίτικης ζωγραφικής των μέσων του 12ου αιώνα. Μεταξύ άλλων εικονίζονται τέσσερις στρατιωτικές μορφές (ανά ζεύγη ανάμεσα στα τοξωτά παράθυρα του βόρειου και νότιου κλίτους), για την ταύτιση των οποίων έχει διατυπωθεί η άποψη ότι πιθανώς πρόκειται για προσωπογραφίες μελών της οικογένειας των Κομνηνών.</p> \n<p>Ο ναός όχι άδικα αποτελεί ένα από τα σημαντικότερα βυζαντινά μνημεία της Ελλάδας και έχει καθιερωθεί ως το<strong>Παγκόσμιο Προσκύνημα των απανταχού Θρακών</strong>.</p>",
"restaurant": null,
"source_cat": "http://www.religiousgreece.gr",
"time_spent": null,
"amea_access": null,
"per_enotita": null,
"s_descr_eng": null,
"subcategory": [],
"ticket_cost": null,
"interchanges": null,
"accommodation": null,
"certification": null,
"info_material": null,
"observatories": null,
"difficulty_lvl": null,
"full_descr_eng": null,
"interchange_id": [],
"kallikratis_id": null,
"historic_period": null,
"outdoor_company": null,
"nearest_neighbour": null,

```

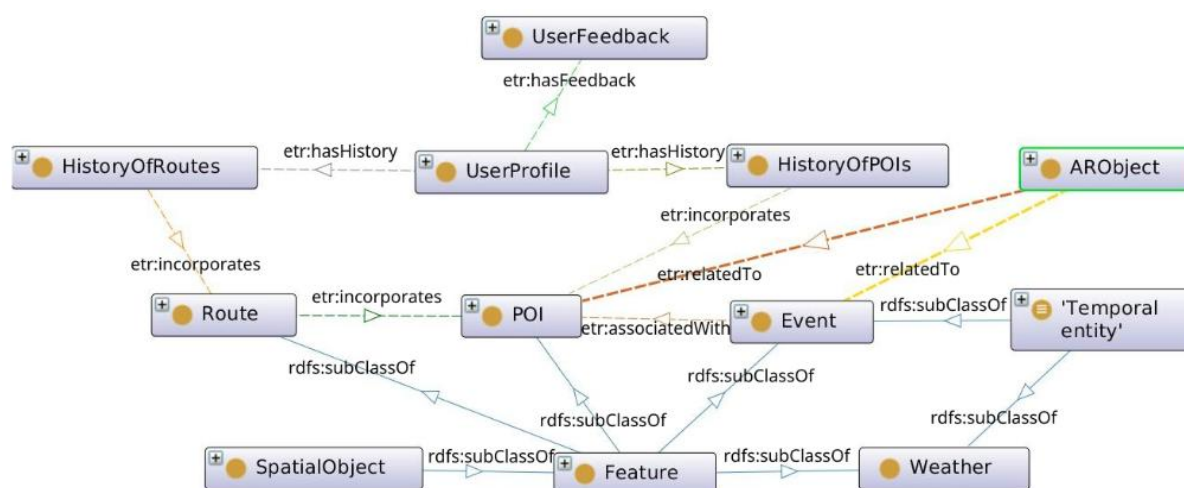
```

