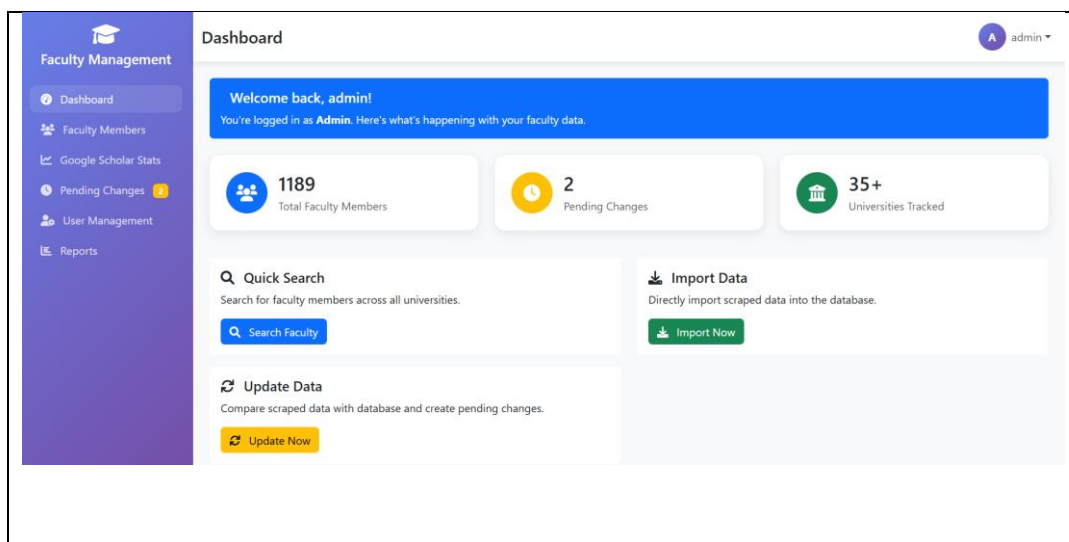


ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Αυτόματος προσδιορισμός χαρακτηριστικών
ακαδημαϊκού προσωπικού μέσω ανάπτυξης web
services»



Φοιτητής
Μόσχος Αλέξανδρος Βίκτωρας
Αρ. Μητρώου: 185233

Επιβλέπων
Ονοματεπώνυμο Ουγιάρογλου
Στέφανος
Βαθμίδα Επίκουρος Καθηγητής

Σεπτέμβριος 2025

Αυτόματος προσδιορισμός χαρακτηριστικών ακαδημαϊκού προσωπικού μέσω ανάπτυξης web services

Κωδικός Π.Ε. 22301

Ονοματεπώνυμο φοιτητή Μόσχος Αλέξανδρος Βίκτωρας

Ονοματεπώνυμο εισηγητή Ουγιάρογλου Στέφανος

Ημερομηνία ανάληψης Π.Ε. 27/10/2022

Ημερομηνία περάτωσης Π.Ε. 12/09/2025

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως πτυχιακή εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Μόσχου Αλέξανδρου Βίκτωρα που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης, ο συγγραφέας/δημιουργός εκχωρεί στο Διεθνές Πανεπιστήμιο της Ελλάδος άδεια χρήσης του δικαιώματος αναπαραγωγής, δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσης της εργασίας διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος. Η ανοικτή πρόσβαση στο πλήρες κείμενο της εργασίας, δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού, ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, πώληση, εμπορική χρήση, διανομή, έκδοση, μεταφόρτωση (downloading), ανάρτηση (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραίτητως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

Πρόλογος

Η επιλογή αυτής της Πτυχιακής Εργασίας προέκυψε από την ανάγκη για την αυτοματοποίηση μιας διαδικασίας που γινόταν από μέλη του τμήματος με σκοπό την οργάνωση ακαδημαϊκών profile απο τμήματα Πληροφορικής καθώς και παρακολούθηση των επιδόσεων οποιων δραστηριοποιούνται στο Google Scholar. Ένα μεγάλο μέρος της ενημέρωσης των δεδομένων γίνεται χειροκίνητα, όπως οι προσλήψεις, αποχωρήσεις αλλά και οι προαγωγές καθηγητών. Ο αυτοματισμός υπήρχε μόνο στην ενημέρωση των δεδομένων από το Google Scholar, με την προϋπόθεση ότι έχει καταχωρηθεί στην βάση δεδομένων το Google Scholar ID του ακαδημαϊκού μέλους από πριν. Γι' αυτό σκεφτήκαμε πως χρειαζόμαστε μια αυτοματοποιημένη λύση όσο για την ανάκτηση δεδομένων από διάφορες ιστοσελίδες τμημάτων, όσο και για την εύκολη προβολή των στοιχείων αυτών. Στόχος της πτυχιακής είναι η δημιουργία ενός scraping microservice για την άντληση των στοιχείων ακαδημαϊκού προσωπικού χωρίς συγκεκριμένα μοντέλα για την κάθε ιστοσελίδα, καθώς και μιας web εφαρμογής για την διαχείριση και προβολή των στοιχείων αυτών.

Περίληψη

Η παρούσα πτυχιακή εργασία παρουσιάζει την ανάπτυξη ενός ολοκληρωμένου συστήματος διαχείρισης ακαδημαϊκού προσωπικού που αποτελείται από δύο βασικές υπηρεσίες: ένα web scraping microservice και μια web εφαρμογή διαχείρισης δεδομένων. Το scraping microservice υλοποιήθηκε σε Python χρησιμοποιώντας βιβλιοθήκες ανάλυσης web περιεχομένου και HTTP requests, ενώ διαθέτει ένα ευέλικτο σύστημα διαμόρφωσης που επιτρέπει την προσαρμογή σε διαφορετικές δομές ιστοσελίδων τμημάτων χωρίς την ανάγκη για συγκεκριμένα μοντέλα. Το σύστημα ενσωματώνει έξυπνους αλγόριθμους ανάλυσης HTML για την αυτόματη ανάκτηση στοιχείων ακαδημαϊκού προσωπικού, ενώ διατηρεί την συμμόρφωση με τους κανόνες web crawling. Η web εφαρμογή αναπτύχθηκε με σύγχρονο web framework και παρέχει RESTful API endpoints για τη διαχείριση χρηστών, ακαδημαϊκού προσωπικού και αλλαγών. Το σύστημα περιλαμβάνει έλεγχο ταυτότητας με secure tokens, διαχείριση ρόλων και ασφαλή επεξεργασία κωδικών πρόσβασης. Χρησιμοποιήθηκε η υπάρχουσα βάση δεδομένων που περιείχε πληροφορίες σχετικά με το ακαδημαϊκό προσωπικό και Google Scholar, ενώ το σύστημα ενσωματώνει αυτόματα τα δεδομένα από το Google Scholar μέσω έξυπνων αλγόριθμων fuzzy matching. Το σύστημα διαθέτει δυνατότητες όπως την αυτόματη ανάλυση στατιστικών επιδόσεων, τη παρακολούθηση αλλαγών από τις ιστοσελίδες, τη δημιουργία αναφορών και την οπτικοποίηση των δεδομένων μέσω ενός responsive web interface. Η αρχιτεκτονική του microservice επιτρέπει την εύκολη επέκταση και προσαρμογή σε νέες πηγές δεδομένων, ενώ η χρήση caching mechanisms βελτιώνει σημαντικά την απόδοση, σεβασζόμενη τις πολιτικές crawling της κάθε ιστοσελίδας.

« Automatic Determination of Academic Staff Characteristics through Web Services Development »

«Alexandros Viktoras Moschos»

Abstract

This thesis presents the development of a comprehensive academic staff management system consisting of two core services: a web scraping microservice and a web application for data management. The scraping microservice was implemented in Python using web content analysis libraries and HTTP requests, while featuring a flexible configuration system that allows adaptation to different department website structures without the need for specific models. The system integrates intelligent HTML parsing algorithms for automatic retrieval of academic staff information, while maintaining compliance with web crawling policies. The web application was developed using a modern web framework and provides RESTful API endpoints for user management, academic staff, and changes. The system includes authentication with secure tokens, role management, and secure password processing. The existing database containing information about academic staff and Google Scholar was utilized, while the system automatically integrates data from Google Scholar through intelligent fuzzy matching algorithms. The system features capabilities such as automatic performance statistics analysis, monitoring changes from websites, report generation, and data visualization through a responsive web interface. The microservice architecture allows easy expansion and adaptation to new data sources, while the use of caching mechanisms significantly improves performance, respecting the crawling policies of each website.

Περιεχόμενα

Πρόλογος.....	4
Περίληψη.....	5
Abstract	6
Περιεχόμενα	8
Κατάλογος Σχημάτων	11
Κατάλογος Πινάκων.....	11
Συντομογραφίες.....	13
Κεφάλαιο 1ο: Εισαγωγή	14
1.1 Εισαγωγή.....	14
1.2 Το Ακαδημαϊκό Προσωπικό στην Ελλάδα.....	14
1.3 Επιστημομετρία.....	15
1.4 Διαδικτυακοί τόποι επιστημομετρίας.....	15
1.5 Κίνητρο	16
1.6 Συνεισφορά.....	16
1.7 Οργάνωση της Εργασίας.....	17
Κεφάλαιο 2ο: Εφαρμογές	17
2.1 OMEA Citations.....	17
2.2 HCSR	18
Κεφάλαιο 3ο: Προσέγγιση.....	19
3.1 Η Προσέγγιση της Γενικής Σχεδίασης	19
3.2 Web Scrapers.....	19
3.3 Web Crawlers vs Scrapers.....	21
3.4 Web Crawlers: Ορισμοί και πλαίσιο χρήσης	22
3.5 Αρχιτεκτονική crawler και ευπρέπεια πρόσβασης.....	22
3.6 Κανονικοποίηση διευθύνσεων και αποφυγή διπλοεγγραφών	22
3.7 Εστιασμένο crawling σε ιστοσελίδες	23
3.8 Δυναμικό περιεχόμενο και headless browsers	23

3.9	Σύνδεση με την κανονικοποίηση βαθμίδων ΔΕΠ στο ελληνικό σύστημα.....	23
Κεφάλαιο 4ο:	Αρχιτεκτονική Συστήματος και Επιλογή Τεχνολογιών	24
4.1	Η Αρχιτεκτονική του Συστήματος	24
4.2	Οι Τεχνολογίες που Επιλέχθηκαν	24
4.3	Python ως Βασική Γλώσσα Προγραμματισμού	25
4.4	FastAPI για το Web Framework	25
4.5	BeautifulSoup4 για το HTML Parsing	26
4.6	Bootstrap για το Frontend Design	26
4.7	MariaDB ως Σύστημα Διαχείρισης Βάσης Δεδομένων	26
4.8	Jinja2 για το Template Engine.....	27
Κεφάλαιο 5ο:	Το Scraping Microservice: Ροές, Διαμόρφωση και Επεξεργασία Δεδομένων	27
5.1	Στόχοι και λειτουργικές απαιτήσεις του microservice	27
5.2	Επισκόπηση αρχιτεκτονικής και ροή δεδομένων	28
5.3	Παραμετροποίηση Web Scraper, δομή και ιδιότητες.....	29
5.4	Δομή και βασικές κατηγορίες παραμέτρων	29
5.5	Ανάλυση παραμέτρου parser type.....	29
5.5.1	Generic	30
5.5.2	Container	31
5.5.3	Table	32
5.5.4	Sequential	34
5.6	Συμμόρφωση με Robots Exclusion Protocol.....	36
5.7	Caching HTML: Σκοπός, Αρχές και Εφαρμογή	36
5.8	Parsing Pipeline: Λήψη, Ανάλυση, Κανονικοποίηση, Εξαγωγή.....	37
Κεφάλαιο 6ο:	Η Web Εφαρμογή - Ρόλος, Ροές και Διεπαφές.....	38
6.1	Σκοπός και ρόλος της web εφαρμογής.....	38
6.2	Αυθεντικοποίηση με JSON Web Tokens (JWT).....	40
6.3	Ροές εργασίας και διασύνδεση με το microservice	41
6.4	API διεπαφές (REST endpoints, σχήματα)	42
6.5	Διάγραμμα βάσης δεδομένων.....	42
6.6	Public Endpoints.....	43

6.7	Endpoints με χρήση JWT	43
6.8	Σελίδα Faculty (κατάλογος, αναζήτηση/φίλτρα/ταξινόμηση, σελιδοποίηση).....	45
6.9	Σελίδα Pending Changes (έλεγχος και έγκριση αλλαγών).....	46
6.10	Σελίδα Google Scholar Stats (μετρικές και αντιστοιχίσεις)	47
6.10.1	Ομοιότητα χαρακτήρων (SequenceMatcher)	48
6.10.2	Δείκτης Jaccard	48
6.10.3	Κανονικοποιημένη επικάλυψη λέξεων	48
6.11	Σελίδα Reports (αναφορές και εξαγωγές δεδομένων).....	49
Κεφάλαιο 7ο:	Αποτελέσματα και Συμπεράσματα	50
7.1	Μεθοδολογία αξιολόγησης και μετρικές.....	50
7.2	Αποτελέσματα: ακρίβεια, αστοχίες και παραδείγματα	51
7.3	Συμπεράσματα.....	53
7.4	Μελλοντικές επεκτάσεις.....	53
7.5	Επίλογος.....	53
BIBΛΙΟΓΡΑΦΙΑ.....		54
ΠΑΡΑΡΤΗΜΑ ΚΩΔΙΚΑ.....		57

Κατάλογος Σχημάτων

Πίνακας 5.1: Δείγμα για name matching (σύγκριση HTML snapshot vs JSON αρχείου)

Πίνακας 5.2: Δείγμα για role matching (σύγκριση HTML snapshot vs JSON αρχείου)

Πίνακας 5.3: Name matching (Summary A)

Πίνακας 5.4: Rank matching στα Matched Names (Summary B)

Κατάλογος Πινάκων

Σχήμα 2.1: Αρχική σελίδα της εφαρμογής OMEA

Σχήμα 2.2: Αρχική σελίδα της εφαρμογής HCSR

Σχήμα 3.1: Διάγραμμα αρχιτεκτονικής Scrapy

Σχήμα 3.2: DOM Model

Σχήμα 4.1: Παράδειγμα interactive API

Σχήμα 4.2: Ιδιότητες ACID

Σχήμα 5.1: Αρχείο παραμετροποίησης scraper

Σχήμα 5.2: Περιεχόμενο με καθηγητή και βαθμίδα

Σχήμα 5.3: Περιεχόμενο χωρίς ιδιότητες αναγνώρισης / τύπου Container

Σχήμα 5.4: Παράδειγμα ρυθμίσεων για τύπου Table

Σχήμα 5.5: Περιεχόμενο τύπου Table

Σχήμα 5.6: Περιεχόμενο τύπου Sequential.

Σχήμα 5.7: Ροή δημιουργίας και ανάκτησης Cached περιεχομένου

Σχήμα 5.8: Διάγραμμα ροής Web Scraper

Σχήμα 6.1: Αρχική οθόνη διαχειριστή

Σχήμα 6.2: Διάγραμμα ροής για αυθεντικοποίηση μέσω JWT

Σχήμα 6.3: Διάγραμμα ER βάσης δεδομένων

Σχήμα 6.4: Αρχική οθόνη σύνδεσης

Σχήμα 6.5: Οθόνη διαχείρισης προσωπικού

Σχήμα 6.6: Οθόνη έγκρισης/απόρριψης προτεινόμενων αλλαγών

Σχήμα 6.7: Οθόνη προβολής βιβλιομετρικών δεδομένων

Σχήμα 6.8: Οθόνη προβολής αναφορών και αναλύσεων

Συντομογραφίες

Δ.Ε.	Διπλωματική Εργασία
ΔΙΠΑΕ	Διεθνές Πανεπιστήμιο Ελλάδος
Π.Ε.	Πτυχιακή Εργασία
GSID	Google Scholar Identifier
ΔΕΠ	Διδακτικό και Ερευνητικό Προσωπικό
ΕΔΠΠ	Ειδικό Διδακτικό Προσωπικό
ΕΤΕΠ	Ειδικό Τεχνικό Εργαστηριακό Προσωπικό
HCSR	Hellenic Computer Science Research
NLP	Natural Language Processing

Κεφάλαιο 1ο: Εισαγωγή

1.1 Εισαγωγή

Η συλλογή και η εξόρυξη δεδομένων (data collection και data mining) αποτελούν θεμέλια για τη δημιουργία αξιόπιστης γνώσης σε κάθε οργανισμό. Στην πράξη, τα δεδομένα προέρχονται από πολλές, ετερογενείς πηγές και με διαφορετικές δομές, απαιτώντας διαδικασίες εντοπισμού, ανάκτησης, καθαρισμού, ενοποίησης και συνεχούς επικαιροποίησης. Οι κύριες προκλήσεις περιλαμβάνουν τον κατακερματισμό της πληροφορίας, την έλλειψη τυποποίησης, τις συχνές αλλαγές περιεχομένου, τους περιορισμούς πρόσβασης και τις πολιτικές των ιστοτόπων, καθώς και ζητήματα ποιότητας, ακρίβειας και ιχνηλασιμότητας (ποιος άλλαξε τι και πότε). Όταν αυτές οι εργασίες εκτελούνται χειροκίνητα, αυξάνεται το κόστος σε χρόνο και σφάλματα, ενώ μειώνεται η συχνότητα ενημέρωσης και η δυνατότητα αξιοποίησης των δεδομένων για ορθή ανάλυση.

Στον ελληνικό ακαδημαϊκό χώρο, απουσιάζει ένα κεντρικό, ελεύθερα προσβάσιμο μητρώο με πληροφορίες για το προσωπικό, με αποτέλεσμα τα σχετικά στοιχεία να είναι διασκορπισμένα σε ιστοσελίδες ιδρυμάτων και τμημάτων, συχνά με διαφορετική μορφοποίηση και βαθμό πληρότητας. Αυτό οδηγεί σε αποσπασματικές και όχι πάντα επίκαιρες καταχωρήσεις, αυξάνει την εξάρτηση από χειροκίνητες διαδικασίες και δυσχεραίνει τόσο τον συντονισμό όσο και την αναλυτική αξιοποίηση των δεδομένων. Η ανάγκη για αυτοματοποίηση και κεντρικοποίηση —μέσω τυποποιημένων ροών συλλογής, ελέγχου ποιότητας και ασφαλούς αποθήκευσης— καθίσταται επιτακτική, ώστε τα δεδομένα να γίνονται πιο αξιόπιστα, διαλειτουργικά και άμεσα αξιοποιήσιμα από την ακαδημαϊκή κοινότητα.

1.2 Το Ακαδημαϊκό Προσωπικό στην Ελλάδα

Στα ελληνικά ΑΕΙ, το ακαδημαϊκό προσωπικό οργανώνεται σε διακριτές κατηγορίες με συγκεκριμένα καθήκοντα και πλαίσια εξέλιξης. Κεντρική κατηγορία αποτελεί το Διδακτικό και Ερευνητικό Προσωπικό (ΔΕΠ), με τρεις ενεργές βαθμίδες: Καθηγητής, Αναπληρωτής Καθηγητής και Επίκουρος Καθηγητής. Η βαθμίδα του Λέκτορα έχει καταργηθεί για νέες εκλογές, με διατήρηση υφιστάμενων θέσεων έως εξέλιξη ή αποχώρηση. Παράλληλα, το Ειδικό Διδακτικό Προσωπικό (ΕΔΠ) επικεντρώνεται στη διδασκαλία και την εκπαιδευτική υποστήριξη, ενώ το Ειδικό Τεχνικό Εργαστηριακό Προσωπικό (ΕΤΕΠ) παρέχει τεχνική και εργαστηριακή υποστήριξη στη διδακτική/ερευνητική διαδικασία. Η θεσμική οριοθέτηση των κατηγοριών και αρμοδιοτήτων τους καθορίζεται από το ισχύον νομοθετικό πλαίσιο των ΑΕΙ.[40][41]

Η ακαδημαϊκή εξέλιξη στα μέλη ΔΕΠ τυπικά ακολουθεί πορεία Επίκουρος → Αναπληρωτής → Καθηγητής, μέσω ανοικτών προκηρύξεων και αξιολόγησης που συνεκτιμά ερευνητικό έργο, διδακτική εμπειρία και διοικητική συμβολή. Στα μέλη ΕΔΠ/ΕΤΕΠ η εσωτερική εξέλιξη (συνήθως από βαθμίδα Β' σε Α') πραγματοποιείται κατόπιν αξιολόγησης και τήρησης τυπικών προϋποθέσεων. Για ανάγκες συστημάτων καταγραφής/ανάλυσης όπως της παρούσας εργασίας, η πολυτυπία τίτλων και ορολογίας (μεταφράσεις, συντμήσεις, ιστορικές βαθμίδες) απαιτεί ομογενοποίηση και κανόνες κανονικοποίησης, ώστε να ανιχνεύονται αξιόπιστα μεταβολές όπως προαγωγές, μετακινήσεις ή αποχωρήσεις και να διασφαλίζεται η συνέπεια των δεδομένων σε βάθος χρόνου.

1.3 Επιστημομετρία

Η επιστημομετρία είναι ο ποσοτικός κλάδος που μελετά την παραγωγή, διάχυση και απήχηση της επιστημονικής γνώσης, αξιοποιώντας δείκτες όπως αναφορές, h-index και i10-index, καθώς και αναλύσεις δικτύων (συγγραφείς, ιδρύματα, θεματικά πεδία). Χρησιμοποιείται για χαρτογράφηση ερευνητικών τάσεων και υποστήριξη αξιολόγησης. Στο παρόν έργο οι δείκτες αντλούνται αυτούσιοι από το Google Scholar, ωστόσο η ερμηνεία τους πρέπει να λαμβάνει υπόψη διαφορές ανά επιστημονικό πεδίο, την κάλυψη/ποιότητα δεδομένων, ζητήματα αποσαφήνισης ονομάτων και τις αυτοαναφορές.

Στο πλαίσιο της παρούσας εργασίας εστιάζουμε στο Google Scholar ως πρωτεύουσα δημόσια πηγή επιστημομετρικών δεδομένων, λόγω της ευρείας κάλυψης και της δωρεάν πρόσβασης που προσφέρει σε πλήθος πεδίων, ιδίως στην Πληροφορική. Η συμβολή του έργου είναι ότι αυτοματοποιεί τον εντοπισμό μελών ΔΕΠ από ετερογενείς ιστοσελίδες τμημάτων/ιδρυμάτων και, στη συνέχεια, παρέχει δυνατότητα συσχέτισης για κάθε μέλος με το αντίστοιχο προφίλ του στο Google Scholar (GSID), ώστε να αντλούνται δείκτες όπως h-index, i10-index και αναφορές χωρίς χειροκίνητη καταχώριση. Μέσω κανονικοποίησης ονομάτων και ελεγχόμενης ροής έγκρισης, το σύστημα μειώνει θόρυβο/διπλοεγγραφές και επιτρέπει συνεχή, αυτόματη ενημέρωση των δεδομένων, υποστηρίζοντας ιδρυματικές αναλύσεις και παρακολούθηση μεταβολών (π.χ. προαγωγές) με ελάχιστο διοικητικό φόρτο.

1.4 Διαδικτυακοί τόποι επιστημομετρίας

Οι διαδικτυακοί τόποι επιστημομετρίας είναι πλατφόρμες που συγκεντρώνουν, οργανώνουν και μετροκοπούν επιστημονικά τεκμήρια, παρέχοντας δείκτες απήχησης και παραγωγικότητας. Ενδεικτικά, το Google Scholar προσφέρει ευρεία κάλυψη και απλή αναζήτηση, ενώ πόροι όπως OpenAlex και Crossref λειτουργούν ως ανοιχτές υποδομές μεταδεδομένων με έμφαση σε DOI/ταυτοποίηση. Επιμελημένες βάσεις όπως Scopus και Web of Science διαθέτουν αυστηρότερη επιμέλεια περιεχομένου αλλά συχνά υπόκεινται σε άδειες/συνδρομές. Οι διαφορές σε εύρος κάλυψης, επιμέλεια, καθυστερήσεις ενημέρωσης, μοντέλα πρόσβασης (API/σελιδοποίηση) και ταυτοποίηση οντοτήτων (ORCID, DOI, GSID) επηρεάζουν άμεσα τη συγκρισιμότητα και τη χρησιμότητα των δεικτών.

Από τεχνική πλευρά, η αξιοποίηση αυτών των τόπων κυμαίνεται από επίσημα APIs έως εξαγωγή από HTML, με ζητήματα όπως όρια ρυθμού, μεταβολές DOM Tree για περιπτώσεις εξαγωγής από HTML, πολιτικές χρήσης και ανάγκη αποφυγής διπλοεγγραφών. Καίρια είναι η αποσαφήνιση ταυτότητας συγγραφέων (name disambiguation), η ενοποίηση παραλλαγών ονομάτων/ιδρυμάτων και η ιχνηλασιμότητα ενημερώσεων (logs). Στο παρόν έργο, εστιάζουμε στο Google Scholar: αυτοματοποιούμε τον εντοπισμό μελών ΔΕΠ από ετερογενείς ιστοσελίδες τμημάτων και, στη συνέχεια, διευκολύνουμε τη συσχέτισή τους με προφίλ GS (GSID) για την άντληση δεικτών χωρίς χειροκίνητη καταχώριση, με έμφαση στη σταθερότητα διαδικασίας, στον έλεγχο αλλαγών και στη βιώσιμη περιοδική συντήρηση δεδομένων.

1.5 Κίνητρο

Πίνακας 1.1: Πίνακας με στοιχεία καθηγητών

	A-Z gsid	A-Z name	A-Z position	A-Z inst	123 hindex	123 publications	123 citations
1	_56WDOoAAAAJ	Anyia Sotiropoulou	Lab Lecturer	cs@ionio	6	25	81
2	_6k57BEAAAAJ	Christos Katsanos	Assistant Professor	csd@auth	18	71	1,264
3	_6KPd_0AAAAJ	Maria Louka	Lab Lecturer	di@uoa	4	10	47
4	_6MQFHsAAAAJ	Kostas Vassilakis	Professor	ece@hmu	11	55	447
5	_h6qzbcAAAAJ	Angelos Giakoumis	Assistant Professor	iee@ihu	5	18	192
6	_HksGRoAAAAJ	Panagiotis Kokkinos	Assistant Professor	ds@uop	19	115	1,136
7	_jPeBJ0AAAAJ	-Michael T Chrysomallis	Professor	ee@duth	15	84	1,513
8	_muGENEAAAAJ	Zafeiris Vasileios	Lab Lecturer	cs@aueb	3	10	99
9	_nvjnSoAAAAJ	D. Magos	Professor	ice@uniwa	13	58	461
10	_pWMKhgAAAAJ	Konstantinos Politopoulos	Associate Professor	ece@ntua	12	55	648
11	_Q_wCtYAAAAJ	Spyros Denazis	Professor	ece@upatras	22	155	2,643
12	_q7ATEYAAAAJ	PARIS MASTOROCOSTAS	Professor	ice@uniwa	14	73	1,208
13	_raNYW4AAAAJ	Kostas Goulianas	Associate Professor	iee@ihu	5	12	47
14	_vIFOPgAAAAJ	Sotirios Xydis	Assistant Professor	dit@hua	19	138	1,413
15	_0BLhicAAAAJ	Evgenia Adamopoulou	Lab Lecturer	ece@ntua	15	100	918

Μέχρι σήμερα η επικαιροποίηση των στοιχείων ακαδημαϊκού προσωπικού γίνεται χειροκίνητα από μέλη του τμήματος καθώς χρειάζεται να γίνει επίσκεψη στις ιστοσελίδες πολλών διαφορετικών τμημάτων/ιδρυμάτων, οπτική σύγκριση λιστών και καταγραφή μεταβολών (νέες προσλήψεις, προαγωγές, αποχωρήσεις). Για κάθε μέλος υπάρχουν πολλά πεδία προς έλεγχο και ενημέρωση (βασικά στοιχεία ταυτότητας και θέσης, αλλά και επιστημονομετρικά όπως h-index, αναφορές, δημοσιεύσεις κ.λπ.), με αποτέλεσμα ο χρόνος ανά εγγραφή να είναι υψηλός. Ο Πίνακας 1.1 αναδεικνύει πώς ο όγκος των πεδίων, σε συνδυασμό με το πλήθος μελών ΔΕΠ, πολλαπλασιάζει τον συνολικό χρόνο. Η διαδικασία, όντας επαναληπτική και κουραστική, είναι επιρρεπής σε λάθη, ασυνέπειες (π.χ. παραλλαγές ονομάτων) και αστάθεια στην περιοδικότητα, με συνέπεια τα δεδομένα να υστερούν σε επικαιρότητα.

Αυτοί οι πρακτικοί περιορισμοί δημιουργούν ένα σαφές κίνητρο για αυτοματοποίηση: ελάφρυνση του διοικητικού φόρτου, ταχύτερη και συχνότερη ενημέρωση, ομογενοποίηση ονομάτων/βαθμίδων και μείωση λαθών. Ένα σύστημα που αντλεί πληροφορία απευθείας από τις πηγές, εντοπίζει μεταβολές και υποβοηθά την επικύρωσή τους μπορεί να διασφαλίσει ότι η βάση δεδομένων αντικατοπτρίζει την πραγματικότητα με συνέπεια και ιχνηλασιμότητα, απελευθερώνοντας πολύτιμο χρόνο από τα μέλη του προσωπικού και βελτιώνοντας την αξιοπιστία των αναλύσεων και των αναφορών.

1.6 Συνεισφορά

Η παρούσα εργασία παρέχει ένα ενιαίο, ενημερωμένο και ομογενοποιημένο αποθετήριο δεδομένων ακαδημαϊκού προσωπικού, με τυποποιημένα πεδία (όνομα, βαθμίδα, τύπος προσωπικού, τμήμα/ίδρυμα, GSID) και ροή ανίχνευσης αλλαγών (νέες προσθήκες, προαγωγές, αποχωρήσεις) μέσω εκκρεμών αλλαγών προς έγκριση. Με αυτόν τον τρόπο μειώνεται ο φόρτος εργασίας και τα ανθρώπινα λάθη, αυξάνεται η συχνότητα/συνέπεια ενημερώσεων και διασφαλίζεται ότι η βάση αντικατοπτρίζει την πραγματικότητα με ιχνηλασιμότητα (logs) και διαβαθμισμένα δικαιώματα πρόσβασης.

Πρακτικά, τα αποτελέσματα εκτίθενται σε μηχανικώς αναγνώσιμη μορφή JSON με σταθερό σχήμα (π.χ. name, rank, staff_type, university, department, gsid), γεγονός που επιτρέπει άμεση τροφοδότηση βάσεων δεδομένων και εργαλείων ανάλυσης μέσω απλών διεπαφών. Η τυποποίηση των πεδίων διευκολύνει την ένταξη σε υφιστάμενες ροές, την προγραμματισμένη ενημέρωση και τη δημιουργία αναπαραγωγίμων pipelines για επικαιροποιημένα στοιχεία προσωπικού.

1.7 Οργάνωση της Εργασίας

Η εργασία αυτή παρουσιάζει ένα ολοκληρωμένο σύστημα για αυτόματη συλλογή, ομογενοποίηση και συνεχή επικαιροποίηση στοιχείων ακαδημαϊκού προσωπικού από ετερογενείς ιστοσελίδες τμημάτων/ιδρυμάτων. Η ροή περιλαμβάνει παραμετρική εξαγωγή δομημένων δεδομένων, ενιαία αναπαράσταση (name, rank, staff_type, university, department, gsid), ανίχνευση μεταβολών (νέες προσθήκες, προαγωγές, αποχωρήσεις) ως εκκρεμείς αλλαγές προς έγκριση, και προαιρετική συσχέτιση με προφίλ Google Scholar για άντληση βασικών δεικτών. Η web εφαρμογή προσφέρει αναζήτηση, φίλτρα, σελιδοποίηση, ρόλους/δικαιώματα και ιχνηλασιμότητα μέσω logs, ενώ τα αποτελέσματα εκτίθενται σε τυποποιημένο JSON ώστε να τροφοδοτούνται εύκολα βάσεις δεδομένων και εργαλεία ανάλυσης.

Κατόπιν της εισαγωγής, η εργασία προχωρά στην αποτύπωση των υπαρχόντων εφαρμογών που σχετίζονται με επιστημονικά δεδομένα, ώστε να τεκμηριωθούν οι απαιτήσεις και τα κενά που καλύπτει το προτεινόμενο σύστημα. Ακολουθεί η παρουσίαση της αρχιτεκτονικής και των τεχνολογικών επιλογών, της ροής συλλογής/ομογενοποίησης δεδομένων και της web εφαρμογής.

Συγκεκριμένα, στο Κεφάλαιο 2 παρουσιάζονται αναλυτικά δύο αντιπροσωπευτικές πλατφόρμες (2.1, 2.2), με έμφαση στις λειτουργίες τους και στο πώς συνδυάζονται με ενημερωμένα στοιχεία προσωπικού. Το Κεφάλαιο 3 καλύπτει την αρχιτεκτονική του συστήματος και τις τεχνολογίες. Το Κεφάλαιο 4 περιγράφει το scraping microservice (παραμετροποίηση, συμμόρφωση robots, caching, parsing pipeline). Το Κεφάλαιο 5 παρουσιάζει τη web εφαρμογή και τέλος, το Κεφάλαιο 6 περιλαμβάνει μεθοδολογία αξιολόγησης, αποτελέσματα/παραδείγματα, συμπεράσματα και δυνατότητες βελτίωσης.

Κεφάλαιο 2ο: Εφαρμογές

2.1 OMEA Citations

Google Scholar citations

See departments' stats

☒Professor

☒Associate Professor

☒Assistant Professor

☒Lecturer

☐Lab Lecturer

☐Former member

select all

Click column head to sort

Members:30

P:2142

C:39721

Avg H:14.4

P5:696

C5:19287

Avg C per M per Y:57.3

Avg H5:10

Avg P per M:71.4

Avg C per M:1324

Avg age:23.8

Avg P per M per Y:3.2

☐ceid@upatras

☐cs@aub

☐cs@hu

☐cs@ionio

☐cs@unipi

☐cs@uoc

☐cs@uoi

☐cs@uowm

☐cs@uth

☐csd@auth

☐dai@uom

☐di@uoa

☐db@uth

☐dit@hua

☐dit@uoi

☐dit@uop

☐ds@unipi

☐ds@uop

☐ds@uth

☐e@hmu

☐ece@ntua

☐ece@tuc

☐ece@uop

☐ece@uowm

☐ece@upatras

☐ece@uth

☐ee@auth

☐ee@uth

☐ee@hmu

☐eee@uniwa

☐ice@uniwa

☐ic@d@aegean

☐ict@hu

☒lee@hu

select all

clear all

Name	Position	Dept	P	C	H	P5	C5	H5	2025	2024	2023	2022	2021	2020	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990	none	avgP	σ		
Periklis Oudonizos	Professor	lee@hu	431	8186	41	95	3560	25	9355	10520	17575	28647	15715	18744	17748	14755	18643	18449	35240	24248	25152	16174	8169	8195	7150	2163	3176	8159	1059	1042	50	30	-	10	-	-	-	-	-	-	-	-	-	9089	17.2	3			
Kostas Diamantaris	Professor	lee@hu	224	7307	30	58	3109	21	7567	11863	10599	14493	8495	8343	9273	5253	4225	7121	6125	4132	3172	3147	6157	8183	3191	5179	7228	8202	3204	3210	7195	3157	6150	797	7100	795	566	852	831	536	429	220	318	59	12169	6.4	2		
Boas Vasilatos	Professor	lee@hu	72	2787	23	11	615	12	875	3136	1133	1115	308	468	853	884	8190	895	8195	2150	8111	5140	3170	3186	3163	2174	3187	7183	5115	553	944	815	410	40	-	-	-	-	-	-	-	-	-	-	205	2.8	1		
Isidoros Korkkakis	Professor	lee@hu	30	2119	18	0	490	14	837	864	875	880	8115	8119	897	8113	8132	8138	8151	3140	3158	3139	897	189	283	271	153	358	427	119	216	214	813	88	85	812	819	20	-	-	-	-	-	-	-	116	1	7	
Charalambos Bratis	Assistant Professor	lee@hu	136	2190	22	27	1063	14	3132	8206	4183	4199	7181	1162	6129	10136	7116	18118	2110	4104	899	1475	1256	429	824	316	38	10	30	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1329	6.5	1	
Maria Papadopoulou	Assistant Professor	lee@hu	88	1699	12	60	1548	11	7275	8450	19404	18259	9286	740	18	82	84	83	-	22	82	18	13	21	26	47	14	127	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	816	3.8	7	
Constantinos Marios Angelopoulos	Assistant Professor	lee@hu	91	1476	22	21	1039	18	8227	1229	4176	3156	5134	8181	1084	219	374	458	964	329	619	39	85	20	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1921	3.8	6	
Alexios Sidiroglou	Associate Professor	lee@hu	42	1413	13	11	364	7	193	481	850	133	353	246	871	285	370	288	296	294	190	197	8113	198	168	387	429	518	34	-	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	827	1.9	6
Epifanio Korkkakis	Associate Professor	lee@hu	54	1320	22	22	1101	18	1193	2234	2218	8287	7149	494	848	743	117	127	528	320	29	37	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	234	4.2	1
Evdiki Karamerousi	Professor	lee@hu	51	1158	13	16	1054	11	1334	2344	3224	673	138	324	819	22	23	18	26	82	17	517	31	22	82	93	22	83	-	412	81	410	51	14	13	82	10	-	-	-	-	-	-	-	-	-	208	1.8	4
George Korkkakis	Assistant Professor	lee@hu	95	1001	18	31	677	13	8183	3106	3108	3119	5123	8115	17132	14108	1536	819	48	13	12	20	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	321	7	7