    "outdoor_activities": null,
    "recreational_areas": null
  }
}
]
}

```

6.2 Οντολογία

Η χρήση οντολογιών στα πλαίσια του e-ΧΝΗΛΑΤΗ εξυπηρετεί σκοπούς που απαιτούν αποθήκευση, διαχείριση, επεξεργασία και συλλογιστική δεδομένων και ευρύτερης γνώσης με πλούσια σημασιολογία. Στα πλαίσια αυτών των αναγκών, αναπτύσσεται κι ανανεώνεται σε τακτά χρονικά διαστήματα η οντολογία που υποστηρίζει όλα τα προαναφερθέντα. Για τη διαμόρφωσή της, είχε προηγηθεί μια εκτενής έρευνα σχετικά με πόρους που μπορούν να επαναχρησιμοποιηθούν σε συνδυασμό με ορισμό εννοιών εκ νέου τόσο σε επίπεδο οντοτήτων όσο και επίπεδο σχέσεων και ιδιοτήτων. Μερικές από τις οντολογίες που χρησιμοποιήθηκαν είναι οι : LODE για απλοϊκή περιγραφή γεγονότων, η GeoSPARQL για την προσθήκη φυσικής γεωχωρικής υπόστασης σε αντικείμενα και οντότητες, η OWL-Time για προσθήκη χρονικότητας σε αντικείμενα και οντότητες, και τέλος το User Profile για την περιγραφή του προφίλ του χρήστη. Για την πρακτική υλοποίηση των παραπάνω χρησιμοποιήθηκε το ολοκληρωμένο περιβάλλον ανάπτυξης Protégé 5.5.0 και ως γνωσιακή βάση επιλέχθηκε έπειτα από έρευνα η GraphDB 9.1.1 Free Edition. Χρησιμοποιήθηκε επίσης ως γλώσσα η OWL 2 (Web Ontology Language) και η λογική εγκυρότητα της οντολογίας δοκιμάστηκε με διάφορους ενσωματωμένους reasoners στο περιβάλλον ανάπτυξης, όπως επίσης και με τη χρήση της γλώσσας SHACL (Shapes Constraint Language). Στην εικόνα 8 παρουσιάζεται ο συνδυασμός εννοιών και σχέσεων σε ανώτερο επίπεδο.



Εικόνα 8: Ανώτερο επίπεδο εννοιών και σχέσεων οντολογίας e-ΧΝΗΛΑΤΗ

Γίνεται περιγραφή εννοιών με αντίστοιχη πληθώρα ιδιοτήτων όπως: σημείο ενδιαφέροντος (POI), ιστορικό σημείων ενδιαφέροντος (HistoryOfPOIs), γεγονός (Event), αντικείμενο επαυξημένης πραγματικότητας (ARObjct), χωρικό αντικείμενο (SpatialObject), χρονική οντότητα (temporal entity), διαδρομή (Route), ιστορικό

διαδρομών (HistoryOfRoutes), προφίλ χρήστη (UserProfile), ανάδραση χρήστη (UserFeedback) και καιρός (Weather). Φυσικά εκτός από το άνω επίπεδο οντοτήτων υπάρχουν διασυνδέσεις και σε χαμηλότερο επίπεδο για να καλύψουν εξειδικευμένες ανάγκες και να εξυπηρετήσουν ειδικούς σκοπούς.

Τέλος, εκτός από τη γνώση που προέρχεται άμεσα από διαδικασίες ανάκτησης από το διαδίκτυο, με χρήση τόσο αυτόματων μεθόδων όσο και χειροποίητα κατασκευασμένων εξάγεται νέα γνώση από την ήδη προϋπάρχουσα με τη χρήση ερωτημάτων Construct της γλώσσας SPARQL. Για περαιτέρω πληροφορίες, ανάλυση και επεξήγηση αναφορικά με τα σημασιολογικά μοντέλα και τις τεχνικές σημασιολογικού συσχετισμού για τη δημιουργία διαδρομών ενδείκνυται το παραδοτέο Π2.2.1 (M12).

7 Σύνολα δεδομένων

Σε αυτήν την ενότητα παρουσιάζονται τα σύνολα δεδομένων που δημιουργήθηκαν χρησιμοποιώντας τους crawlers του e-ΧΝΗΛΑΤΗ. Οι πόροι που συλλέχθηκαν καλύπτουν τις ανάγκες συλλογής δεδομένων που έχουν καθοριστεί από την αρχή του έργου έως αυτήν την περίοδο αναφοράς.

Σύνολα δεδομένων από ΟΔΥΣΣΕΥΣ

Η ιστοσελίδα του “ΟΔΥΣΣΕΥΣ” πρόκειται για μια επίσημη πηγή πληροφοριών από το ελληνικό υπουργείο πολιτισμού και αθλητισμού. Περιέχει υλικό και πόρους για πολλά μνημεία και μουσεία σε ολόκληρη την ελλαδική επικράτεια. Προσφέρει αναλυτική αναζήτηση μουσείων, μνημείων και αρχαιολογικών χώρων. Πολύ σημαντικό κριτήριο επιλογής για την ανάκτηση πληροφορίας από τον ιστότοπο και τη παραγωγή συνόλου δεδομένων είναι η παροχή γεωχωρικών συντεταγμένων, γραπτών πληροφοριών και πολυμεσικού υλικού που μπορεί να χρησιμοποιηθεί για τη δημιουργία και την προβολή σημείων ενδιαφέροντος στην ακτίνα δράσης του e-ΧΝΗΛΑΤΗ κατά την πρόταση εξατομικευμένων μοναδικών δυναμικών διαδρομών. Παρακάτω παρουσιάζεται ένα παράδειγμα μίας εγγραφής για ένα συγκεκριμένο σημείο ενδιαφέροντος όπως περιεσφαιλήθηκε απευθείας από την επίσημη ιστοσελίδα με την ημι-αυτόματη χρήση λογισμικού.