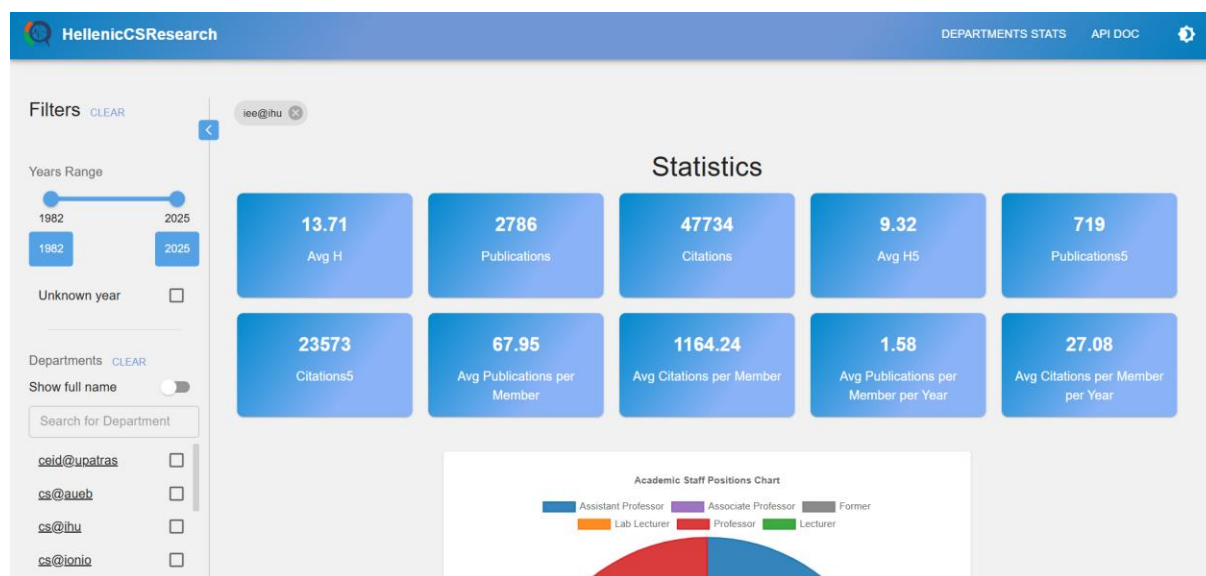
Σχήμα 2.1 : Αρχική σελίδα της εφαρμογής OMEA Citations

Το OMEA Citations είναι μια απλή διαδικτυακή εφαρμογή που αναπτύχθηκε από μέλη του τμήματος και παρουσιάζει επιστημονικά στοιχεία (π.χ. h-index, αναφορές, δημοσιεύσεις) από το Google Scholar για μέλη ακαδημαϊκού προσωπικού τμημάτων Πληροφορικής. Η πλατφόρμα διατηρεί για κάθε μέλος ένα βασικό προφίλ (ταυτότητα/τμήμα/Ιδρυμα) και ένα GSID, και ενημερώνει αυτόματα τις μετρικές από το Scholar, επιτρέποντας γρήγορη επισκόπηση της ερευνητικής απήχησης ανά τμήμα ή ίδρυμα. (Σχήμα 2.1)

Κρίσιμος περιορισμός είναι ότι ο κατάλογος μελών ΔΕΠ/ΕΔΙΠ εισάγεται και επικαιροποιείται χειροκίνητα: οι νέες προσλήψεις, προαγωγές ή αποχωρήσεις συχνά καθυστερούν να αποτυπωθούν, οδηγώντας σε γρήγορη απαξίωση του μητρώου, παρότι οι μετρικές Scholar ανανεώνονται αυτόματα. Η χειροκίνητη ροή αυξάνει τον φόρτο εργασίας και τον κίνδυνο ασυνεπειών (π.χ. παραλλαγές ονομάτων), ειδικά σε πολλαπλά τμήματα Πληροφορικής. Μια τυποποιημένη και αυτόματη τροφοδότηση με ενημερωμένο μητρώο θα μπορούσε να ενσωματωθεί απευθείας στην εφαρμογή, μειώνοντας δραστικά τις χειροκίνητες παρεμβάσεις και διασφαλίζοντας ότι οι λίστες προσωπικού παραμένουν σύγχρονες.

2.2 HCSR

Το HCSR (Hellenic Computer Science Research) αποτελεί εξελιγμένη εκδοχή πλατφόρμας παρουσίασης επιστημονικών δεδομένων, με σύγχρονη και καθαρή διεπαφή, διαδραστικά γραφήματα και πιο πλούσιες αναλύσεις για τμήματα Πληροφορικής. Παρέχει δημόσιο API για προγραμματιστική πρόσβαση σε στοιχεία (ερευνητές, δείκτες, συγκεντρωτικά), διευκολύνοντας ενσωματώσεις σε dashboards, αναφορές και εξωτερικά εργαλεία. Η έμφαση σε ευχρηστία και οπτικοποίηση καθιστά ευκολότερη την εξερεύνηση τάσεων, συγκρίσεων και κατανομών. (Σχήμα 2.2)



Σχήμα 2.2 :Αρχική σελίδα της εφαρμογής HCSR

Παρά τα παραπάνω, το βασικό ζήτημα παραμένει η έγκαιρη επικαιροποίηση του μητρώου μελών ΔΕΠ/ΕΔΙΠ: όταν οι λίστες τροφοδοτούνται/διορθώνονται χειροκίνητα, οι αλλαγές (νέες προσθήκες,

προαγωγές, αποχωρήσεις) μπορεί να καθυστερούν να αποτυπωθούν, με αποτέλεσμα απόκλιση μεταξύ πραγματικότητας και εμφανιζόμενων δεδομένων, ακόμη και εάν οι μετρικές του Scholar ενημερώνονται αυτόματα. Μια τυποποιημένη, μηχανικώς αναγνώσιμη ροή με σταθερά πεδία θα επέτρεπε αξιόπιστη τροφοδότηση του HCSR και συνεχή ευθυγράμμιση του δημόσιου API με το τρέχον μητρώο.

Κεφάλαιο 3ο: Προσέγγιση

3.1 Η Προσέγγιση της Γενικής Σχεδίασης

Μόλις αναγνωρίστηκε η ανάγκη για τον αυτοματισμό της διαδικασίας ενημέρωσης στοιχείων, η πρόκληση ήταν η σχεδίασή ενός συστήματος που θα μπορούσε να αυτοματοποιήσει την διαδικασία, χωρίς την ύπαρξη ειδικού κώδικα για την κάθε διαφορετική ιστοσελίδα τμήματος. Παρατηρώντας τις διάφορες σελίδες αυτές, φάνηκε πως παρά τις διαφορές τους στη δομή και την εμφάνιση, υπήρχαν συγκεκριμένα μοτίβα, τα οποία οδήγησαν στην ανάπτυξη του αυτοματοποιημένου συστήματος για ενημέρωση στοιχείων καθηγητών.

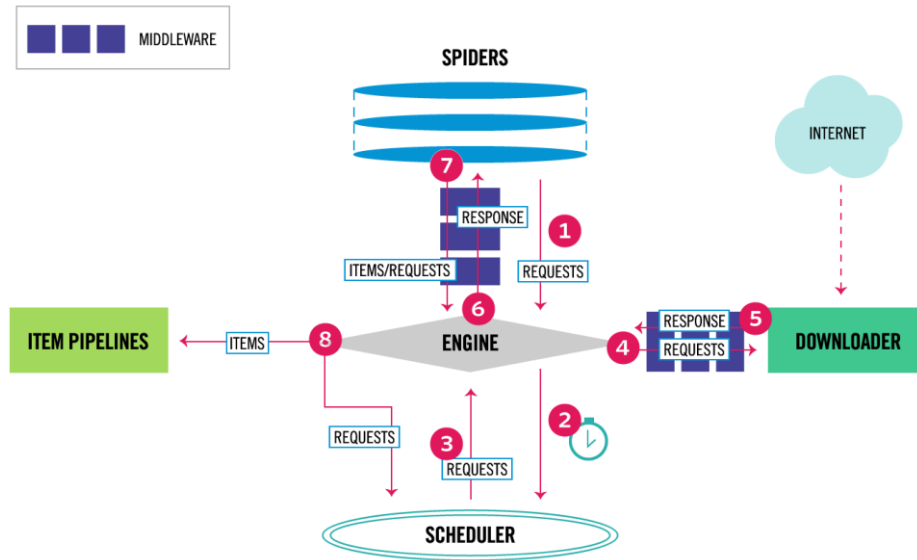
Η κεντρική ιδέα ήταν να δημιουργηθεί ένα ευέλικτο scraping microservice που ρυθμίζεται με ελάχιστες παραμέτρους, αντί να απαιτείται ξεχωριστός κώδικας ανά ιστοσελίδα. Για κάθε τμήμα χρησιμοποιείται ένα μικρό αρχείο ρυθμίσεων (JSON) με τα απολύτως απαραίτητα: URLs, βασικούς CSS selectors για ονόματα και βαθμίδες, λέξεις-κλειδιά και ενδείξεις για τη δομή της σελίδας. Με βάση αυτά επιλέγεται η κατάλληλη μέθοδος ανάγνωσης (πίνακες, λίστες, containers ή διαδοχικό κείμενο), τα οποία θα αναλυθούν περισσότερο παρακάτω, και ως αποτέλεσμα παράγονται ομοιόμορφα δεδομένα απ' όλες τις σελίδες σε ένα ενιαίο αρχείο. Έτσι, η προσθήκη νέου τμήματος μεταφράζεται απλώς σε ένα νέο αρχείο ρυθμίσεων, χωρίς αλλαγές στον κορμό του κώδικα. Παράλληλα, διατηρείται cache των σελίδων που έχουν ήδη ανακτηθεί και τηρούνται οι κανόνες crawling κάθε ιστότοπου: πριν από κάθε αίτημα ανακτάται το αρχείο robots.txt (Robots Exclusion Protocol)[1] στο /robots.txt, ελέγχονται τα επιτρεπόμενα/απαγορευμένα μονοπάτια για τον δηλωμένο user-agent και εφαρμόζεται τυχόν χρονική καθυστέρηση (crawl-delay). Με αυτόν τον τρόπο μειώνεται το φορτίο στους εξυπηρετητές, αποφεύγονται αποκλεισμοί και τηρούνται οι πολιτικές λειτουργίας των ιστοτόπων.

Χωρίσαμε το σύστημα σε τρία καθαρά κομμάτια με ξεκάθαρο ρόλο ώστε να συνεργάζονται, αλλά να τρέχουν ανεξάρτητα. Το scraping microservice «φέρνει» τα δεδομένα: επισκέπτεται τις σελίδες, αναλύει το HTML και παράγει τυποποιημένα αποτελέσματα. Η web εφαρμογή τα «δείχνει» και τα οργανώνει: προβολή, αναζήτηση, επιβεβαίωση αλλαγών από το scraping microservice, σύνδεση με στοιχεία από το Google Scholar και βασική διαχείριση χρηστών. Η βάση δεδομένων διατηρεί: εγγραφές μελών, ιστορικό ενημερώσεων και συσχετίσεις. Αυτός ο πρακτικός διαχωρισμός μας επιτρέπει να αναβαθμίζουμε ή να κλιμακώνουμε κάθε μέρος ξεχωριστά (π.χ. να τρέχει ο scraper συχνότερα χωρίς να επηρεάζεται το web), να δοκιμάζουμε αλλαγές με ασφάλεια και να εντοπίζουμε άμεσα την πηγή τυχόν προβλημάτων.

3.2 Web Scrapers

Οι web scrapers είναι λογισμικά που αυτοματοποιούν την εξαγωγή δομημένων δεδομένων από ιστοσελίδες. Σε αντίθεση με τους web crawlers, που εστιάζουν κυρίως στην πλοήγηση και ευρετηρίαση μεγάλων όγκων σελίδων, οι scrapers επικεντρώνονται στην ανάγνωση του περιεχομένου μιας σελίδας

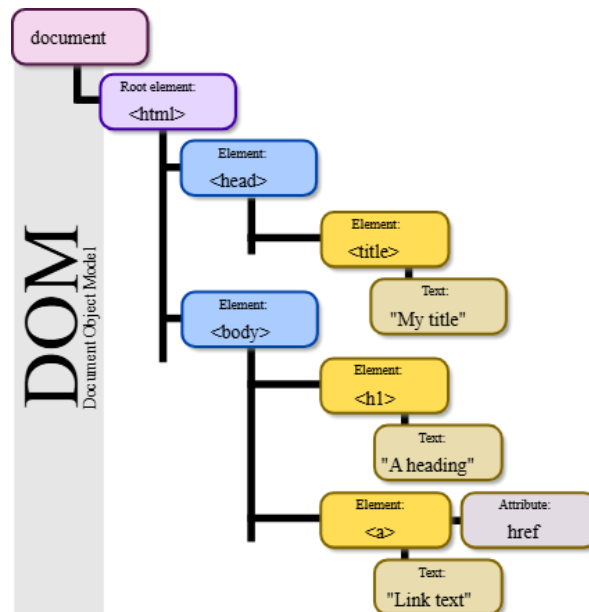
και στη μετατροπή του σε αξιοποιήσιμη, τυποποιημένη μορφή (π.χ. πίνακες ονομάτων/βαθμίδων)[22][23].



Σχήμα 3: Διάγραμμα αρχιτεκτονικής Scrapy

Οι web scrapers είναι προγράμματα που αυτοματοποιούν την εξαγωγή δομημένων δεδομένων από ιστοσελίδες. Διαφέρουν από τους web crawlers, οι οποίοι εστιάζουν στην ευρετηρίαση και πλοήγηση μεγάλων συνόλων σελίδων. Οι scrapers επικεντρώνονται στο ίδιο το περιεχόμενο μιας σελίδας και το μετατρέπουν σε αξιοποιήσιμη μορφή (π.χ. πίνακες με ονόματα και βαθμίδες)[22]. Στην πράξη χρησιμοποιούνται από ερευνητές, οργανισμούς και επιχειρήσεις για χρήσεις όπως σύγκριση τιμών, παρακολούθηση αλλαγών περιεχομένου και συγκέντρωση δεικτών ή μεταδεδομένων που διατίθενται δημόσια[23].

Η συνήθης ροή εργασίας περιλαμβάνει προγραμματισμό των διευθύνσεων προς άντληση, ανάκτηση περιεχομένου μέσω HTTP με σεβασμό στις πολιτικές πρόσβασης των ιστοτόπων (robots.txt)[1], ανάλυση της δομής HTML/DOM και εντοπισμό των σχετικών στοιχείων με CSS selectors ή XPath (Σχήμα 3.2)[29], καθαρισμό/κανονικοποίηση κειμένου και αποθήκευση σε ενιαίο σχήμα. Frameworks όπως το Scrapy παρέχουν έτοιμη αρχιτεκτονική (scheduler, downloader, parsers, pipelines), κάνοντας τη διαδικασία πιο αξιόπιστη και επεκτάσιμη (Σχήμα 3.1)[30].



Σχήμα 3 :DOM Model

Όταν το περιεχόμενο παράγεται δυναμικά από JavaScript, χρησιμοποιούνται εργαλεία αυτοματοποίησης «χωρίς κεφάλι» (headless browsers) ώστε η σελίδα να αποδοθεί πριν την ανάλυση, με συνήθεις επιλογές το Selenium WebDriver και το Playwright[38][39]. Παράλληλα, τηρούνται αρχές ευπρέπειας και συμμόρφωσης: σεβασμός στο Robots Exclusion Protocol (robots.txt)[1], λογικός ρυθμός αιτημάτων και προσοχή στους όρους χρήσης. Ζητήματα νομιμότητας γύρω από τη συλλογή δημόσιου περιεχομένου έχουν απασχολήσει τα δικαστήρια, με χαρακτηριστική την υπόθεση hiQ Labs v. LinkedIn (9th Cir., 2022)[31]. Συνολικά, οι scrapers λειτουργούν ως γέφυρα που μετατρέπει ετερογενείς ιστοσελίδες σε συνεκτικά, ενημερώσιμα σύνολα δεδομένων έτοιμα για ανάλυση και ενσωμάτωση σε εφαρμογές[22][23].

3.3 Web Crawlers vs Scrapers

Οι web crawlers (spiders) είναι προγράμματα που «περνούν» συστηματικά σελίδες μια-προς-μια, ακολουθώντας συνδέσμους από αρχικά URLs, με στόχο την ανακάλυψη μεγάλων τμημάτων του ιστού. Διαχειρίζονται ουρά διευθύνσεων (URL frontier), εφαρμόζουν πολιτικές επανάληψης/επανεπίσκεψης, κανονικοποίηση διευθύνσεων και αποφυγή διπλοτήτων, αξιοποιούν sitemaps όπου υπάρχουν, και τηρούν τις πολιτικές πρόσβασης (Robots Exclusion Protocol/robots.txt, ρυθμός αιτημάτων)[22]. Το κύριο παραδοτέο τους είναι ευρετήρια, σύνολα URLs ή/και αντίγραφα σελίδων με μεταδεδομένα για αναζήτηση και ανάλυση μεγάλης κλίμακας.

Σε αντιπαράβολή, οι web scrapers στοχεύουν «σε βάθος» σε συγκεκριμένες γνωστές σελίδες, εξάγοντας πεδία ενδιαφέροντος και αποθηκεύοντάς τα σε δομημένη μορφή (π.χ. CSV/JSON ή βάση δεδομένων), έτοιμη για άμεση αξιοποίηση σε εφαρμογές ή αναφορές[23]. Οι crawlers υπηρετούν την ευρεία ανακάλυψη και ευρετηρίαση, ενώ οι scrapers την στοχευμένη εξαγωγή περιεχομένου. Στο παρόν έργο, το ζητούμενο είναι κυρίως η δεύτερη προσέγγιση (scraping) πάνω σε συγκεκριμένες σελίδες τμημάτων.

Οι web crawlers και οι web scrapers εξυπηρετούν συμπληρωματικούς αλλά διακριτούς στόχους. Οι crawlers αποσκοπούν στην ευρεία ανακάλυψη και επανεπίσκεψη σελίδων, διαμορφώνοντας έναν «χάρτη» του ιστού ή ενός επιλεγμένου υποσυνόλου του. Οι scrapers επικεντρώνονται στην εξαγωγή συγκεκριμένων πεδίων από γνωστές σελίδες, σε δομημένη μορφή. Στο παρόν έργο το βάρος πέφτει στη δεύτερη λειτουργία (εξαγωγή), ωστόσο η κατανόηση και η στοχευμένη αξιοποίηση τεχνικών crawling είναι κρίσιμες για να εντοπίζονται με σταθερό τρόπο οι σελίδες ενδιαφέροντος (π.χ. σελίδες λίστας προσωπικού και detail pages με αναλυτικό προφίλ).

3.4 Web Crawlers: Ορισμοί και πλαίσιο χρήσης

Οι web crawlers αποσκοπούν στην ανακάλυψη και επανεπίσκεψη ιστοσελίδων, δημιουργώντας και διατηρώντας έναν «χάρτη» πηγών με οργανωμένο ιστορικό μεταδεδομένων (πότε επισκέφθηκαν μια σελίδα, εάν άλλαξε, ποιο είναι το μέγεθος/τύπος περιεχομένου κ.ο.κ.). Αντίθετα, οι web scrapers επικεντρώνονται στην εξαγωγή συγκεκριμένων πεδίων από γνωστές σελίδες και στη μετατροπή τους σε τυποποιημένα δεδομένα. Στο παρόντα εργασία, παρότι το κύριο ζητούμενο είναι η ακριβής εξαγωγή, η κατανόηση και η στοχευμένη αξιοποίηση τεχνικών crawling είναι καθοριστικές για τον σταθερό εντοπισμό των κατάλληλων σελίδων (λίστες προσωπικού, σελίδες προφίλ) που τροφοδοτούν με ποιότητα το στάδιο της εξαγωγής.

Η διάκριση αυτή δεν είναι θεωρητική αλλά λειτουργική. Ο crawler, ακόμη και όταν εφαρμόζεται «εστιασμένα» σε ένα domain, αναλαμβάνει τον προγραμματισμό επισκέψεων, την ευπρέπεια πρόσβασης και την τήρηση πολιτικών, ώστε να διασφαλίζεται πληρότητα και επικαιρότητα πηγών χωρίς υπερφόρτωση. Ο scraper αξιοποιεί τις πηγές που ανακάλυψε/επικύρωσε ο crawler για να πράξει καθαρά και κανονικοποιημένα δεδομένα. Ο σωστός συντονισμός των δύο ρόλων είναι αυτός που, στην πράξη, μειώνει σφάλματα, κόστοι και καθυστερήσεις.

3.5 Αρχιτεκτονική crawler και ευπρέπεια πρόσβασης

Στον πυρήνα ενός crawler βρίσκεται το «frontier» διευθύνσεων (URL frontier), μια δομή που οργανώνει τι θα επισκεφθεί στη συνέχεια με βάση προτεραιότητες, πολιτικές ανά host και αποφυγή διπλοτήτων. Η ευπρέπεια πρόσβασης επιβάλλει σεβασμό στο Robots Exclusion Protocol (robots.txt), σε συνδυασμό με ρυθμό αιτημάτων ανά προέλευση και προσαρμογή σε ενδείξεις υπερφόρτωσης. Η διαχείριση ανά domain (per-host queues) επιτρέπει την τήρηση ορίων, ενώ η αποθήκευση μεταδεδομένων (κωδικοί HTTP, χρόνοι απόκρισης, τελευταίο hash περιεχομένου) βελτιώνει τη λειτουργική ορατότητα.

Η ευπρέπεια δεν εξαντλείται στο robots.txt. Οι «robots meta» ετικέτες του HTML και οι αντίστοιχοι HTTP headers (x-robots-tag) δηλώνουν πρόσθετες επιθυμίες των διαχειριστών (π.χ. noindex/nofollow), ενώ οι απαντήσεις με Status Code 429/503 και τα μοτίβα rate limiting καθοδηγούν την προσαρμογή. Η πρακτική υλοποίηση απαιτεί backoff στρατηγικές, επαναλήψεις με όρια και έξυπνη προτεραιοποίηση, ώστε ο crawler να είναι «καλός πολίτης» του ιστού και να διατηρεί πρόσβαση μακροπρόθεσμα.

3.6 Κανονικοποίηση διευθύνσεων και αποφυγή διπλοεγγραφών

Πολλές σελίδες είναι προσπελάσιμες από πολλαπλά URLs λόγω παραμέτρων ή εναλλακτικών διαδρομών. Η κανονικοποίηση αφαιρεί ιχνηλατικές παραμέτρους, ενοποιεί σχήματα και αξιοποιεί τα canonical tags όπου υπάρχουν, για να προληφθεί ο τεχνητός πολλαπλασιασμός του ίδιου περιεχομένου. Έτσι, η «κάλυψη» αποτυπώνει πραγματική ποικιλία σελίδων και όχι παραλλαγές της ίδιας. [36]

Συμπληρωματικά, η ανίχνευση διπλοτύπων μέσω αποτυπωμάτων περιεχομένου (hashing/fingerprints) προστατεύει τον crawler από περιττές επανεπισκέψεις και εξασφαλίζει ορθότητα στις μετρήσεις. Για έργα όπως το παρόν, όπου ενδιαφέρει η ορθή αναγνώριση λίστες/προφίλ προσωπικού, η αποφυγή διπλοεγγραφών είναι προϋπόθεση για αξιόπιστη downstream επεξεργασία και σύγκριση στο χρόνο.

3.7 Εστιασμένο crawling σε ιστοσελίδες

Σε αντίθεση με την αδιάκριτη περιήγηση, το εστιασμένο crawling περιορίζει το πεδίο σε συγκεκριμένα μονοπάτια και πρότυπα (π.χ. /faculty/, /people/, /staff/), εφαρμόζοντας allow/deny λίστες, όρια βάθους και φίλτρα τύπων περιεχομένου. Με αυτόν τον τρόπο, εντοπίζονται οι κρίσιμες σελίδες χωρίς έκθεση σε άσχετες ενότητες ή μεγάλα αρχεία, με άμεσο όφελος στην ταχύτητα και τη σταθερότητα. [37]

Η πρακτική αξία είναι διπλή: αφενός ανακαλύπτονται σελίδες λίστας με συνοπτική πληροφορία, αφετέρου εντοπίζονται πρόσθετες πληροφορίες (π.χ βαθμίδα) με καθαρότερη διατύπωση και μερικές φορές χρονικά μεταδεδομένα. Η ύπαρξη αυτών των σελίδων προφίλ λειτουργεί ως «δεύτερη πηγή αλήθειας», βελτιώνοντας την ακρίβεια εξαγωγής των δεδομένων.

3.8 Δυναμικό περιεχόμενο και headless browsers

Πολλές σύγχρονες ιστοσελίδες παράγουν τελικό περιεχόμενο μέσω JavaScript (AJAX, client-side rendering). Σε τέτοιες περιπτώσεις, ένας «headless» browser αποδίδει τη σελίδα όπως ένας πραγματικός φυλλομετρητής, επιτρέποντας την εκτέλεση scripts και τη σταθεροποίηση του DOM πριν από την ανάλυση. Το Selenium WebDriver προσφέρει διεπαφή αυτοματοποίησης που ελέγχει προγράμματα πλοήγησης (Chrome/Firefox κ.ά.), ενώ το Playwright εστιάζει στη σταθερότητα selectors και την υποστήριξη πολλαπλών μηχανών rendering με έμφαση στην ταχύτητα και αξιοπιστία. [38][39]

Η επιλογή τέτοιων εργαλείων γίνεται με έρευνα και μελέτη, λόγω αυξημένου κόστους και ευαισθησίας σε πολιτικές ιστότοπων. Όπου είναι δυνατό, προτιμάται στατικό HTML με ελεγχόμενους selectors και caching. Όπου είναι αναγκαίο, τα headless εργαλεία περιορίζονται σε λίγες, στοχευμένες σελίδες, με ρητές αναμονές για συγκεκριμένους selectors και σαφή διαχείριση σφαλμάτων, ώστε να ελαχιστοποιείται η επιβάρυνση.

3.9 Σύνδεση με την κανονικοποίηση βαθμίδων ΔΕΠ στο ελληνικό σύστημα

Στο ακαδημαϊκό πλαίσιο, η ιεραρχία μελών ΔΕΠ ακολουθεί διαδρομή από Επίκουρο Καθηγητή (assistant professor) σε Αναπληρωτή Καθηγητή (associate professor) και καταλήγει σε Καθηγητή (full professor). Ο «Ομότιμος Καθηγητής» (emeritus) απονέμεται τιμητικά μετά τη συνταξιοδότηση και συχνά συνυπάρχει σε λίστες, χωρίς να αποτελεί ενεργή βαθμίδα εξέλιξης. Ιστορικές/μεταβατικές κατηγορίες όπως ο «Λέκτορας» μπορεί να απαντώνται σε παλαιότερες ή μη επικαιροποιημένες σελίδες, ενώ κατηγορίες ειδικού/τεχνικού προσωπικού (ΕΔΙΠ/ΕΕΠ/ΕΤΕΠ), «Adjunct/Visiting», ή συμβασιούχοι (π.δ. 407/80) συχνά εμφανίζονται στον ίδιο χώρο με τις βασικές βαθμίδες, δημιουργώντας αμφισημία.

Η κανονικοποίηση που απαιτεί το σύστημα αντιμετωπίζει διγλωσσία (Καθηγητής/Professor, Αναπληρωτής/Associate, Επίκουρος/Assistant), συντομογραφίες («Αν. Καθηγητής», «Επικ. Καθηγητής», «Assoc. Prof.»), ποικίλες πτώσεις, αλλά και διαφοροποιήσεις δομής (επικεφαλίδες, λεζάντες, breadcrumbs). Ο εστιασμένος crawler αυξάνει την πιθανότητα εντοπισμού σελίδων προφίλ όπου η βαθμίδα διατυπώνεται καθαρότερα, ενώ ο μηχανισμός χαρτογράφησης σε ελεγχόμενο λεξιλόγιο

και η πρόβλεψη «Unknown» με ανθρώπινη επιβεβαίωση λειτουργούν ως πρακτικές δικλίδες ποιότητας. Με αυτόν τον τρόπο εξασφαλίζεται ότι η ενιαία βάση αποτυπώνει συνεκτικά την πραγματική ιεραρχία εξέλιξης, επιτρέποντας αξιόπιστη σύγκριση και παρακολούθηση στο χρόνο.

Κεφάλαιο 4ο: Αρχιτεκτονική Συστήματος και Επιλογή Τεχνολογιών

4.1 Η Αρχιτεκτονική του Συστήματος

Η αρχιτεκτονική του συστήματος στηρίζεται στις αρχές των μικροϋπηρεσιών[2], ώστε κάθε υποσύστημα να μπορεί να αναπτύσσεται, να αναβαθμίζεται και να λειτουργεί αυτόνομα. Με αυτόν τον σχεδιασμό επιτυγχάνονται ευελιξία, επεκτασιμότητα και ευκολότερη συντήρηση, ενώ διευκολύνεται η προσαρμογή σε νέες απαιτήσεις χωρίς εκτεταμένες αλλαγές στη συνολική υποδομή. Το σύστημα χωρίστηκε σε τρία διακριτά μέρη με σαφείς ρόλους:

- Το scraping microservice, που ανακτά τα δεδομένα από τις ιστοσελίδες των τμημάτων και τα μετατρέπει σε τυποποιημένη μορφή.
- Τη web εφαρμογή, που αναλαμβάνει την παρουσίαση, την οργάνωση και τις βασικές ροές διαχείρισης (π.χ. αναζήτηση, έλεγχος/επιβεβαίωση αλλαγών, σύνδεση με στοιχεία Google Scholar).
- Τη βάση δεδομένων, που αποθηκεύει επίσημα τις εγγραφές, το ιστορικό ενημερώσεων και τις μεταξύ τους συσχετίσεις.

Η διασύνδεση των μερών υλοποιείται μέσω σαφώς ορισμένων APIs[3] και RESTful κλήσεων[4], με ανταλλαγή δεδομένων σε τυποποιημένη μορφή (JSON). Η επιλογή αυτή επιτρέπει ανεξάρτητη ανάπτυξη και αναβάθμιση ανά υποσύστημα, ευκολότερο εντοπισμό προβλημάτων, καθώς και μείωση των χρόνων εκτέλεσης (π.χ. ο scraper μπορεί να εκτελείται περιοδικά ή ασύγχρονα, χωρίς να επηρεάζεται η web διεπαφή).

Στο επίπεδο αποθήκευσης, εφαρμόστηκαν αρχές κανονικοποίησης[5] για τη μείωση πλεονασμού και την ενίσχυση της ακεραιότητας των δεδομένων. Συμπληρωματικά, χρησιμοποιείται μηχανισμός προσωρινής αποθήκευσης (caching)[6] για τη μείωση επαναλαμβανόμενων αιτημάτων και τη βελτίωση της απόδοσης, ιδίως κατά την ανάκτηση δεδομένων από εξωτερικές πηγές. Η χρήση HTTP ως κοινού πρωτοκόλλου επιτρέπει στα υποσυστήματα να συνεργάζονται ανεξάρτητα από γλώσσες και πλατφόρμες, απλοποιεί το deployment σε διαφορετικά περιβάλλοντα, δίνει δυνατότητα κλιμάκωσης ανά υπηρεσία και καθιστά εύκολες τις δοκιμές ανά microservice, χάρη σε καλά ορισμένα APIs.

4.2 Οι Τεχνολογίες που Επιλέχθηκαν

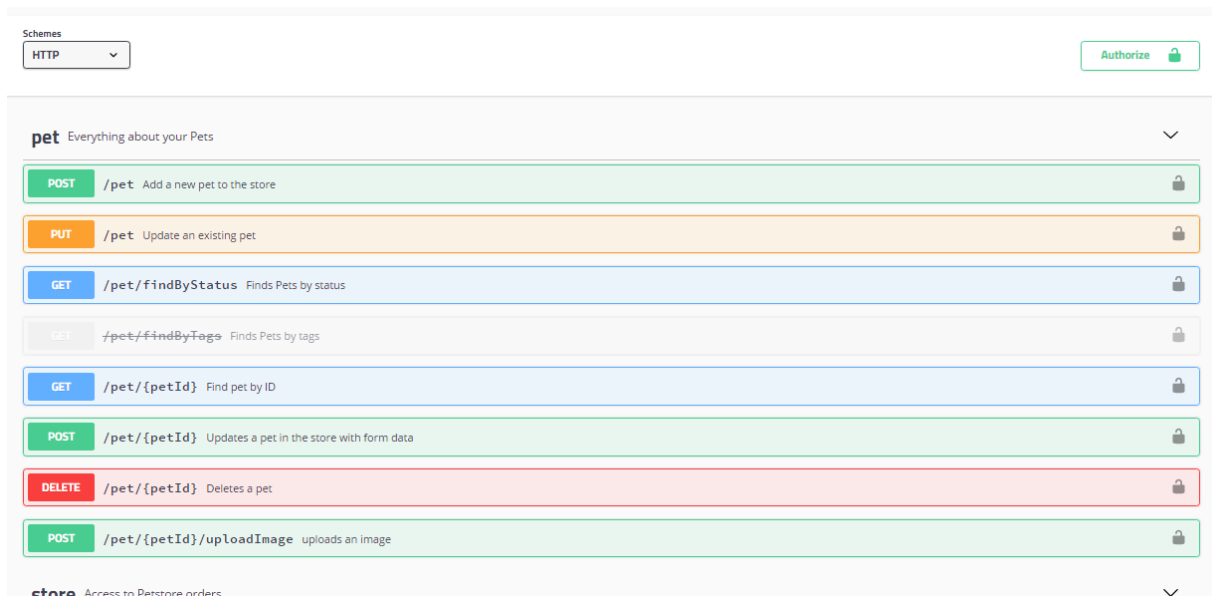
Για τις ανάγκες του έργου απαιτήθηκε πλατφόρμα που να επιτρέπει ανάλυση ιστοσελίδων και εξαγωγή/μετασχηματισμό δεδομένων με σχετική ευκολία: βιβλιοθήκες για HTTP και HTML parsing, ισχυρός χειρισμός κειμένου/Unicode, εγγενής υποστήριξη JSON, καθώς και δυνατότητες ασύγχρονης εκτέλεσης και γρήγορης πρωτοτυποποίησης. Προτιμήθηκε προσέγγιση υψηλού επιπέδου αντί γλωσσών χαμηλού επιπέδου, ώστε να μην απαιτείτε πολύ κώδικας και να επιταχυνθεί ο κύκλος ανάπτυξης. Σε αυτό το πλαίσιο, η Python ανταποκρίνεται πλήρως στα παραπάνω κριτήρια χάρη στην απλή, εκφραστική σύνταξη της και το πλούσιο οικοσύστημα εργαλείων για δικτυακές κλήσεις, επεξεργασία κειμένου και αυτοματοποίηση [7].