```
{
  "id": {"$oid": "5c04f45d01c6b83488a1660e"},
  "description": "Στο Αρχαιολογικό Μουσείο Αβδήρων εκτίθενται αντικείμενα που βρέθηκαν στην πόλη και τα νεκροταφεία των αρχαίων Αβδήρων. Η έκθεση καλύπτει τη χρονική περίοδο από τον 7ο αι. π.Χ. μέχρι τον 13ο αι. μ.Χ. Η παρουσίαση των εκθεμάτων γίνεται σε τρεις κύριες ενότητες: δημόσιος βίος, ιδιωτικός βίος και ταφικά έθιμα, οι οποίες υποδιαιρούνται σε επιμέρους θεματικές ομάδες. Τα εκθέματα πλαισιώνονται από επεξηγηματικά κείμενα, φωτογραφίες, σχέδια, χάρτες και διαγράμματα. Το μουσείο έχει έντονο εκπαιδευτικό χαρακτήρα. Κύριος στόχος είναι να συμβάλλει στην επιμόρφωση των μαθητών Πρωτοβάθμιας και Δευτεροβάθμιας εκπαίδευσης, και συγχρόνως να ικανοποιεί τον απαιτητικό επισκέπτη. Η αρχιτεκτονική μελέτη του κτηρίου εκπονήθηκε από τους αρχιτέκτονες μηχανικούς του Υπουργείου Πολιτισμού Γ. Πολυχρονίου και Ν. Φιλιππίδη. Στο εξωτερικό του υπάρχουν αναφορές στα παραδοσιακά κτίσματα της Θράκης. Το κτήριο αποτελείται από το υπόγειο (στο οποίο βρίσκονται τα εργαστήρια, οι αποθήκες αρχαίων και υλικών, οι ηλεκτρολογικές και μηχανολογικές εγκαταστάσεις) το ισόγειο (όπου υπάρχουν τα εργαστήρια συντήρησης, οι χώροι εξυπηρέτησης του κοινού, η αίθουσα πολλαπλών χρήσεων και εκθεσιακός χώρος) και τον όροφο (που περιλαμβάνει γραφεία και εκθεσιακό χώρο). Στα γραφεία κατά τις ώρες 7:00 - 15:00 εξυπηρετείται το κοινό για διοικητικές υποθέσεις. Το Μουσείο υπάγεται στη ΛΑ' Εφορεία Προϊστορικών και Κλασικών Αρχαιοτήτων Ξάνθης. Πρόθεση της Εφορείας είναι να μεταβληθεί σε χώρο πολλαπλών πολιτιστικών δραστηριοτήτων, ώστε να συμβάλλει δυναμικότερα στην πνευματική ανάπτυξη της περιοχής. Έχουν οργανωθεί επιστημονικές ξεναγήσεις και εκπαιδευτικά προγράμματα. Στην αίθουσα πολλαπλών χρήσεων διεξήχθη στις 20-21 Οκτωβρίου 2001 συμπόσιο με θέμα \"Κλαζομενές, Τέως και Αβδηρα: Μητροπόλις και αποικία\". Τον Ιούνιο του 2002 φιλοξενήθηκε συνάντηση Ελλήνων και Ισπανών στο πλαίσιο γενικότερης συνεργασίας της Περιφέρειας Ανατολικής Μακεδονίας - Θράκης με την αντίστοιχη Περιφέρεια της Adra και το Πανεπιστήμιο της Granada. Στους άμεσους πραγματοποιήσιμους στόχους συγκαταλέγεται η οργάνωση φωτογραφικών εκθέσεων, περιοδικών εκθέσεων αρχαιολογικού ή εικαστικού περιεχομένου και η εγκατάσταση οπτικοακουστικού συστήματος.\" , \"hours\": \" Ώρες Λειτουργίας Δευτέρα-Κυριακή: 08:00-16:00 Για περισσότερες πληροφορίες, παρακαλούμε επικοινωνήστε στα τηλέφωνα της Εφορείας Αρχαιοτήτων Ξάνθης (+30 25410 51003, +30 25410 51783). Ημέρες Αργίας 1 Ιανουαρίου 25 Μαρτίου Μ.Παρασκευή: Η έκθεση είναι ανοικτή 12:00 -17.00. 1 Μαΐου: Η έκθεση είναι κλειστή την Τρίτη 03 Μαΐου 2016, λόγω μεταφοράς του εορτασμού της Πρωτομαγιάς. Κυριακή του Πάσχα Χριστούγεννα, 25 Δεκεμβρίου 26 Δεκεμβρίου\", \"image\": \"http://odysseus.culture.gr/java/image?foto_id=1921\\u0026size=m2\", \"name\": \"Αβδήρων Αρχαιολογικό Μουσείο\", \"source\": \"http://odysseus.culture.gr/h/1/gh151.jsp?obj_id=3391\"}
```

Σύνολα δεδομένων από religious greece

Η συγκεκριμένη ιστοσελίδα της ιεράς αρχιεπισκοπής Αθηνών παρέχει πληροφορίες για ορθόδοξους Χριστιανικούς θρησκευτικούς τουριστικούς προορισμούς ανά την

ελλαδική επικράτεια. Πολύ σημαντικό κριτήριο για αυτήν την επιλογή όπως και στην προηγούμενη είναι η παροχή γεωχωρικών συντεταγμένων και στίγματος, γραπτών πληροφοριών για την κάθε οντότητα που παρουσιάζει καθώς και πολυμεσικό υλικό που δυνητικά δύναται να χρησιμοποιηθεί για εμπλουτισμό της εμπειρίας του τελικού χρήστη στην πλατφόρμα του e-ΧΝΗΛΑΤΗ. Παρακάτω παρουσιάζεται ένα παράδειγμα μιας εγγραφής για ένα συγκεκριμένο σημείο ενδιαφέροντος όπως περιεγράφηκε απευθείας από την επίσημη ιστοσελίδα.

```
{
  "id": null,
  "type": "Feature",
  "geometry": {
    "type": "Point",
    "coordinates": [
      22.116404,
      40.408802
    ]
  },
  "properties": {
    "wc": null,
    "food": null,
    "name": "Ιερά Μονή Παναγίας Σουμελάς",
    "dimos": null,
    "email": null,
    "image":
"http://www.religiousgreece.gr/image/journal/article?img_id=344339&t=1445775749050&width=940",
    "is_ar": null,
    "awards": null,
    "refuge": null,
    "routes": null,
    "unesco": null,
    "address": "Καστανιά, Τ.Κ. 59100",
    "parking": null,
    "s_descr": null,
    "website": "http://www.panagiasoumela.gr/",
    "category": [
      "02"
    ],
    "cultural": null,
    "name_eng": null,
    "shopping": null,
    "periferia": null,
    "protected": null,
    "published": null,
    "telephone": "23310 88139",
    "exhibition": null,
    "full_descr": "<p style=\"text-align: justify;\">Στις καταπράσινες πλαγιές του όρους
Βέρμιο, κοντά στο χωριό Καστανιά, βρίσκεται η φημισμένη <strong>Μονή της Παναγίας
Σουμελά</strong>, το πνευματικό κέντρο του ποντιακού ελληνισμού και ένας από τους
σπουδαιότερους προσκυνηματικούς τόπους όλης της Ελλάδας.</p> \n<p style=\"text-align:
justify;\">Η μονή ιδρύθηκε με πρωτοβουλία του ποντιακής καταγωγής συγγραφέα <em><strong>Φίλωνα
Κτενίδη</strong></em> και του σωματούχου <em><strong>Παναγία Σουμελά</strong></em>, στην προσπάθεια αναβίωσης
της ιστορικής ομώνυμης μονής που βρίσκεται στο όρος Μελά, κοντά στην Τραπεζούντα. Η μονή
εκείνη, σύμφωνα με την παράδοση, είχε ιδρυθεί στα τέλη του 4ου αιώνα, και για αιώνες
λειτουργήσε με την εύνοια των βυζαντινών αυτοκρατόρων, έχοντας πλούτο και προνόμια, που
διατήρησε και μετά την κατάλυση της βυζαντινής αυτοκρατορίας, αποκτώντας παράλληλα και
συμβολική αξία μεταξύ των Ποντίων για σχεδόν δεκαέξι αιώνες. Καταστράφηκε από τους Τούρκους το
1922 και τα κειμήλιά της, μαζί με την εικόνα της Παναγίας, μεταφέρθηκαν στην Ελλάδα το 1931
έπειτα από τη μεσολάβηση του <em>Ελευθερίου Βενιζέλου</em> και με τη συμβολή μοναχών που
γνώριζαν τον τόπο φύλαξής τους. Η μονή του Βερμίου θεμελιώθηκε το 1951 σε έκταση που
παραχώρησαν οι κάτοικοι του χωριού Καστανιά, που κατάγονται από τον Πόντο, και τα εγκαίνια της
πραγματοποιήθηκαν το 1952.&nbsp;</p> \n<p style=\"text-align: justify;\">Το συγκρότημα σήμερα
περιλαμβάνει δύο ναούς, πολλούς ξενώνες και κάθε είδους βοηθητικές εγκαταστάσεις για την
εξυπηρέτηση επισκεπτών και προσκυνητών.</p> \n<p style=\"text-align: justify;\">Το καθολικό
είναι μεγαλοπρεπής ναός βυζαντινού ρυθμού και στο εσωτερικό του φυλάσσεται η
<strong>θαυματουργή εικόνα της Παναγίας</strong>, που θεωρείται έργο του
<em><strong>Ευαγγελιστή Λουκά</strong></em>και προέρχεται από την ιστορική μονή του Πόντου.
Επίσης, λειτουργεί πλούσια <strong>βιβλιοθήκη</strong>, καθώς και <strong>εκθετήριο
</strong>πολύτιμων εκκλησιαστικών σκευών, εικόνων και αμφίων.</p> \n<p style=\"text-align:
```



```
justify;\">Η μονή είναι αφιερωμένη στην Κοίμηση της Θεοτόκου και πανηγυρίζει στις 15
Αυγούστου.</p>\",
  \"restaurant\": null,
  \"source_cat\": \"http://www.religiousgreece.gr\",
  \"time_spent\": null,
  \"amea_access\": null,
  \"per_enotita\": null,
  \"s_descr_eng\": null,
  \"subcategory\": [],
  \"ticket_cost\": null,
  \"interchanges\": null,
  \"accommodation\": null,
  \"certification\": null,
  \"info_material\": null,
  \"observatories\": null,
  \"difficulty_lvl\": null,
  \"full_descr_eng\": null,
  \"interchange_id\": [],
  \"kallikratis_id\": null,
  \"historic_period\": null,
  \"outdoor_company\": null,
  \"nearest_neighbour\": null,
  \"outdoor_activities\": null,
  \"recreational_areas\": null
}
}
```

Σύνολα δεδομένων από Very Macedonia

Η ιστοσελίδα <http://verymacedonia.gr/> είναι επίσημα διαχειριζόμενη από την “Περιφέρεια Κεντρικής Μακεδονίας” και πιο συγκεκριμένα από τη “Γενική Διεύθυνση Ανάπτυξης και Περιβάλλοντος”. Το υλικό που διαθέτει, όπως είναι πρόδηλο και από τον φορέα διαχείρισης του, αφορά τοποθεσίες που έγκειται μέσα στην περιφέρεια κεντρικής Μακεδονίας με κατηγορίες όπως: παραλίες, βουνά, φύση, θρησκεία, τη μητρόπολη της Θεσσαλονίκης, γαστρονομία και άλλα. Σημαντικό κριτήριο πέρα από την επισημότητα της ιστοσελίδας αποτελεί η παροχή πλούσιου κατηγοριοποιημένου πολυμεσικού υλικού όπως επίσης και η ύπαρξη επίσημου γεωχωρικού στίγματος για το κάθε σημείο ξεχωριστά. Παρακάτω παρουσιάζεται ένα παράδειγμα μίας εγγραφής για ένα συγκεκριμένο σημείο ενδιαφέροντος όπως αυτό ανακτήθηκε από την ιστοσελίδα με χρήση ημι-αυτόματου λογισμικού.

```
{
  \"company\": [{\"Αξιοθέατο\": \"Άγιος Ιωάννης (ΠΕ Πιερίας)\"}],
  \"metrics\": [[
    {
      \"name\": \"Πληροφορίες\",
      \"citeyear\": 2019,
      \"source\": \"http://verymacedonia.gr/portfolio-item/agios-giannis/?lang=el\",
      \"type\": \"textual\",
      \"value\": \"Γραφική ακτή με πλούσια και χρυσή άμμο, δίπλα στην οργανωμένη ακτή
του Μακρυγιάλου, στην οποία μπορείτε να απολαύσετε την ηρεμία του τοπίου. Ξενοδοχεία και
Ενοικιαζόμενα Δωμάτια περιμένουν να σας φιλοξενήσουν στους οργανωμένους οικισμούς του
Μακρυγιάλου και της Μεθώνης πολύ κοντά στην ακτή. Η ακτή βρίσκεται μόλις λίγα μέτρα βόρεια του
Μακρυγιάλου. http://pieria-tourism.gr/beaches.html\"
    },
    {
      \"name\": \"Κατηγορία\",
      \"citeyear\": 2019,
      \"source\": \"http://verymacedonia.gr/portfolio-item/agios-giannis/?lang=el\",
      \"type\": \"textual\",
      \"value\": \"Παραλίες\"
    },
    {
      \"name\": \"lat-long\",
      \"citeyear\": 2019,
```