Συμπληρωματικά, αξιοποιούνται εξειδικευμένα εργαλεία ανά επίπεδο: ως HTML parser χρησιμοποιείται το BeautifulSoup για στοχευμένη πλοήγηση στο DOM και ανάκτηση στοιχείων με selectors, με ανοχή σε «ατελές» HTML και υποστήριξη διαφορετικών parsers [9][13]. Το επίπεδο web υπηρεσιών βασίζεται στο FastAPI, το οποίο αξιοποιεί τα type hints της Python για αυτόματο OpenAPI schema, ενσωματωμένη επικύρωση δεδομένων και ασύγχρονη εξυπηρέτηση αιτημάτων [8]. Η αποθήκευση δεδομένων υλοποιείται με MariaDB, μια σχεσιακή βάση δεδομένων όπου μας παρέχει ακεραιότητα δεδομένων, καθώς και υποστήριξη ευρετηρίων και βελτιστοποίησης ερωτημάτων [20][21]. Σε λειτουργικούς όρους: ο HTML parser παράγει κανονικοποιημένα σχήματα, τα οποία εκτίθενται μέσω σταθερών REST APIs και επιμένουν σε σχεσιακή αποθήκη, ο σαφής διαχωρισμός ρόλων και οι καθαρές διεπαφές προγραμματισμού επιτρέπουν ανεξάρτητη εξέλιξη υποσυστημάτων και ομαλή ένταξη νέων πηγών ή δυνατοτήτων.

4.3 Python ως Βασική Γλώσσα Προγραμματισμού

Η Python επιλέχθηκε ως κύρια γλώσσα λόγω της απλότητας σύνταξης, της ταχύτητας και του πλούσιου οικοσυστήματος για ανάλυση κειμένου, δικτυακές κλήσεις και χειρισμό δομών δεδομένων[7]. Ως high-level, interpreted γλώσσα με δυναμική τυποποίηση και αυτόματη διαχείριση μνήμης, ευνοεί τον γρήγορο πειραματισμό (rapid prototyping) και την υλοποίηση σύνθετων ροών χωρίς υπερβολικό «τελετουργικό» κώδικα. Το αντικειμενοστραφές μοντέλο της, σε συνδυασμό με τα ισχυρά εργαλεία για κανονικές εκφράσεις, χειρισμό Unicode και JSON, την καθιστούν ιδιαίτερα κατάλληλη για εργασίες web scraping και μετασχηματισμού ημι-δομημένων δεδομένων σε τυποποιημένες εγγραφές[10]. Πρακτικά, αυτό μεταφράζεται σε αξιόπιστη εξαγωγή ονομάτων/βαθμίδων από HTML, καθαρισμό θορυβωδών δεδομένων και ομαλοποίηση κειμένου, με ελάχιστο κώδικα και υψηλή αναγνωσιμότητα.

4.4 FastAPI για το Web Framework



Σχήμα 4: Παράδειγμα interactive API

Για την υλοποίηση των web υπηρεσιών χρησιμοποιήθηκε το FastAPI, ένα σύγχρονο framework που αξιοποιεί πλήρως τα type hints της Python για αυτόματο schema generation (OpenAPI), ενσωματωμένο validation και άρτια τεκμηρίωση των endpoints[8]. Χτίζεται πάνω στο Starlette (ASGI toolkit), προσφέροντας ασύγχρονη εκτέλεση με `async/await` και υψηλή απόδοση σε σενάρια με παράλληλα αιτήματα[11]. Για τον ορισμό και τον έλεγχο των μοντέλων δεδομένων αξιοποιείται το Pydantic, το οποίο πραγματοποιεί έλεγχο τύπων και τιμών σε χρόνο εκτέλεσης και παράγει συνεπές σχήμα ανταλλαγής δεδομένων[12]. (Σχήμα 4.1) Στο πλαίσιο του έργου, αυτό σημαίνει ότι τα RESTful endpoints εκτίθενται με σαφή συμβόλαια, τα αιτήματα/απαντήσεις επικυρώνονται αυτόματα, και η web διεπαφή μπορεί να εξελίσσεται ανεξάρτητα από τον scraper, με σταθερό API, υψηλή αξιοπιστία και εύκολη ολοκλήρωση δοκιμών.

4.5 BeautifulSoup4 για το HTML Parsing

Για την ανάλυση και εξαγωγή δεδομένων από ιστοσελίδες χρησιμοποιείται η βιβλιοθήκη BeautifulSoup4, η οποία διευκολύνει την πλοήγηση και αναζήτηση στο DOM των HTML/XML εγγράφων με απλές και εκφραστικές μεθόδους[9]. Προτιμάται επειδή είναι ανεκτική σε «ατελές» ή κακώς διαμορφωμένο HTML, υποστηρίζει CSS selectors και παρέχει καθαρό API για εντοπισμό στοιχείων και ιδιοτήτων, κάτι που είναι κρίσιμο σε ετερογενείς ιστότοπους.

Η BeautifulSoup4 συνεργάζεται με parsers όπως το lxml ή το ενσωματωμένο `html.parser` για αποδοτική ανάλυση[13]. Διαθέτει πλούσιο σύνολο μεθόδων (`find`, `find_all`, `select`) για στόχευση στοιχείων με selectors, ονόματα tags ή custom συναρτήσεις. Επιπλέον, υποστηρίζει σωστό χειρισμό κωδικοσελίδων (encodings) και «καθαρισμό» προβληματικού HTML, ζητήματα που συναντώνται συχνά σε πραγματικά σενάρια scraping.

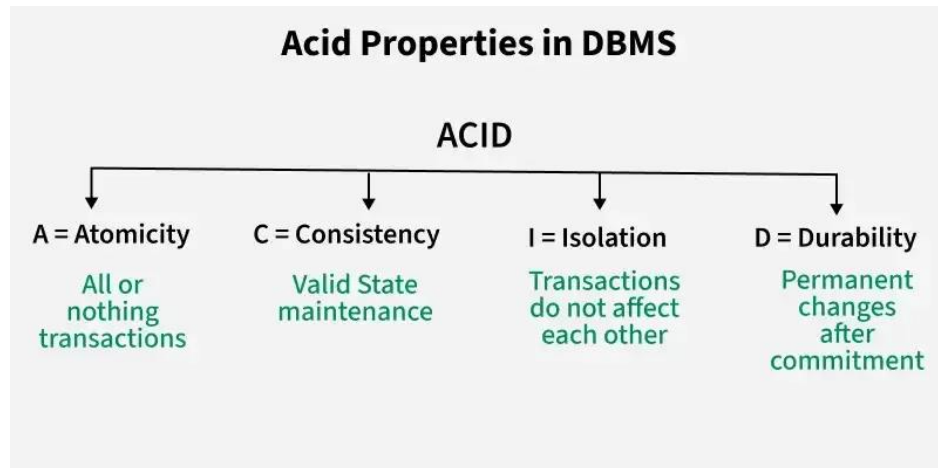
4.6 Bootstrap για το Frontend Design

Για το responsive user interface αξιοποιείται το Bootstrap, ένα ώριμο CSS/JS framework με έτοιμα components και utilities για ταχεία ανάπτυξη διεπαφών που λειτουργούν ομοιόμορφα σε διαφορετικές συσκευές[14]. Το grid σύστημά του (βασισμένο σε Flexbox/CSS Grid) επιτρέπει προσαρμόσιμες διατάξεις και συνεπή στοίχιση περιεχομένου, μειώνοντας τον χρόνο υλοποίησης και τα λάθη διάταξης.

Το Bootstrap προσφέρει ένα εκτενές οικοσύστημα από προκατασκευασμένα στοιχεία (κουμπιά, φόρμες, πίνακες, πλοήγηση) και βοηθητικές κλάσεις (spacing, typography, visibility) που επιταχύνουν την κατασκευή UI και διασφαλίζουν συνέπεια στην εμφάνιση[15]. Έτσι, η διεπαφή παραμένει καθαρή, προσβάσιμη και εύχρηστη, με ελάχιστο «χειροποίητο» CSS.

4.7 MariaDB ως Σύστημα Διαχείρισης Βάσης Δεδομένων

Για την αποθήκευση και διαχείριση δεδομένων χρησιμοποιείται το MariaDB, ένα ανοικτού κώδικα RDBMS με υψηλή απόδοση και ενεργή κοινότητα, συμβατό με το οικοσύστημα MySQL[20]. Παρέχει transactional storage engines (π.χ. InnoDB) με υποστήριξη ACID (Σχήμα 4.2), μηχανισμούς ακεραιότητας (π.χ. foreign keys) και εξελιγμένα χαρακτηριστικά για αξιόπιστη και ασφαλή αποθήκευση πληροφοριών[21].



Σχήμα 4: Ιδιότητες ACID

Το MariaDB βασίζεται στη γλώσσα SQL για ορισμό/επεξεργασία δεδομένων, προσφέρει δείκτες (indexes) για επιτάχυνση αναζητήσεων σε μεγάλα σύνολα και υποστηρίζει βελτιστοποίηση σύνθετων ερωτημάτων[21]. Στο παρόν σύστημα, αυτό επιτρέπει αποδοτική αναζήτηση προσωπικού, συσχετίσεις με εξωτερικές πηγές (π.χ. Google Scholar) και διατήρηση ιστορικότητας/ενημερώσεων με εγγυήσεις συνέπειας.

4.8 Jinja2 για το Template Engine

Για τη δημιουργία HTML templates χρησιμοποιείται το Jinja2, ένα σύγχρονο και αποδοτικό template engine για Python[18]. Υποστηρίζει σύνθεση σύνθετης λογικής εμφάνισης (loops, conditionals, filters, macros), κληρονομικότητα templates (template inheritance) και αυτόματο escaping για αποτροπή XSS, επιτρέποντας καθαρό διαχωρισμό παρουσίασης και δεδομένων[19].

Η σύνταξή του είναι οικεία και εκφραστική, διευκολύνοντας τη συντήρηση και την επαναχρησιμοποίηση δομικών στοιχείων της διεπαφής[19]. Σε συνδυασμό με το web framework, προσφέρει σταθερό και γρήγορο κύκλωμα απόδοσης δυναμικού HTML με ελάχιστη επιβάρυνση.

Κεφάλαιο 5ο: Το Scraping Microservice: Ροές, Διαμόρφωση και Επεξεργασία Δεδομένων

5.1 Στόχοι και λειτουργικές απαιτήσεις του microservice

Στόχος του scraping microservice είναι η αυτόματη, παραμετρική συλλογή και τυποποίηση στοιχείων ακαδημαϊκού προσωπικού από ετερογενείς ιστοσελίδες, χωρίς ανάπτυξη εξειδικευμένου κώδικα ανά ιστότοπο. Η λειτουργία του καθοδηγείται από αρχεία διαμόρφωσης (JSON) με ελάχιστες, σαφείς παραμέτρους (URLs, selectors, λέξεις-κλειδιά, μέθοδος parsing), ώστε να προσαρμόζεται σε διαφορετικές δομές HTML και να παράγει ομοιόμορφα αποτελέσματα. Τα δεδομένα που εξάγονται

κανονικοποιούνται (ονόματα, βαθμίδες, τύπος προσωπικού, συνάφεια), συγχωνεύονται όταν απαιτείται και διατίθενται ως αρχεία εξόδου (JSON) προς αποθήκευση στη βάση και κατανάλωση από το API.

Κατά τον αρχικό σχεδιασμό, ο στόχος ήταν ένα πλήρως αυτοματοποιημένο, configuration-driven σύστημα, όπου η ενσωμάτωση νέας πηγής επιτυγχάνεται με δηλωτική διαμόρφωση (JSON) και ελάχιστες παραμέτρους αντί για εξειδικευμένο κώδικα. Δόθηκε έμφαση στη σταθερότητα και τυποποίηση του σχήματος εξόδου, στην αναπαραγωγικότητα εκτελέσεων, στην ανθεκτικότητα απέναντι σε ετερογενές ή μη τυπικό HTML/DOM και στη συμμόρφωση με πολιτικές πρόσβασης (robots.txt, rate limiting). Υιοθετήθηκε η αρχή «configuration over code» και ισχυρός αποσυζευγμός διαμόρφωσης-υλοποίησης, ώστε η συντήρηση να περιορίζεται σε ενημερώσεις selectors/κανόνων και όχι στον κορμό του λογισμικού, ενισχύοντας επεκτασιμότητα και συντηρησιμότητα.

5.2 Επισκόπηση αρχιτεκτονικής και ροή δεδομένων

Η εκτέλεση ξεκινά με την ανάγνωση των ρυθμίσεων ιστοσελιδών (JSON) και τη δημιουργία ενός context που ορίζει ποιες πηγές θα αντληθούν και με ποια μέθοδο (Σχήμα 5.1). Κάθε αρχείο ρυθμίσεων (JSON) περιλαμβάνει τα ελάχιστα απαιτούμενα πεδία για να αναγνωριστεί η δομή της σελίδας (Selectors, λέξεις-κλειδιά, μέθοδος parsing), έτσι ώστε ο πυρήνας να παραμένει κοινός και η συμπεριφορά να προκύπτει δηλωτικά από τη ρύθμιση.

```
{
  "urls": [
    "https://www.cs.uoi.gr/staff/teaching-faculty/?cat=155&lang=en"
  ],
  "name_selector": "h4 a",
  "container_class": "article.plainbackground",
  "staff_type": "faculty",
  "debug_print": false,
  "use_cached": true,
  "rank_selector": "h3",
  "rank_keywords": [
    "Professor",
    "Associate Professor",
    "Assistant Professor"
  ],
  "search_parent_siblings_for_rank": false,
  "enabled": true,
  "rank_mapping": {
    "Professor": "Professor",
    "Associate Professor": "Associate Professor",
    "Assistant Professor": "Assistant Professor",
    "Assisant Professor": "Assistant Professor"
  },
  "parser_type": "container"
}
```

Σχήμα 5.1: Αρχείο παραμετροποίησης scraper

Πριν από κάθε ανάκτηση, εφαρμόζονται πολιτικές ευπρέπειας και συμμόρφωσης για το microservice: έλεγχος robots.txt και διαθέσιμων κανόνων, ρυθμός αιτημάτων ανά domain και αξιοποίηση cache όπου επιτρέπεται για να περιορίζονται επαναλαμβανόμενες λήψεις και να μειώνεται το φορτίο. Η ανάκτηση περιεχομένου γίνεται μέσω HTTP, με χειρισμό εξαιρέσεων (timeouts, αποτυχημένες συνδέσεις) και επαναπροσπάθειες με ελεγχόμενο backoff, ώστε η ροή να παραμένει ανθεκτική. Σε περίπτωση που κάποιοι ιστότοποι απαγορεύουν την είσοδο οποιουδήποτε crawler/scrapper, υπάρχει η εναλλακτική της ανάγνωσης ενός cached HTML όπου έχει αποθηκευτεί χειροκίνητα για την ανάγνωση του.

Αφού ληφθεί το HTML, επιλέγεται δυναμικά η κατάλληλη parsing μέθοδος, ανάλογα με τα χαρακτηριστικά της σελίδας (π.χ. «πίνακας», «λίστα», «container», «διαδοχικό κείμενο»). Η εξαγωγή βασίζεται σε selectors/κανόνες που εντοπίζουν τα ονόματα και τις βαθμίδες (με εναλλακτικές οδεύσεις, όπως από γονέα/αδερφικό στοιχείο ή μέσω mapping λέξεων-κλειδιών). Στη συνέχεια εφαρμόζεται κανονικοποίηση: ενοποίηση μορφών ονομάτων, αντιστοίχιση βαθμίδων σε τυποποιημένες τιμές, εξαγωγή τύπου προσωπικού και καθαρισμός θορύβου.

Το αποτέλεσμα περνά από μεταγενέστερους ελέγχους ποιότητας: αποφυγή διπλοεγγραφών και συγχώνευση όπου απαιτείται, βασική επικύρωση πεδίων και σταθεροποίηση του σχήματος. Η έξοδος παράγεται σε ενιαία μορφή JSON, συμβατή με τις διεπαφές προγραμματισμού (API) και με την καταχώρηση στη βάση. Παράλληλα, καταγράφονται συμβάντα και μετρικές (π.χ. χρόνοι, μέθοδος για την ανάκτηση στοιχείων, σφάλματα), ώστε να τεκμηριώνεται η συμπεριφορά του συστήματος και να διευκολύνεται η βελτίωση ή η διάγνωση θεμάτων στις εκτελέσεις.

5.3 Παραμετροποίηση Web Scraper, δομή και ιδιότητες

Η συμπεριφορά του microservice ορίζεται δηλωτικά μέσω αρχείων διαμόρφωσης (JSON), ένα ανά ιστότοπο/τμήμα ή ένα ανά σελίδα προσωπικού ενός τμήματος (Καθηγητές/Lab Staff) Η λογική είναι “configuration over code”: οι κανόνες εξαγωγής γράφονται ως παράμετροι, ώστε η ένταξη ή προσαρμογή νέας πηγής να μην απαιτεί αλλαγές στον κορμό. Κάθε αρχείο περιγράφει:

- τις πηγές προς άντληση (URLs) και βασικές ρυθμίσεις (enabled, use_cached, debug),
- τη μέθοδο parsing (π.χ. πίνακας/λίστα/container/διαδοχικό κείμενο),
- τους κανόνες εντοπισμού ονόματος/βαθμίδας και την κανονικοποίησή τους (selectors, keywords, mapping).

Με αυτόν τον τρόπο επιτυγχάνονται αναπαραγωγιμότητα (ίδιες ρυθμίσεις → ίδιο αποτέλεσμα), αποσύζευξη ρύθμισης-υλοποίησης και εστίαση της συντήρησης σε selectors/κανόνες αντί σε κώδικα.

5.4 Δομή και βασικές κατηγορίες παραμέτρων

Οι ιδιότητες εντάσσονται σε τέσσερις κατηγορίες: (α) πηγές/γενικές ρυθμίσεις, (β) μέθοδος parsing, (γ) εξαγωγή ονοματεπώνυμου, (δ) εξαγωγή βαθμίδας και κανονικοποίηση. Προαιρετικά, υποστηρίζονται και “select-based” σελίδες (όπου η ομαδοποίηση/φιλτράρισμα γίνεται από dropdown) καθώς και κάποιες προαιρετικές παραμέτρους για την κάλυψη των περισσότερων δομών.