```
        "source": "http://verymacedonia.gr/portfolio-item/agios-giannis/?lang=el",  
        "type": "other",  
        "value": "(40.426536,22.605078)"  
    }  
  ]  
}
```


8 Συμπεράσματα & επομενα βήματα

Σε αυτό το παραδοτέο, παρουσιάστηκε η αρχική έκδοση των μεθόδων που αναπτύχθηκαν για τους crawlers του e-ΧΝΗΛΑΤΗ, ένα γενικό πλαίσιο που ενσωματώνει και συνδέει καταληκτικά ένα σύνολο διαφορετικών πόρων βασισμένο στο μόντουλο του scrapping περιεχομένου. Ο στόχος για τον οποίο αναπτύχθηκε το συγκεκριμένο πλαίσιο για το έργο είναι η συλλογή χρήσιμου πολυμεσικού περιεχομένου και μεταδεδομένων που μπορούν εν συνεχεία να αποθηκευτούν και να συμπεριληφθούν σε αλγορίθμους δρομολόγησης, προσωποποίησης και συλλογιστικής από τις υποενότητες ΥΕ2.3 και ΥΕ2.4. Πρόκειται για το πρώτο βήμα μιας ροής διεργασιών προ-επεξεργασίας που δημιουργεί περιεχόμενο που θα χρησιμοποιηθεί από την πλατφόρμα του έργου.

Παρουσιάστηκαν εκτενώς οι στοχευμένες επιλογές διαδικτυακών πόρων που επιλέχθηκαν ύστερα από εκτεταμένη έρευνα και ανάλυση των αναγκών, καθώς και λεπτομέρειες για τις ιστοσελίδες και τον τρόπο με τον οποίο ανακτήθηκαν τα δεδομένα από αυτές, είτε στην περίπτωση της διαδικασίας του scrapping, είτε στην περίπτωση παροχής API από την ίδια την ιστοσελίδα προς εξυπηρέτηση παροχής στοχευμένων πληροφοριών.

Επιπλέον, έγινε λόγος για τη δόμηση της μοντελοποίησης των δεδομένων και τη σημασία της, καθώς κάθε ξεχωριστή διαδικτυακή πηγή είχε εξίσου ξεχωριστά δομημένη τη ζητούμενη πληροφορία προς ανάκτηση. Εν συνεχεία, παρουσιάστηκαν τα τωρινά σύνολα δεδομένων που σχηματίστηκαν από τις διαδικασίες που περιγράφηκαν παραπάνω.

Αναφορικά με τον μελλοντικό σχεδιασμό και τα επόμενα βήματα, σαφώς και θα επιλεγθούν με την ίδια μεθοδικότητα και σχολαστικότητα επίσημες πηγές πληροφοριών για σημεία ενδιαφέροντος ή και ελεύθερα προσβάσιμες γεωβάσεις και ανάλογα θα πραγματοποιηθούν επεκτάσεις του υπάρχοντος λογισμικού πλαισίου προς εξειδικευμένη ανάκτηση εάν αυτό καταστεί απαραίτητο. Πέρα από τα τεχνικώς μοναδικά ζητήματα που καλούμαστε να αντιμετωπίσουμε, ορισμένες φορές συναντώνται προβλήματα όρων χρήσης των πόρων και πνευματικών δικαιωμάτων, γεγονός που μειώνει αισθητά τις επιλογές μας. Θα διερευνηθεί φυσικά και η δυνατότητα ανάκτησης πληροφοριών και από επιπρόσθετες κοινωνικές πλατφόρμες και στοχευμένες ιστοσελίδες. Τέλος, θα διαπιστωθεί και η αναγκαιότητα και δυνατότητα συμπερίληψης μεθόδων ανάκτησης σχετικά με οπτικοακουστικό περιεχόμενο, όπου αυτό προσφέρεται, ενισχύοντας κατά αυτόν τον τρόπο τη συνολική εμπειρία του χρήστη της πλατφόρμας του e-ΧΝΗΛΑΤΗ.

Κατά τη χρονική περίοδο που συντάσσεται το παρόν έγγραφο και έως το μήνα 34, η επιπρόσθετη τεχνολογική εξέλιξη των τεχνικών αυτόματης αναζήτησης και εξόρυξης γνώσης από κοινωνικά δίκτυα και τον ιστό για την υποενότητα εργασίας ΥΕ2.1 θα περιγραφεί εκτενώς στο αντίστοιχο ακόλουθο παραδοτέο Π2.1.2 κατά τον αντίστοιχο μήνα.

αναφορές

Gkatziaki, V., Papadopoulos, S., Mills, R., Diplaris, S., Tsampoulatidis, I., & Kompatsiaris, I. (2018). easIE: Easy-to-use information extraction for constructing CSR databases from the web. *ACM Transactions on Internet Technology (TOIT)*, 18(4), 1-21.

Shaw, R., Troncy, R., & Hardman, L. (2009, December). Lode: Linking open descriptions of events. In *Asian semantic web conference* (pp. 153-167). Springer, Berlin, Heidelberg.

Battle, R., & Kolas, D. (2011). Geosparql: enabling a geospatial semantic web. *Semantic Web Journal*, 3(4), 355-370.

Hobbs, J. R., & Pan, F. (2006). Time ontology in OWL. *W3C working draft*, 27, 133.

Maria, G., Akrivi, K., Costas, V., George, L., & Constantin, H. (2007). Creating an ontology for the user profile: Method and applications. In *Proceedings AI* AI Workshop RCIS*.

Stathopoulos, E. A., Paliokas, I., Meditskos, G., Diplaris, S., Tsafaras, S., Valkouma, E., ... & Tzovaras, D. (2019, October). Smart Discovery of Cultural and Natural Tourist Routes. In *IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume* (pp. 208-214).

Stathopoulos, E. A., Kokkalas, A., Mitsopoulou, E. E., Patenidis, A. T., Meditskos, G., Diplaris, S., ... & Kompatsiaris, I. (2020, June). Knowledge-Based Management and Reasoning on Cultural and Natural Touristic Routes. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 355-367). Springer, Cham.

Olston, C., & Najork, M. (2010). Web crawling. *Foundations and Trends in Information Retrieval*, 4(3), 175-246.

Motik, B., Grau, B. C., Horrocks, I., Wu, Z., Fokoue, A., & Lutz, C. (2009). OWL 2 web ontology language profiles. *W3C recommendation*, 27, 61.

Knublauch, H., & Kontokostas, D. (2017). Shapes Constraint Language (SHACL), *W3C Recommendation*. World Wide Web Consortium.

Παράρτημα

Παράρτημα Α: Παράδειγμα Αρχείου διαμόρφωσης για την ιστοσελίδα της Ιεράς Αρχιεπισκοπής Αθηνών:

```
{
  "url": {
    "base_url": "http://www.religiousgreece.gr"
  },
  "group_of_urls": ["/web/guest/west-central-macedonia", "/web/guest/east-macedonia-thrace"],
  "source_name": "religiousgreece",
  "table_selector": "#dv-top-attr-container div.dv-carousel div.owl-item",
  "company_info": [
    {
      "label": "name",
      "value": {
        "selector": "span.dv-title",
        "type": "text"
      }
    }
  ],
  "metrics": [
    {
      "label": "crawl_to",
      "value": {
        "selector": "div.dv-content > a:nth-child(1)",
        "type": "link"
      }
    }
  ],
  "crawl": {
    "metrics": [{
      "label": "full_descr",
      "value": {
        "selector": ".dv-content-area",
        "type": "text"
      }
    },
    {
      "label": "lat",
      "value": {
        "selector": "a.dv-widget-link",
        "type": "href",
        "replace": {
          "regex": [".*@", " ", "[^,]*z", ".*", ""],
          "with": ["", "", "", ""]
        }
      }
    }
  ],
  {
    "label": "lng",
    "value": {
      "selector": "a.dv-widget-link",
      "type": "href",
      "replace": {
        "regex": [".*@", " ", "[^,]*z", ".*", ""],
        "with": ["", "", "", ""]
      }
    }
  ],
  {
    "label": "Info",
    "value": {
      "selector": ".dv-widget-content:eq(1) li",
```

```
        "type": "text"
    }
    ],
    },
    "store": {
        "format": "json",
        "hd_path": "D:/Documents/CERTH/Crawler easie/Religious
Greece/results_group_url_religious.json"
    },
    "dynamic_page": true
}
```

Πηγαίος κώδικας StaticWrapperGenerator:

```
public class StaticWrapperGenerator extends WrapperGenerator {
    public StaticWrapperGenerator(Configuration configuration) {
        super(configuration);
    }

    @Override
    public void execute() throws InterruptedException, Exception {
        if (configuration.url.relative_url != null) {
            StaticHTMLExtractor wrapper = new StaticHTMLExtractor(
                configuration.url.base_url,
                configuration.url.relative_url
            );

            if (configuration.next_page_selector != null) {
                PaginationIterator pagination = new PaginationIterator(
                    configuration.url.base_url,
                    configuration.url.relative_url,
                    configuration.next_page_selector
                );
                extraction_handler.execute(pagination, configuration);
            } else {
                extraction_handler.execute(wrapper, configuration);
            }
        } else if (configuration.group_of_urls != null && configuration.url.base_url != null &&
configuration.url.relative_url == null) {
            if (configuration.next_page_selector != null) {
                throw new Exception("Pagination can not be conducted in Group Wrapper mode");
            }
            GroupHTMLExtractor wrapper = new GroupHTMLExtractor(configuration.group_of_urls,
configuration.url.base_url);
            extraction_handler.execute(wrapper, configuration);
        }
    }
}
```

Πηγαίος κώδικας DynamicWrapperGenerator:

```
public class DynamicWrapperGenerator extends WrapperGenerator {
    private String ChromeDriverPath;
    public DynamicWrapperGenerator(Configuration configuration, String ChromeDriverPath) {
        super(configuration);
        this.ChromeDriverPath = ChromeDriverPath;
    }

    @Override
    public void execute() throws InterruptedException, Exception {

        if (configuration.url.relative_url != null) {
            DynamicHTMLExtractor wrapper = new DynamicHTMLExtractor(
                configuration.url.base_url,
                configuration.url.relative_url,

```

```

        ChromeDriverPath
    );
    if (configuration.events != null) {
        this.execute_events(wrapper);
    } else {
        extraction_handler.execute(wrapper, configuration);
    }

    wrapper.browser_emulator.close();
    } else if (configuration.group_of_urls != null && configuration.url.base_url != null &&
configuration.url.relative_url == null) {
        HashSet<String> group_of_urls = configuration.group_of_urls;

        Iterator<String> it = group_of_urls.iterator();
        while (it.hasNext()) {
            String current_url = it.next().replace(configuration.url.base_url, "");

            DynamicHTMLExtractor wrapper = new DynamicHTMLExtractor(
                configuration.url.base_url,
                current_url,
                ChromeDriverPath
            );

            if (configuration.events != null) {
                this.execute_events(wrapper);
            } else {
                extraction_handler.execute(wrapper, configuration);
            }

            wrapper.browser_emulator.close();
        }
    } else {
        throw new Exception("relative_url is not defined");
    }
}

private void execute_events(DynamicHTMLExtractor wrapper) throws InterruptedException,
Exception {
    if (configuration.events instanceof ArrayList) {
        ArrayList<Configuration.Event> list_of_events = (ArrayList<Configuration.Event>)
configuration.events;
        for (int i = 0; i < list_of_events.size(); i++) {
            if (list_of_events.get(i).type.equals("CLICK")) {
                int times_to_repeat = list_of_events.get(i).times_to_repeat;
                for (int j = 0; j < times_to_repeat; j++) {
                    wrapper.browser_emulator.clickEvent(list_of_events.get(i).selector);
                }
                extraction_handler.execute(wrapper, configuration);
            } else if (list_of_events.get(i).type.equals("SCROLL_DOWN")) {
                int times_to_repeat = list_of_events.get(i).times_to_repeat;
                for (int j = 0; j < times_to_repeat; j++) {
                    wrapper.browser_emulator.scrollDownEvent();
                }
                extraction_handler.execute(wrapper, configuration);
            }
        }
    } else {
        Configuration.Event event = (Configuration.Event) configuration.events;
        System.out.println(event.extraction_type);
        if (event.type.equals("CLICK")) {
            if (event.extraction_type.equals("AFTER_ALL_EVENTS")) {
                for (int i = 0; i < event.times_to_repeat; i++) {
                    wrapper.browser_emulator.clickEvent(event.selector);
                }
                extraction_handler.execute(wrapper, configuration);
            } else {
                extraction_handler.execute(wrapper, configuration);
                for (int i = 0; i < event.times_to_repeat; i++) {
                    wrapper.browser_emulator.clickEvent(event.selector);
                    extraction_handler.execute(wrapper, configuration);
                }
            }
        }
    }
}

```

```
    }  
    } else {  
  
        if (event.times_to_repeat != null) {  
            wrapper.browser_emulator.scrollDownEvent(event.times_to_repeat);  
        } else {  
            wrapper.browser_emulator.scrollDownEvent();  
        }  
        extraction_handler.execute(wrapper, configuration);  
    }  
}  
}
```

Παράδειγμα απόκρισης Ticketmaster Event Search API για ένα γεγονός

```
{  
  "_embedded": {  
    "events": [  
      {  
        "name": "Harry Styles | Platinum",  
        "type": "event",  
        "id": "Z698xZu0Za675",  
        "test": false,  
        "url": "https://www.ticketmaster.cz/event/harry-styles-%7C-platinum-tickets/28939?language=en-us",  
        "locale": "en-us",  
        "images": [  
          {  
            "ratio": "3_2",  
            "url": "https://s1.ticketm.net/dam/a/fc3/2d35621d-71b3-46c6-87af-e5163f3c9fc3_1206431_TABLET_LANDSCAPE_3_2.jpg",  
            "width": 1024,  
            "height": 683,  
            "fallback": false  
          }  
        ],  
        { },  
        { },  
        { },  
        { },  
        { },  
        { },  
        { },  
        { },  
        { }  
      ],  
      "sales": {  
        "public": {  
          "startDateTime": "2019-11-22T10:44:00Z",  
          "startTBD": false,  
          "endDateTime": "2021-02-23T22:00:00Z"  
        }  
      },  
      "dates": {  
        "access": {  
          "startDateTime": "2021-02-23T18:30:00Z",  
          "startApproximate": false,  
          "endApproximate": false  
        }  
      },  
      "start": {  
        "localDate": "2021-02-23",  
        "localTime": "20:00:00",  
        "dateTime": "2021-02-23T19:00:00Z",  
        "dateTBD": false,  
      }  
    }  
  }  
}
```

```

        "dateTBA":false,
        "timeTBA":false,
        "noSpecificTime":false
    },
    "timezone":"Europe/Prague",
    "status":{
        "code":"rescheduled"
    },
    "spanMultipleDays":false
},
"classifications":[
    {
        "primary":true,
        "segment":{
            "id":"KZFzniwnSyZfZ7v7nJ",
            "name":"Music"
        }
    },
    "genre":{
        "id":"KnvZfZ7vAeA",
        "name":"Rock"
    },
    "subGenre":{
        "id":"KZazBEonSMnZfZ7v6F1",
        "name":"Pop"
    },
    "family":false
}
],
"promoter":{
    "id":"923",
    "name":"LIVE NATION CZECH REPUBLIC s.r.o."
},
"promoters":[
    {
        "id":"923",
        "name":"LIVE NATION CZECH REPUBLIC s.r.o."
    }
]
,
"_links":{
    "self":{
        "href":"/discovery/v2/events/Z698xZu0Za675?locale=en-us"
    }
},
"attractions":[
    {
        "href":"/discovery/v2/attractions/K8vZ9174XZ0?locale=en-us"
    }
],
{ }
],
"venues":[
    {
        "href":"/discovery/v2/venues/Z598xZu0Zdv61?locale=en-us"
    }
]
],
"_embedded":{
    "venues":[

```



```

        {
            "name": "O2 Arena",
            "type": "venue",
            "id": "Z598xZu0Zdv61",
            "test": false,
            "url": "https://www.ticketmaster.cz/venue/o2-arena-praha-9-
vstupenky/o2arena/105",
            "locale": "en-us",
            "images": [
                {
                    "ratio": "16_9",
                    "url": "https://media.ticketmaster.eu/czechrepublic/b18a88b1506370ac889a0bd8d26da3bb.jpg",
                    "width": 205,
                    "height": 115,
                    "fallback": false
                }
            ],
            "postalCode": "19000",
            "timezone": "Europe/Prague",
            "city": {
                "name": "Praha 9"
            },
            "country": {
                "name": "Czech Republic",
                "countryCode": "CZ"
            },
            "address": {
                "line1": "Českomoravská 2345/17"
            },
            "location": {
                "longitude": "14.49377",
                "latitude": "50.10479"
            },
            "upcomingEvents": {
                "_total": 32,
                "mfx-cz": 32
            },
            "_links": {
                "self": {
                    "href": "/discovery/v2/venues/Z598xZu0Zdv61?locale=en-us"
                }
            }
        },
        {
            "attractions": [
                {
                    "name": "Harry Styles",
                    "type": "attraction",
                    "id": "K8vZ9174XZ0",
                    "test": false,
                    "url": "http://www.ticketmaster.cz/artist/984352",
                    "locale": "en-us",
                    "externalLinks": {
                        "youtube": [
                            {
                                "url": "https://www.youtube.com/channel/UCbOCbp5gXL8jigIBZLqMPrw"
                            }
                        ]
                    }
                }
            ],
            "twitter": [

```

```
        {
            "url": "https://twitter.com/Harry_Styles"
        }
    ],
    "itunes": [
        {
            "url": "https://itunes.apple.com/us/artist/id471260289"
        }
    ],
    "lastfm": [
        {
            "url": "https://www.last.fm/music/Harry%20Styles"
        }
    ],
    "wiki": [
        {
            "url": "https://en.wikipedia.org/wiki/Harry_Styles"
        }
    ],
    "facebook": [
        {
            "url": "https://www.facebook.com/harrystyles"
        }
    ],
    "instagram": [
        {
            "url": "https://instagram.com/harrystyles/"
        }
    ],
    "musicbrainz": [
        {
            "id": "7eb1ce54-a355-41f9-8d68-e018b096d427"
        }
    ],
    "homepage": [
        {
            "url": "http://hstyles.co.uk/"
        }
    ]
},
    "images": [
        {
            "ratio": "3_2",
            "url": "https://s1.ticketm.net/dam/a/fc3/2d35621d-71b3-46c6-87af-e5163f3c9fc3_1206431_TABLET_LANDSCAPE_3_2.jpg",
            "width": 1024,
            "height": 683,
            "fallback": false
        },
        { },
        { },
        { },
        { },
        { },
        { }
    ],
```