Οι βασικές υποχρεωτικές ρυθμίσεις που πρέπει να οριστούν είναι:

- Urls (List of Strings): λίστα σελίδων προς επεξεργασία, συχνά ανά κατηγορία προσωπικού.
- enabled, use_cached, debug_print (boolean): ενεργοποίηση σάρωσης, χρήση cache, διαγνωστικά logs.
- staff_type: ετικέτα τυποποίησης (π.χ. faculty, lab_lecturer), χρησιμοποιείτε για την διαφοροποίηση σελίδων ανά τμήμα.

5.5 Ανάλυση παραμέτρου parser type

Η παράμετρος parser_type καθορίζει τον τρόπο ανάγνωσης της σελίδας. Υποστηρίζονται τέσσερις μέθοδοι: generic, container, table, sequential. Κάθε μέθοδος στοχεύει σε διαφορετικό μοτίβο DOM και αξιοποιεί διαφορετικό υποσύνολο παραμέτρων.

5.5.1 Generic

Η προεπιλεγμένη, πιο ευέλικτη μέθοδος. Εφαρμόζεται όταν τα στοιχεία προσωπικού εντοπίζονται με σταθερούς selectors, χωρίς αυστηρή ενιαία δομή (π.χ. κάρτες, λίστες ή ελεύθερη ροή HTML). Δουλεύει είτε πάνω σε όλο το έγγραφο είτε μέσα σε συγκεκριμένα containers αν δοθεί container_class. Υποστηρίζει σύνθετες περιπτώσεις όπως όνομα και βαθμίδα στο ίδιο στοιχείο (split), πλοήγηση από το node του ονόματος προς το στοιχείο που περιέχει την βαθμίδα, ή χρήση σταθερής βαθμίδας για όπου δεν υπάρχει πληροφορία αυτής.

Βασικοί παράμετροι της generic:

- name_selector: CSS selector για το στοιχείο του ονόματος (π.χ. "h3.name-imagia").
- container_class (προαιρετικό): περιορισμός της αναζήτησης σε συγκεκριμένα containers (π.χ. "card-imagia").
- rank_selector (προαιρετικό): selector για ανάγνωση βαθμίδας κοντά στο όνομα ή στο container (π.χ. "p.subtitle-imagia").
- rank_mapping: mapping με σκοπό την κανονικοποίηση/καθαρισμό κειμένου βαθμίδας (π.χ. "Καθηγητής" → "Professor").
- rank_keywords: Λεξιλόγιο για την κανονικοποίηση του κειμένου βαθμίδας, όταν δεν υπάρχει ρητή αντιστοίχιση στο mapping. π.χ. ["Associate Professor", "Assistant Professor", "Lecturer", "Professor"])

Προαιρετικές παράμετροι (για σύνθετες περιπτώσεις):

- name_concat_selectors: λίστα selectors για συνένωση τμημάτων ονόματος (π.χ. [".fname", ".lname"]). Χρησιμοποιείτε όταν το ονοματεπώνυμο είναι χωρισμένο σε πολλά στοιχεία.
- name_split_separator: custom διαχωριστής κατά την εξαγωγή κειμένου (π.χ. " " για ένωση).
- name_rank_split_separator: διαχωρισμός όταν όνομα/βαθμίδα συνυπάρχουν (π.χ. ", " στο "Ι. Παπαδόπουλος, Αναπλ. Καθηγητής").
- rank_navigation_steps: βήματα πλοήγησης από το node του ονόματος προς τη βαθμίδα (π.χ. [{"action": "parent"}, {"action": "select_one", "selector": ".rank"}]).
- fixed_rank: σταθερή βαθμίδα όταν δεν παρέχεται (π.χ. "Lecturer").

```
<div class="card-imagia">
  <div class="front-imagia">
    <div class="cover-imagia cover-gradient"></div>
    <div class="user-imagia">⋮</div>
    <div class="content-imagia">
      <h3 class="name-imagia">
        Avgoustos Tsinakos
      </h3>
      <p class="subtitle-imagia">
        Professor
      </p>
      <p class="text-center">⋮</p>
      <o class="text-center"> </o>
    </div>
  </div>
</div>
```

Σχήμα 5.2: Περιεχόμενο με καθηγητή και βαθμίδα

Περιγραφή ροής: Ο parser (generic) εντοπίζει τα containers με class "card-imagia", βρίσκει μέσα τους τα ονόματα τύπου "h3" με CSS class name "name-imagia" και τη βαθμίδα μέσω "p.subtitle-imagia" (Σχήμα 5.2). Αν η βαθμίδα δεν ανακτηθεί απευθείας, η επιλογή search_parent_siblings_for_rank την επιτρέπει αναζήτηση σε γειτονικά ή γονικά στοιχεία. Το κείμενο βαθμίδας κανονικοποιείται με το rank_keywords. Το αποτέλεσμα βγαίνει ως ομοιόμορφες εγγραφές {name, rank, staff_type} με σκοπό την αποθήκευση/κατανάλωση τους.

5.5.2 Container

Ο container parser εφαρμόζεται όταν η σελίδα οργανώνει τα μέλη σε “ενότητες/κουτιά” και τα ονόματα βρίσκονται μέσα σε αυτά τα containers, ενώ η βαθμίδα προκύπτει είτε από κείμενο/στοιχείο εντός του container είτε από ιδιότητα του ίδιου του container (π.χ. id, data-attribute). Η λογική είναι “βρες πρώτα το σύνολο” (container), έπειτα εξήγα τα ονόματα και εντόπισε/παράγαγε τη βαθμίδα στο πλαίσιο του κάθε container. Χρησιμοποιείτε συνήθως όταν δεν υπάρχει κάποιο στοιχειώδες πεδίο για την αναγνώριση ονόματος ή βαθμίδας ώστε να ψάξει γειτονικά στοιχεία.

Βασικοί παραμέτροι της Container:

- container_class (απαραίτητο): κλάση που ορίζει τα “κουτιά” με τα μέλη (π.χ. "article.plainbackground", "group", "profile-box"). Περιορίζει τη σάρωση σε σαφώς ορισμένα τμήματα της σελίδας.
- name_selector (απαραίτητο): selector για το όνομα εντός του container (π.χ. "h4 a", ".member-name a").
- rank_selector: στοιχείο/selector μέσα στον container που περιέχει το κείμενο βαθμίδας (π.χ. "h3", ".subtitle", ".role").
- rank_mapping / rank_keywords: κανονικοποίηση κειμένου βαθμίδας. Πρώτα εφαρμόζεται η ρητή αντιστοίχιση (mapping), αλλιώς καθαρισμός με βάση το λεξιλόγιο.

Προαιρετικές παράμετροι (σύνθετες περιπτώσεις):

- rank_container / rank_element: όταν η βαθμίδα «κληρονομείται» από το ίδιο το container. Παράδειγμα: rank_container: {"attr": "data-rank"} → "associate" → rank_mapping → "Associate Professor".
- use_sibling_rank / rank_sibling_selector: όταν η βαθμίδα βρίσκεται σε αδελφικό node του ονόματος (π.χ. μετά το <a> υπάρχει ...).
- rank_is_parent / use_parent_rank_with_distance / max_siblings_to_check / search_parent_siblings_for_rank: όταν η ένδειξη βαθμίδας βρίσκεται σε γονική επικεφαλίδα ή σε γειτονικά blocks, επιτρέπεται ελεγχόμενη πλοήγηση για την εύρεσή της.
- rank_from_select / select_element / select_value_attr / select_text_attr / select_all_value: όταν η σελίδα ομαδοποιεί/φιλτράρει μέσω dropdown και οι ομάδες αντιστοιχούν σε βαθμίδες (mapping τιμή/κειμένου option σε τυποποιημένη βαθμίδα).

Παράδειγμα:

- Κάθε container <div class="profile-box" data-rank="associate"> περιέχει <h4 class="name"><a>Όνομα</h4> περιέχει <h4 class="name"><a>Όνομα</h4> και

χωρίς εμφανές κείμενο βαθμίδας. Με `rank_container: {"attr":"data-rank"}` ανακτάται η «associate», κανονικοποιείται μέσω `rank_mapping` → "Associate Professor", ενώ το όνομα εξάγεται με `name_selector: "h4.name a"`.

- Σε σελίδα με headers ανά βαθμίδα και κάτω «κουτιά» με ονόματα, αν η βαθμίδα δεν υπάρχει εντός του box, ενεργοποιείται `use_parent_rank_with_distance` για να ληφθεί από γονική επικεφαλίδα.

```
<article class="plainbackground">
  <div class="media" style="padding:10px;">
    <div class="media-left"></div>
    <div class="media-body">
      <h4 style="color:#813223 !important">
        <a href="https://www.cse.uoi.gr/teams/stergios-anastasiadis/?lang=en" style="color: #813223!important;">Stergios Anastasiadis</a>
      </h4>
      <h3 style="margin-top:5px;color: !important">Professor</h3>
      <article class="col-md-12 biography" style="margin-bottom:0;margin-left:0;padding-left:0;"></article>
    </div>
  <div class="contactdiv"></div>
  <div class="col-lg-12 marginleft"></div>
</div> == $0
</article>
```

Σχήμα 5.3: Περιεχόμενο χωρίς ιδιότητες αναγνώρισης / τύπου Container

Περιγραφή ροής container parser (Σχήμα 5.3):

- Εντοπίζονται όλα τα containers με class "article.plainbackground".
- Για κάθε container, ανακτάται το όνομα από "h4 a" και η βαθμίδα από "h3".
- Το κείμενο της βαθμίδας πρώτα ελέγχεται στο `rank_mapping` (ώστε να διορθωθούν τυπογραφικά ή παραλλαγές). Αν δεν υπάρχει αντιστοίχιση, εφαρμόζεται καθαρισμός με `rank_keywords`.
- Παράγονται ομοιόμορφες εγγραφές {name, rank, staff_type} έτοιμες για αποθήκευση και κατανάλωση από το API.

5.5.3 Table

Ο table parser εφαρμόζεται όταν το προσωπικό παρουσιάζεται σε κλασικούς HTML πίνακες (table/tr/td), με εγγραφές ανά γραμμή. Η λογική είναι “διάβασε γραμμή-γραμμή, βρες το κελί ονόματος και, όπου απαιτείται, εντόπισε/παράγαγε τη βαθμίδα από κοντινά στοιχεία ή επικεφαλίδες”.

```
{
  "urls": ["https://www.iee.ihu.gr/en/staff_category/faculty-member/"],
  "name_selector": "td b a",
  "container_class": "table",
  "staff_type": "faculty",
  "debug_print": true,
  "is_table_data": true,
  "rank_keywords": [
    "Associate Professor",
    "Assistant Professor",
    "Lecturer",
    "Professor"
  ],
  "search_parent_siblings_for_rank": false,
  "enabled": true,
  "use_cached": true,
  "parser_type": "table"
}
```

Σχήμα 5.4: Παράδειγμα ρυθμίσεων για τύπου Table

Ρυθμίσεις για ανάγνωση table data (Σχήμα 5.4):

- `urls`: ["https://www.iee.ihu.gr/en/staff_category/faculty-member/"] — σελίδα πίνακα με τα μέλη.
- `is_table_data`: `true` — δηλώνει tabular περιεχόμενο.
- `container_class`: "table" — φιλτράρει/στοχεύει πίνακες με αυτή την κλάση (αν απαιτείται).
- `name_selector`: "td b a" — εντοπίζει το όνομα στο κελί της γραμμής.
- `rank_keywords`: ["Associate Professor", "Assistant Professor", "Lecturer", "Professor"] — λεξιλόγιο κανονικοποίησης βαθμίδας για να ανακτηθεί κείμενο βαθμίδας.

```

<table>
  <thead>...</thead>
  <tbody>
    <tr>...</tr>
    <tr>
      <td valign="middle" style="text-align:center;width:80px;">...</td>
      <td>
        <b>
          <a href="https://www.iee.ihu.gr/staff/antoniou-efstathios/" role="link">Αντωνίου Ευστάθιος</a>
        </b>
        <br>
        " (Πρόεδρος Τμήματος)"
        <br>
        <span class="dashicons dashicons-welcome-learn-more">...</span>
        " Αριθμητικές και Συμβολικές Υπολογιστικές Μέθοδοι για τη Μαθηματική Θεωρία Συστημάτων"
        <br>
        <span class="dashicons dashicons-location">...</span>
        " Κτίριο Η, Ισόγειο, Γραφείο Προϊσταμένου"
      </td>
    </tr> == $0
  </tbody>
</table>

```

Σχήμα 5.5: Περιεχόμενο τύπου Table

Περιγραφή ροής:

Κατά την εκτέλεση, εντοπίζεται ο σχετικός HTML πίνακας (με ενδεχόμενο φιλτράρισμα μέσω `container_class`) και αναλύονται οι γραμμές του, συνήθως εξαιρώντας την επικεφαλίδα. Για κάθε γραμμή, το όνομα ανακτάται με τον `name_selector`, ενώ η βαθμίδα εντοπίζεται είτε στο ίδιο κελί/γειτονικό στοιχείο μέσω `rank_selector` είτε, όταν δηλώνεται σε επικεφαλίδα τμήματος/πίνακα, με ελεγχόμενη πλοήγηση προς γονικά ή γειτονικά στοιχεία (`rank_is_parent`, `use_parent_rank_with_distance`). (Σχήμα 5.5) Το κείμενο της βαθμίδας τυποποιείται κατά προτεραιότητα με `rank_mapping` (ρητές αντιστοιχίσεις) και, όπου δεν υπάρχει ταύτιση, καθαρίζεται/αναγνωρίζεται με βάση το `rank_keywords`. Το αποτέλεσμα αποδίδεται ως ομοιόμορφες εγγραφές `{name, rank, staff_type}`, έτοιμες για αποθήκευση και κατανάλωση από το API.

- Όταν η βαθμίδα δεν εμφανίζεται σε κάθε γραμμή αλλά δηλώνεται σε επικεφαλίδα τμήματος/πίνακα, αξιοποιούνται `'rank_is_parent'` ή `'use_parent_rank_with_distance'` για πλοήγηση και κληρονόμηση της βαθμίδας στις γραμμές.
- Αν υπάρχουν πολλοί πίνακες ανά σελίδα, το `'table_class'` (όπου εφαρμόζεται) βοηθά να περιοριστεί η σάρωση στον σωστό πίνακα.

5.5.4 Sequential

Ο `sequential parser` χρησιμοποιείται όταν η σελίδα είναι δομημένη σε “ενότητες βαθμίδας” (`headers/sections`) και τα ονόματα ακολουθούν κατά ομάδες κάτω από τον αντίστοιχο τίτλο. Σε αντίθεση με το `container`, δεν απαιτεί η βαθμίδα να υπάρχει μέσα σε κάθε “κουτί” μέλους· αντιθέτως “κληρονομείται” από τον πιο πρόσφατο τίτλο ενότητας που εντοπίστηκε. Έτσι, η αντιστοίχιση

rank→ονόματα γίνεται συμφραζομενικά (context-driven) με βάση τη ροή του DOM, κάτι που λύνει περιπτώσεις όπου το κάθε entry δεν φέρει ρητά τη βαθμίδα.

[illegible]

Σχήμα 5.6: Περιεχόμενο τύπου Sequential.

Παράδειγμα (e-ce.uth.gr) (Σχήμα 5.6)

Η σελίδα περιέχει τίτλους ενότητας βαθμίδας (π.χ. μέσα σε `div.heading h2`) και, από κάτω, “κουτιά” με στοιχεία μέλους (π.χ. `div.toggle-wrap`) που περιλαμβάνουν το όνομα (π.χ. `h3.trigger a`). Η ρύθμιση `parser_type: "sequential"`, με `rank_selector: "div.heading h2"`, `container_class: "toggle-wrap"`, `name_selector: "h3.trigger a"` και περιορισμό πεδίου μέσω `main_content_selector: "div.entry"`, επιτρέπει να τεθεί “τρέχουσα βαθμίδα” όταν συναντηθεί ο τίτλος ενότητας και να αποδοθεί αυτή η βαθμίδα σε όλα τα επόμενα `entries` μέχρι τον επόμενο τίτλο. Το `rank_mapping` ενοποιεί παραλλαγές/γλωσσικά ισοδύναμα, ενώ το `rank_keywords` λειτουργεί ως λεξιλόγιο καθαρισμού όταν δεν υπάρχει ακριβής αντιστοίχιση. Προαιρετικά, μπορεί να γίνει και τοπική υπέρβαση μέσω `use_sibling_rank` όταν η βαθμίδα εμφανίζεται δίπλα στο όνομα σε ορισμένα `entries`.

Περιγραφή ροής:

Ο parser περιορίζει πρώτα τον χώρο αναζήτησης (π.χ. div.entry) και διατρέχει σειριακά τα DOM nodes. Όταν εντοπίσει element που ταιριάζει στον “rank_selector”, θέτει τη “τρέχουσα” τυποποιημένη βαθμίδα (με rank mapping/rank keywords). Κατόπιν, για κάθε στοιχείο που ταιριάζει στο “container class”,

εξάγει το όνομα με `name_selector` και αποδίδει σε αυτό το entry τη τρέχουσα βαθμίδα. Αν σε μεμονωμένες περιπτώσεις η βαθμίδα βρίσκεται δίπλα στο όνομα, μπορεί να γίνει στοχευμένη ανάγνωση (π.χ. `sibling`) και να υπερισχύσει τοπικά. Το τελικό αποτέλεσμα είναι ομοιόμορφες εγγραφές `{name, rank, staff_type}` που απηχούν πιστά τη δομή “τίτλος ενότητας → λίστα μελών”.

5.6 Συμμόρφωση με Robots Exclusion Protocol

Το Robots Exclusion Protocol ορίζει έναν απλό μηχανισμό δήλωσης πολιτικών πρόσβασης μέσω του αρχείου `/robots.txt`, με κανόνες ανά user-agent (π.χ. `User-agent`, `Disallow`, `Allow`, `Sitemap`)[1]. Δεν αποτελεί μηχανισμό ασφάλειας, αλλά σύμβαση ευπρέπειας που υποδεικνύει τι επιτρέπεται να αντληθεί και με ποιον ρυθμό. Η τήρησή του μειώνει το φορτίο στους εξυπηρετητές, αποτρέπει αποκλεισμούς και προάγει συνετή, προβλέψιμη συμπεριφορά υπηρεσιών που περιηγούνται στον ιστό.

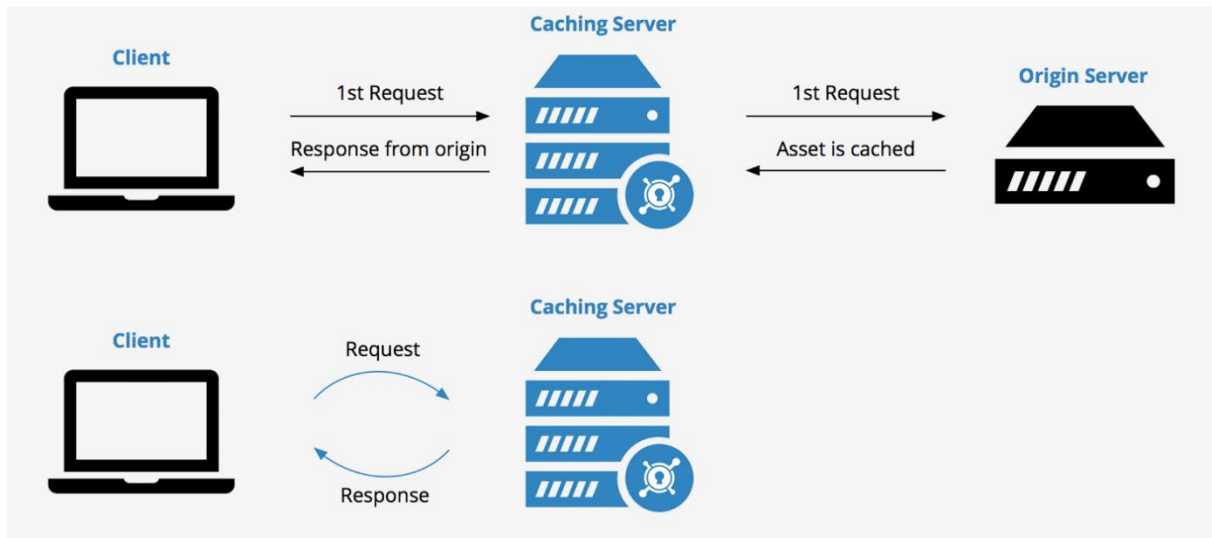
Στο παρόν έργο το microservice σέβεται τις πολιτικές crawling ως εξής[1]:

- Πριν από κάθε αίτημα ανακτά και αναλύει το `/robots.txt` του εκάστοτε domain (caching ανά domain) και ελέγχει αν ο δηλωμένος user-agent επιτρέπεται να προσπελάσει το URL (`can_fetch`). Αν δεν επιτρέπεται, η σελίδα παραλείπεται.
- Εφόσον υπάρχει ένδειξη καθυστέρησης (`crawl-delay`) για τον συγκεκριμένο user-agent, αυτή εφαρμόζεται πριν από το επόμενο αίτημα· όπου δεν παρέχεται, εφαρμόζεται συντηρητική προεπιλογή καθυστέρησης.
- Γίνεται χρήση τοπικής cache HTML (όπου επιτρέπεται) ώστε να αποφεύγονται περιττές επαναληπτικές ανακτήσεις του ίδιου περιεχομένου.

Σημειώνεται ότι ορισμένες οδηγίες (όπως το `crawl-delay`) δεν είναι τμήμα του βασικού προτύπου και δεν υποστηρίζονται ομοιόμορφα από όλους τους εξυπηρετητές· όπου παρέχονται, λαμβάνονται υπόψη. Σε περιπτώσεις που το `/robots.txt` δεν είναι προσβάσιμο (π.χ. προσωρινό σφάλμα), η υπηρεσία συνεχίζει με προσεκτικό ρυθμό αιτημάτων και ενεργή cache, χωρίς να επιχειρεί παράκαμψη περιορισμών. Συνολικά, η προσέγγιση εντάσσεται σε πολιτικές «ευπρεπούς» crawling: έλεγχος δικαιωμάτων πρόσβασης ανά URL, σεβασμός ρυθμών και περιορισμός του φορτίου μέσω caching.

5.7 Caching HTML: Σκοπός, Αρχές και Εφαρμογή

Το web caching αποσκοπεί στη μείωση των καθυστερήσεων και του φορτίου στους εξυπηρετητές, στη βελτίωση της απόδοσης και στη διατήρηση συνεπούς εμπειρίας μέσω αποθήκευσης και επαναχρησιμοποίησης προηγουμένως ανακτημένου περιεχομένου. Σε επίπεδο προτύπων, οι μηχανισμοί καθορίζονται από την οικογένεια RFC για το HTTP, όπου ορίζονται οι πολιτικές cache (π.χ. `Cache-Control`, `Freshness`, `Revalidation`), οι κανόνες επικαιρότητας και οι αλληλεπιδράσεις με `conditional requests` (`ETag/If-None-Match`, `Last-Modified/If-Modified-Since`)[23][25]. Πρακτικά, οδηγίες όπως η φρεσκάδα, η επαλήθευση και η συμπεριφορά «stale» περιγράφονται ως πολιτικές που οι πελάτες και οι ενδιάμεσες cache μπορούν να τηρούν για αποδοτικότερες ανακτήσεις[24].



Σχήμα 5.7: Ροή δημιουργίας και ανάκτησης Cached περιεχομένου

Στο παρόν έργο, το caching HTML ενεργοποιείται δηλωτικά (`use_cached`) και εφαρμόζεται σε επίπεδο αρχείων ανά URL/πηγή. Όταν υπάρχει τοπικό στιγμιότυπο, χρησιμοποιείται για ανάλυση χωρίς εκ νέου δίκτυο· διαφορετικά, γίνεται ανάκτηση μέσω HTTP (Σχήμα 5.7), σέβοντας το `robots.txt` και αποθήκευση στιγμιότυπου για μελλοντική χρήση. Η προσέγγιση αυτή επιταχύνει κύκλους παραμετροποίησης (ιδίως κατά τον ορισμό selectors), μειώνει τα αιτήματα προς τρίτους ιστότοπους και ενισχύει την αναπαραγωγικότητα (ίδια είσοδος → ίδιο αποτέλεσμα). Για μελλοντική βελτίωση προβλέπεται αξιοποίηση conditional requests (ETag/Last-Modified) και πολιτικών λήξης/επαλήθευσης, σύμφωνα με τις προδιαγραφές HTTP caching[23][25], ώστε να επιτυγχάνεται καλύτερη ισορροπία μεταξύ επικαιρότητας και απόδοσης.

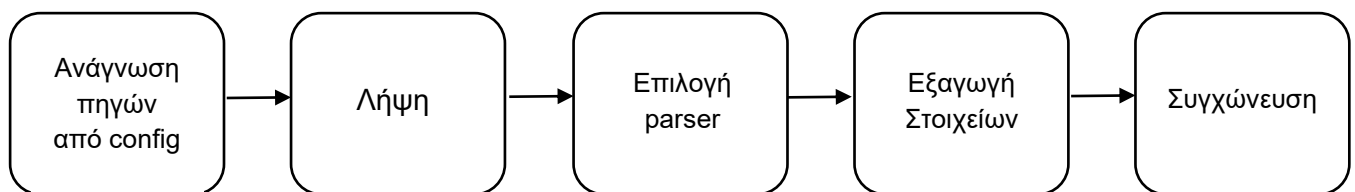
Μια πρόσθετη χρησιμότητα του μηχανισμού cache είναι ότι επιτρέπει «offline» επεξεργασία στιγμιότυπων σελίδων όταν η απομακρυσμένη πρόσβαση από αυτοματοποιημένους agents απορρίπτεται (π.χ. 403/anti-bot, προσωρινοί περιορισμοί). Σε τέτοιες περιπτώσεις μπορεί να ληφθεί, με κανονικό φυλλομετρητή ή εγκεκριμένο εργαλείο αρχειοθέτησης, ένα στατικό στιγμιότυπο της σελίδας (fully rendered HTML) και να τοποθετηθεί στο τοπικό cache (με ενεργό `use_cached`), ώστε το microservice να το αναλύσει χωρίς επιπλέον δικτυακά αιτήματα. Η προσέγγιση αυτή διατηρεί την αναπαραγωγικότητα (σταθερό εισαγόμενο HTML → σταθερό αποτέλεσμα), μειώνει το φορτίο προς τρίτους εξυπηρετητές και επιτρέπει συνέχιση δοκιμών/παραμετροποίησης selectors ακόμη και όταν η ζωντανή πηγή μπλοκάρει αυτόματα αιτήματα. Σημειώνεται ότι η πρακτική αυτή προϋποθέτει σεβασμό στους όρους χρήσης και στα δικαιώματα πρόσβασης του εκάστοτε ιστοτόπου και πρέπει να χρησιμοποιείται αποκλειστικά για νόμιμο, επιτρεπτό και δημόσια διαθέσιμο περιεχόμενο.

5.8 Parsing Pipeline: Λήψη, Ανάλυση, Κανονικοποίηση, Εξαγωγή

Η εκτέλεση ξεκινά από το «ορχηστρικό» στρώμα, το οποίο διαβάζει τα ενεργά αρχεία διαμόρφωσης (JSON) και παράγει το πρόγραμμα άντλησης. Για κάθε URL εφαρμόζονται έλεγχος πρόσβασης μέσω `robots.txt` και τυχόν καθυστέρηση (`crawl-delay`), έπειτα επιλέγεται πηγή περιεχομένου: αν το

"use_cached" είναι ενεργό και υπάρχει στιγμιότυπο, χρησιμοποιείται το τοπικό HTML, διαφορετικά γίνεται HTTP ανάκτηση και ταυτόχρονη αποθήκευση στο cache. Σε αυτό το στάδιο καταγράφονται βασικές μετρικές (π.χ. χρόνοι, επιτυχία/αποτυχία αιτημάτων) ώστε να είναι εφικτή η διάγνωση.

Ακολουθεί η ανάλυση DOM. Με βάση το parser_type επιλέγεται η κατάλληλη στρατηγική: στον generic περιορίζεται (αν δοθεί) ο χώρος αναζήτησης σε συγκεκριμένα containers και στοχεύονται ευθέως τα στοιχεία με selectors· στον container εντοπίζονται πρώτα τα «κουτιά» και μέσα σε καθένα εξάγεται το όνομα· στον table αναλύονται οι γραμμές ενός HTML πίνακα· στον sequential ορίζεται «τρέχουσα» βαθμίδα από headers/sections και αποδίδεται στα επόμενα entries μέχρι να αλλάξει ενότητα. Η εξαγωγή του ονόματος γίνεται με τον name_selector, με υποστήριξη σε περιπτώσεις όπου το ονοματεπώνυμο είναι σπασμένο σε τμήματα (συνένωση/διαχωρισμός). Η εξαγωγή της βαθμίδας είναι δομική (selector, parent/sibling πλοήγηση ή κληρονόμηση από section), όχι απλή αναζήτηση λέξεων.



Σχήμα 5.8: Διάγραμμα ροής Web Scraper

Στη συνέχεια πραγματοποιείται κανονικοποίηση. Το κείμενο της βαθμίδας πρώτα αντιστοιχίζεται μέσω rank_mapping (ρητές αντιστοιχίσεις/διορθώσεις) και, αν δεν προκύψει αποτέλεσμα, καθαρίζεται/αναγνωρίζεται με «λεξιλόγιο» rank_keywords ώστε να παραχθεί τυποποιημένη τιμή. Παράλληλα εφαρμόζονται καθαρισμοί στο όνομα (π.χ. αφαίρεση τίτλων/παρενθέσεων, ομαλοποίηση διαστημάτων/Unicode) και ορίζεται το staff_type από τη διαμόρφωση. Όπου η πληροφορία λείπει από τη σελίδα, μπορεί να χρησιμοποιηθεί fixed_rank ως fallback.

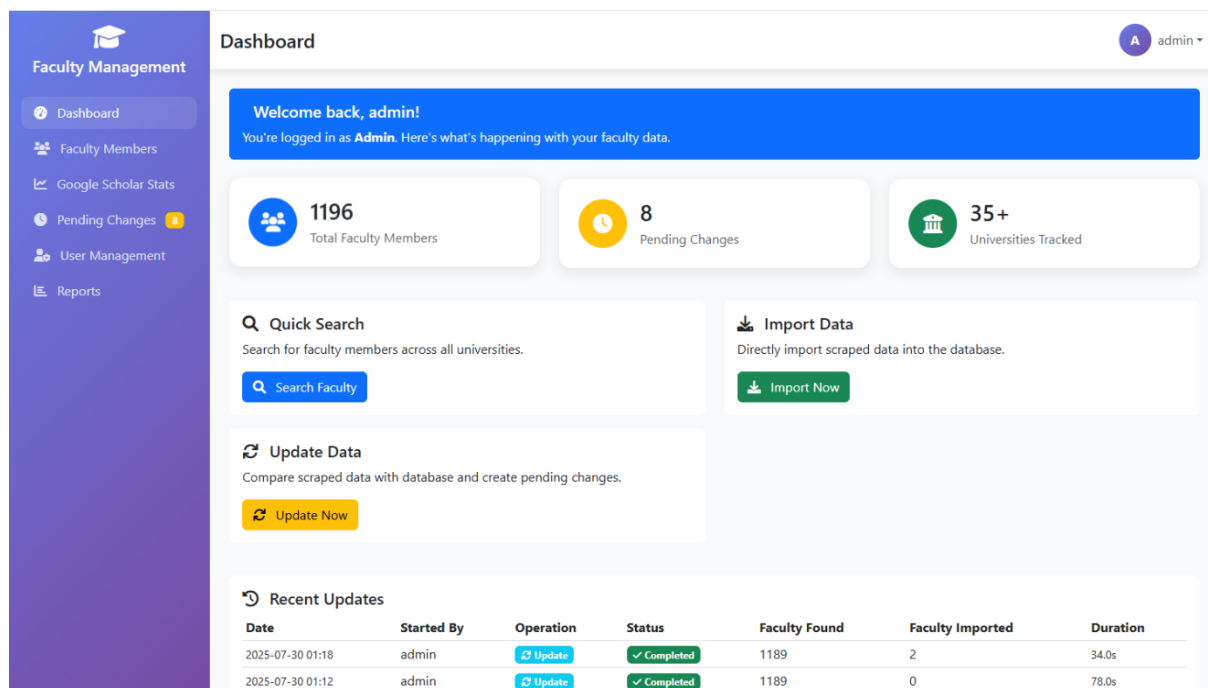
Τέλος, τα αποτελέσματα ενοποιούνται και παράγεται η έξοδος. Εγγραφές από πολλαπλά URLs του ίδιου ιστότοπου συγχωνεύονται όταν αναγνωριστούν ως ταυτόσημες (ομαλοποιημένο όνομα + βαθμίδα), αποφεύγοντας διπλοεγγραφές. Παράγεται τυποποιημένο JSON (ανά πηγή και συνολικό), συμβατό με τα APIs και έτοιμο για αποθήκευση στη βάση (Σχήμα 5.8). Τα σφάλματα καταγράφονται με επαρκές πλαίσιο (URL, selector, στάδιο ροής) και οι εξαιρέσεις αντιμετωπίζονται με ήπια αποτυχία (timeouts, ελεγχόμενα retries), ώστε η διαδικασία να παραμένει προβλέψιμη.

Κεφάλαιο 6ο: Η Web Εφαρμογή - Ρόλος, Ροές και Διεπαφές

6.1 Σκοπός και ρόλος της web εφαρμογής

Η web εφαρμογή λειτουργεί ως το λειτουργικό «control panel» του οικοσυστήματος: καταναλώνει τα τυποποιημένα δεδομένα του microservice, τα οργανώνει σε ενιαίο σχήμα, επιτρέπει επαλήθευση/έγκριση αλλαγών, παρέχει αναζητήσεις και αναφορές, και εκθέτει τις απαραίτητες ροές διαχείρισης. Στόχος είναι η ασφαλής και αποτελεσματική αξιοποίηση των δεδομένων (προβολή, έλεγχος ποιότητας, λήψη αποφάσεων), με σταθερά συμβόλαια API και σαφή όρια ρόλων.

Το Dashboard προσφέρει συνοπτική εικόνα κατάστασης: βασικούς μετρητές (π.χ. πλήθος μελών, πρόσφατες ενημερώσεις, εκκρεμείς αλλαγές), ενδεικτικούς δείκτες αντιστοίχισης Google Scholar και γρήγορες ενέργειες πλοήγησης προς σελίδες ελέγχου (Faculty, Pending Changes, Reports). Λειτουργεί ως σημείο εκκίνησης για τις συνήθεις ροές, χωρίς να υποκαθιστά τις εξειδικευμένες σελίδες.



Σχήμα 6.1: Αρχική οθόνη διαχειριστή

Το σύστημα εφαρμόζει role-based access control (RBAC), ώστε η πρόσβαση σε δεδομένα και ενέργειες να καθορίζεται από τον ρόλο του χρήστη. Η αυθεντικοποίηση εκδίδει διαπιστευτήρια συνεδρίας και στη συνέχεια η εξουσιοδότηση ελέγχει ανά αίτημα αν ο χρήστης διαθέτει τα απαραίτητα δικαιώματα για προβολή ή μεταβολή πόρων. Ευαίσθητες πληροφορίες και ενέργειες τροποποίησης προστατεύονται από ελέγχους στο επίπεδο API, ενώ η διεπαφή προσαρμόζει δυναμικά την ορατότητα κουμπιών/σελίδων βάσει του ρόλου, ακολουθώντας την αρχή του ελάχιστου προνομίου. (Σχήμα 6.1)

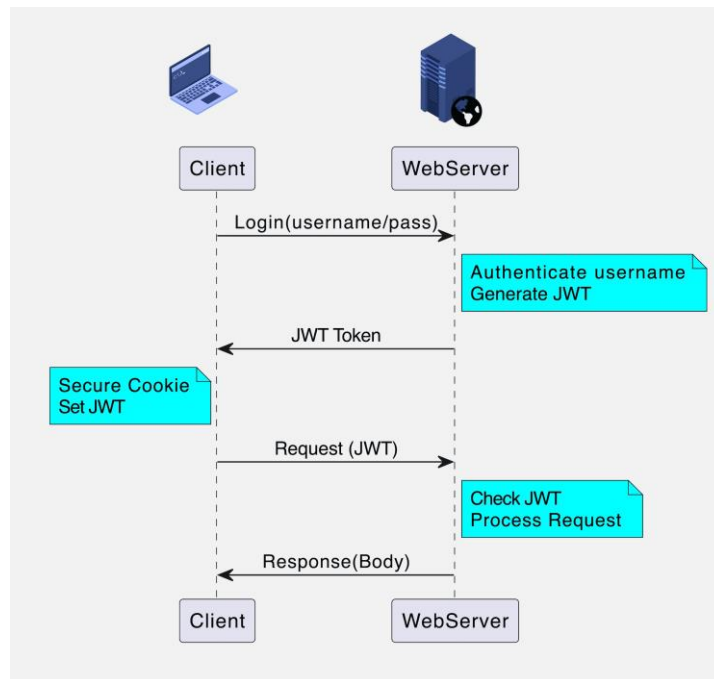
Ρόλοι και άδειες πρόσβασης:

- **Viewer:** ανάγνωση/προβολή σελίδων (Faculty, Google Scholar Stats, βασικά Reports). Δράσεις τροποποίησης και σελίδες διαχείρισης αποκρύπτονται ή απενεργοποιούνται.
- **Approver:** κληρονομεί τα δικαιώματα του Viewer, συν χειρισμός «Pending Changes» (έγκριση/απόρριψη), ώστε να διασφαλίζεται ο έλεγχος πριν από την οριστική ενημέρωση.
- **Admin:** πλήρης πρόσβαση (Users, ρυθμίσεις/ενσωματώσεις, προχωρημένες αναφορές). Βλέπει όλες τις σελίδες και έχει πρόσβαση σε όλες τις ενέργειες.