```

        { },
        { },
        { }
    ],
    "classifications":[
        {
            "primary":true,
            "segment":{
                "id":"KZFzniwnSyZfZ7v7nJ",
                "name":"Music"
            },
            "genre":{
                "id":"KnvZfZ7vAev",
                "name":"Pop"
            },
            "subGenre":{
                "id":"KZazBEonSMnZfZ7vkE1",
                "name":"Pop Rock"
            },
            "type":{
                "id":"KZAyXgnZfZ7v7lt",
                "name":"Event Style"
            },
            "subType":{
                "id":"KZFzBErXgnZfZ7vA6J",
                "name":"Concert"
            },
            "family":false
        }
    ],
    "upcomingEvents":{
        "_total":101,
        "tmr":8,
        "mfx-de":17,
        "mfx-es":3,
        "mfx-at":3,
        "mfx-dk":4,
        "mfx-nl":3,
        "ticketmaster":57,
        "mfx-cz":3,
        "mfx-pl":3
    },
    "_links":{
        "self":{
            "href":"/discovery/v2/attractions/K8vZ9174XZ0?locale=en-us"
        }
    }
},
{
    "name":"King Princess",
    "type":"attraction",
    "id":"K8vZ9179uW0",
    "test":false,
    "url":"http://www.ticketmaster.cz/artist/999960",
    "locale":"en-us",
    "externalLinks":{
        "musicbrainz":[
            {
                "id":"158778fe-55c1-4dd5-8a3a-9d1b968a5e52"
            }
        ]
    }
}

```

```
]
},
    "images": [
        {
            "ratio": "3_2",
            "url": "https://sl.ticketm.net/dam/a/a42/56a913a9-ba02-4f7d-b8b1-
d68fae76da42_1223031_ARTIST_PAGE_3_2.jpg",
            "width": 305,
            "height": 203,
            "fallback": false
        },
        { },
        { },
        { },
        { },
        { },
        { },
        { },
        { },
        { }
    ],
    "classifications": [
        {
            "primary": true,
            "segment": {
                "id": "KZFzniwnSyZfZ7v7nJ",
                "name": "Music"
            },
            "genre": {
                "id": "KnvZfZ7vAev",
                "name": "Pop"
            },
            "subGenre": {
                "id": "KZazBEonSMnZfZ7vk1t",
                "name": "Pop"
            },
            "type": {
                "id": "KZAyXgnZfZ7v71a",
                "name": "Individual"
            },
            "subType": {
                "id": "KZFzBErXgnZfZ7vAd7",
                "name": "Musician"
            },
            "family": false
        }
    ],
    "upcomingEvents": {
        "_total": 6,
        "ticketmaster": 3,
        "mfx-cz": 3
    },
    "_links": {
        "self": {
            "href": "/discovery/v2/attractions/K8vZ9179uW0?locale=en-us"
        }
    }
}
```

```

]
}
},
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { },
    { }
]
},
  "_links":{
    "first":{
      "href":"/discovery/v2/events.json?page=0&size=20"
    },
    "self":{
      "href":"/discovery/v2/events.json"
    },
    "next":{
      "href":"/discovery/v2/events.json?page=1&size=20"
    },
    "last":{
      "href":"/discovery/v2/events.json?page=5918&size=20"
    }
  },
  "page":{
    "size":20,
    "totalElements":118375,
    "totalPages":5919,
    "number":0
  }
}

```

Παράδειγμα απόκρισης Open Weather One Call API για ένα σημείο

```
{
  "lat": 33.44,
  "lon": -94.04,
  "timezone": "America/Chicago",
  "timezone_offset": -18000,
  "current": {
    "dt": 1588935779,
    "sunrise": 1588936856,
    "sunset": 1588986260,
    "temp": 16.75,
    "feels like": 16.07,

```

```
"pressure": 1009,
"humidity": 93,
"dew_point": 15.61,
"uvi": 8.97,
"clouds": 90,
"visibility": 12874,
"wind_speed": 3.6,
"wind_deg": 280,
"weather": [
  {
    "id": 501,
    "main": "Rain",
    "description": "moderate rain",
    "icon": "10n"
  },
  {
    "id": 200,
    "main": "Thunderstorm",
    "description": "thunderstorm with light rain",
    "icon": "11n"
  }
],
"rain": {
  "1h": 2.79
},
"minutely": [
  {
    "dt": 1588935780,
    "precipitation": 2.789
  },
  ...
],
"hourly": [
  {
    "dt": 1588935600,
    "temp": 16.75,
    "feels_like": 13.93,
    "pressure": 1009,
    "humidity": 93,
    "dew_point": 15.61,
    "clouds": 90,
    "wind_speed": 6.66,
    "wind_deg": 203,
    "weather": [
      {
        "id": 501,
        "main": "Rain",
        "description": "moderate rain",
        "icon": "10n"
      }
    ],
    "rain": {
      "1h": 2.92
    }
  },
  ...
]
"daily": [
  {
    "dt": 1588960800,
    "sunrise": 1588936856,
    "sunset": 1588986260,
    "temp": {
      "day": 22.49,
      "min": 10.96,
      "max": 22.49,
      "night": 10.96,
      "eve": 18.45,
      "morn": 18.14
    },
    "feels_like": {
      "day": 18.72,
```

```
    "night": 6.53,  
    "eve": 16.34,  
    "morn": 16.82  
  },  
  "pressure": 1014,  
  "humidity": 60,  
  "dew_point": 14.35,  
  "wind_speed": 7.36,  
  "wind_deg": 342,  
  "weather": [  
    {  
      "id": 502,  
      "main": "Rain",  
      "description": "heavy intensity rain",  
      "icon": "10d"  
    }  
  ],  
  "clouds": 68,  
  "rain": 15.38,  
  "uvi": 8.97  
},  
...  
}
```