Η πλοήγηση και τα UI στοιχεία προσαρμόζονται στον ρόλο: στοιχεία που δεν επιτρέπονται είναι αόρατα ή μόνο-ανάγνωσης, ώστε να αποφεύγονται λάθη και να παραμένει καθαρή η εμπειρία χρήσης. Σε επόμενες ενότητες αναλύονται οι ροές διασύνδεσης με το microservice, τα API και οι βασικές σελίδες.

6.2 Αυθεντικοποίηση με JSON Web Tokens (JWT)

Τα JSON Web Tokens (JWT) είναι συμπαγή, URL-safe διαπιστευτήρια που μεταφέρουν «claims» σε τρία μέρη (header.payload.signature), κωδικοποιημένα σε Base64url και υπογεγραμμένα για ακεραιότητα[26]. Το header δηλώνει τον αλγόριθμο/τύπο, το payload περιέχει δηλώσεις όπως υποκείμενο/ρόλους/λήξη (π.χ. sub, role, exp), ενώ η signature επιτρέπει στον server να επαληθεύει ότι τα claims δεν αλλοιώθηκαν. Το μοντέλο είναι «stateless»: μετά το login, ο server εκδίδει ένα access token και ο πελάτης το στέλνει σε κάθε αίτημα (Σχήμα 6.2). Η προσέγγιση χρησιμοποιείται ευρέως σε APIs, μικροϋπηρεσίες και SPA/mobile εφαρμογές και αποτελεί τυπικό μηχανισμό «bearer token» στο OAuth 2.0[27].



Σχήμα 6.2: Διάγραμμα ροής για αυθεντικοποίηση μέσω JWT

Στην πράξη, τα JWT ενσωματώνουν τόσο στοιχεία ταυτότητας όσο και εξουσιοδότησης. Για RBAC, ο ρόλος (π.χ. admin/approver/viewer) τοποθετείται ως claim και ελέγχεται ανά αίτημα στον web server, επιβάλλοντας την αρχή του ελάχιστου προνομίου. Συνήθη claims ασφαλείας είναι iss (issuer), aud (audience), iat (issued-at), nbf (not-before), exp (expiration). Η επαλήθευση υπογραφής/claims γίνεται σε κάθε κλήση χωρίς αναζήτηση κατάστασης σε session store, διευκολύνοντας οριζόντια κλιμάκωση (πολλαπλά instances/containers)[26].

Σε σύγκριση με παραδοσιακά server-side sessions (stateful), τα JWT επιτρέπουν απλούστερη κλιμάκωση, αλλά απαιτούν προσοχή σε ανάκληση (revocation) και διάρκεια ζωής. Καλές πρακτικές είναι: σύντομη λήξη access tokens, προαιρετικά refresh tokens με rotation, έλεγχοι iss/aud/nbf/iat, ισχυροί αλγόριθμοι (HS256 ή ασύμμετροι RS256/ES256), αποκλειστική χρήση HTTPS, και προσεκτική αποθήκευση (π.χ. HttpOnly/secure cookies ή in-memory και ποτέ σε localStorage για να περιοριστεί ο κίνδυνος XSS)[26][28]. Όταν απαιτείται «αναγκαστική» ανάκληση (π.χ. ύποπτη

δραστηριότητα), μπορεί να χρησιμοποιηθεί λίστα αποκλεισμού (denylist) ή μείωση TTL ώστε να ελαχιστοποιείται το παράθυρο έκθεσης[28].

Στο παρόν σύστημα, ο μηχανισμός αυθεντικοποίησης εκδίδει JWT που ενσωματώνει την ταυτότητα και τον ρόλο του χρήστη. Τα endpoints ελέγχουν τη γνησιότητα/ισχύ του token και τα claims προτού εκτελέσουν ενέργειες. Κατ' αυτόν τον τρόπο, η web εφαρμογή εφαρμόζει RBAC σε επίπεδο API και UI: τα μη επιτρεπτά χειριστήρια/σελίδες κρύβονται ή καθίστανται μόνο-ανάγνωσης, ενώ τα ευαίσθητα endpoints απαιτούν τα κατάλληλα claims (π.χ. admin) για μεταβολές.

6.3 Ροές εργασίας και διασύνδεση με το microservice

Η υπηρεσία scraping είναι ανεξάρτητη και μπορεί να εκτελεστεί αυτόνομα, αλλά και μέσω API Endpoint από τη web εφαρμογή. Η έξοδος της παράγεται ως τυποποιημένα, κανονικοποιημένα αρχεία JSON (ανά πηγή και ένα συγκεντρωτικό), τα οποία η εφαρμογή καταναλώνει και επεξεργάζεται. Η προσέγγιση αυτή απλοποιεί την ενσωμάτωση, καθώς δεν απαιτούνται ειδικές προσαρμογές ή χρήση κάποιου timer μεταξύ εκτέλεσης scraper και ανάγνωσης δεδομένων. Επιπλέον, διευκολύνονται ο προγραμματισμός περιοδικών εκτελέσεων, οι στοχευμένες δοκιμές σε σελίδες και η απομόνωση σφαλμάτων από τη λειτουργία της διεπαφής. Συνολικά, το σχήμα συνεργασίας παραμένει ευέλικτο, επεκτάσιμο και φιλικό σε μελλοντικές ενσωματώσεις.

Η web εφαρμογή εκτελεί τον scraper (χειροκίνητα ή προγραμματισμένα) και στη συνέχεια καταναλώνει τα αποτελέσματά του: τα JSON ανά πηγή και το συγκεντρωτικό “combined” JSON. Δεν υπάρχει σύγχρονη (real-time) κλήση μεταξύ τους· η εφαρμογή διαβάσει τα παραγόμενα αρχεία από τον καθορισμένο φάκελο/μονοπάτι και συνεχίζει τη ροή εισαγωγής.

Λειτουργικά, ο κύκλος έχει ως εξής: το microservice εκτελείται και παράγει τα JSON ανά τμήμα μαζί με ένα συγκεντρωτικό “combined” JSON. Η web εφαρμογή, μέσω διαδικασίας εισαγωγής, διαβάσει το συγκεντρωτικό αρχείο, επικυρώνει το σχήμα/πεδία, εφαρμόζει κανονικοποίηση (ονόματα/βαθμίδες/τύπο προσωπικού) και πραγματοποιεί σύγκριση (diff) με τα ήδη αποθηκευμένα δεδομένα. Οι διαφορές καταχωρίζονται ως «Pending Changes» για έλεγχο, ο αρμόδιος ρόλος (approver/admin) εξετάζει, εγκρίνει ή απορρίπτει τις προτεινόμενες αλλαγές, και μόνο οι εγκεκριμένες μεταβολές εφαρμόζονται στη βάση. Η διαδικασία συνοδεύεται από καταγραφή ενεργειών (audit trail) και ορατότητα σε μετρικές (π.χ. πόσες νέες εγγραφές, πόσες ενημερώσεις) ώστε να είναι διαφανής και ανιχνεύσιμη.

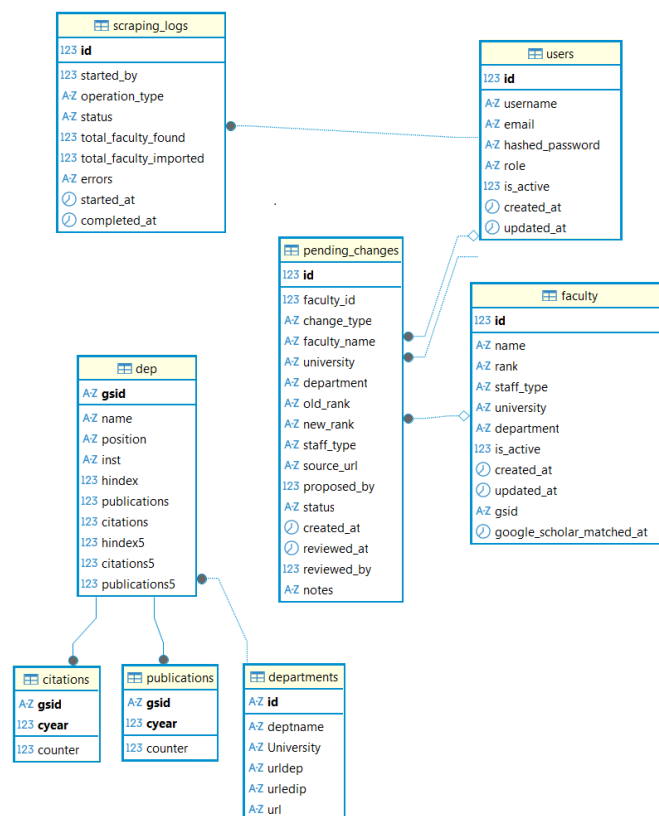
Το περιεχόμενο του combined JSON περιλαμβάνει τα ελάχιστα απαιτούμενα πεδία (name, rank, staff_type, university/department) σε κανονικοποιημένη μορφή (καθαρισμένα ονόματα, ενοποιημένες βαθμίδες). Η κανονικοποίηση είναι κρίσιμη γιατί επιτρέπει ένα ενιαίο σχήμα για δεδομένα που προέρχονται από ετερογενείς πηγές, μειώνει σφάλματα και ασυνέπειες, και διασφαλίζει ότι οι περιοδικές ενημερώσεις της βάσης γίνονται με ασφάλεια και ακρίβεια. Άγνωστες ή ελλιπείς τιμές επισημαίνονται για ανασκόπηση, ενώ κατά την εισαγωγή εφαρμόζονται σταθερά κλειδιά ταυτοποίησης (π.χ. normalized name + department) και έλεγχοι διπλοεγγραφών, ώστε με τα ίδια εισαγόμενα δεδομένα να μην δημιουργούνται νέες μεταβολές. Το microservice παραμένει ανεξάρτητο και επεκτάσιμο, ενώ η web εφαρμογή αναλαμβάνει την ευθυγράμμιση και τον συγχρονισμό των εγγραφών στη βάση, εφαρμόζοντας τους κανόνες ποιότητας και τα δικαιώματα πρόσβασης.

6.4 API διεπαφές (REST endpoints, σχήματα)

Η web εφαρμογή εκθέτει REST API endpoints που καταναλώνονται μέσω του web περιβάλλοντος διεπαφής (HTML σελίδες) και από εσωτερικές διαδικασίες της εφαρμογής. Μέρους των endpoints είναι «δημόσια» με την έννοια ότι σχετίζονται με την έναρξη/λήξη συνεδρίας και προσπελαύνονται χωρίς έλεγχο ταυτότητας, ενώ τα υπόλοιπα είναι προστατευμένα και χρησιμοποιούνται από την εφαρμογή μετά την αυθεντικοποίηση (μέσω JWT cookie) ή/και για εσωτερική λειτουργία. Ακολουθώς, παρουσιάζονται το σχήμα της βάσης καθώς και, συνοπτικά τα δημόσια και τα προστατευμένα endpoints που παρέχονται.

6.5 Διάγραμμα βάσης δεδομένων

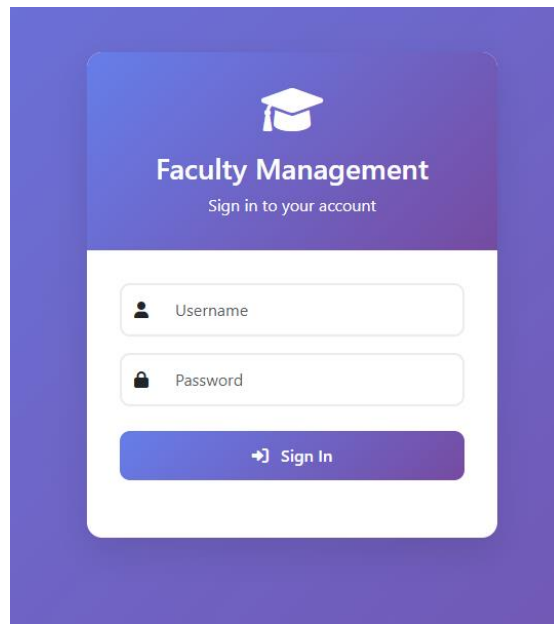
Η βάση δεδομένων οργανώνει τα στοιχεία προσωπικού και τις ροές εργασίας γύρω από έναν κεντρικό πίνακα μητρώου και συμπληρωματικούς πίνακες για ελέγχους, logs και μετρικές. Κεντρικός είναι ο πίνακας «faculty» (μέλη, βαθμίδα, τύπος προσωπικού, τμήμα/ίδρυμα) με πεδίο «gsid» για συσχέτιση σε δείκτες Google Scholar. Οι αλλαγές δεν εφαρμόζονται άμεσα: πρώτα καταχωρίζονται ως εκκρεμείς (pending_changes) και έπειτα εγκρίνονται/απορρίπτονται. Η διαδικασία scraping/σύγκρισης ιχνηλατείται στον πίνακα scraping_logs. Οι μετρικές Scholar αντλούνται από τον πίνακα dep (per-GSID), ενώ οι αντιστοιχίσεις πανεπιστημίων/τμημάτων στηρίζονται στον πίνακα departments. Η ασφάλεια/ρόλοι χρηστών υλοποιούνται μέσω users.



Σχήμα 6.3: Διάγραμμα ER βάσης δεδομένων

Οι σχέσεις είναι σαφείς και ελάχιστες: το «faculty.gsid» γίνεται LEFT JOIN στο «dep.gsid» για να φέρει μετρικές κατά την προβολή, ενώ ο «pending_changes» συνδέεται προαιρετικά με «faculty» και οπωσδήποτε με «users» (προτείνοντες/εγκρίνοντες), το «scraping_logs» συνδέεται με users (ποιος εκκίνησε). Το σχήμα υποστηρίζει αναλυτική παρακολούθηση (timestamps, status), επαναληψιμότητα και εύκολες εξαγωγές JSON, ώστε να τροφοδοτείται απευθείας οποιαδήποτε εφαρμογή ανάλυσης/dashboards. (Σχήμα 6.3)

6.6 Public Endpoints



Σχήμα 6.4: Αρχική οθόνη σύνδεσης

- Αρχική/φόρμα σύνδεσης: 'GET' '/'
Προβάλλεται η φόρμα εισόδου. Δεν απαιτείται προηγούμενη ταυτοποίηση (Σχήμα 6.4).
- Υποβολή σύνδεσης: 'POST' '/login'
Ελέγχονται τα διαπιστευτήρια και, σε επιτυχία, εκδίδεται JWT που αποθηκεύεται ως cookie 'access_token', με ανακατεύθυνση στο "/>

6.7 Endpoints με χρήση JWT

- Πίνακας ελέγχου: 'GET' '/dashboard'
Σελίδα επισκόπησης με βασικές πληροφορίες (ενεργά μέλη, εκκρεμείς αλλαγές) και πρόσφατα logs του scraping. Χρησιμοποιεί ως κεντρικό σημείο πλοήγησης μέσω του Navigation Bar μετά

τη σύνδεση. Οι διαθέσιμες ενέργειες που προβάλλονται στον χρήστη εξαρτώνται από τον ρόλο που έχει (Admin/Approver/Viewer). (Σχήμα 6.1)

- **Λίστα ακαδημαϊκού προσωπικού:** 'GET /faculty'
Κύρια σελίδα αναζήτησης ακαδημαϊκού προσωπικού με δυνατότητες φιλτραρίσματος σε πεδία όπως (όνομα, πανεπιστήμιο, τμήμα, κατηγορία βαθμίδας, τύπος προσωπικού) με server-side pagination. Οι εγγραφές εμπλουτίζονται με ενδείξεις/μετρικές Google Scholar όπου είναι διαθέσιμες.
- **Εκκρεμείς αλλαγές:** 'GET /pending-changes'
Προβάλλονται οι προτεινόμενες αλλαγές (προσθήκες, ενημερώσεις βαθμίδας, αποχωρήσεις) προς έλεγχο. Η προβολή της σελίδας δεν είναι διαθέσιμη για τους "Viewer". οι ενέργειες αποδοχής/απόρριψης πραγματοποιούνται σε ξεχωριστά προστατευμένα API (για ρόλους approver/admin).
- **Στατιστικά Google Scholar:** 'GET /google-scholar-stats'
Σελίδα αναζήτησης ακαδημαϊκού προσωπικού με έμφαση στο Google Scholar. Προσφέρει φιλτράρισμα ανά πανεπιστήμιο/τμήμα και προβολή συγκεντρωτικών αποδόσεων και μετρήσεων. (π.χ. Ποσοστό ταυτοποιημένων καθηγητών με GSID, h-index, citations, publications). Υποστηρίζεται και pagination μέσω του Server όπως στη Λίστα ακαδημαϊκού προσωπικού.
- **Αναφορές:** 'GET /reports'
Η σελίδα συγκεντρώνει σε ένα σημείο τις βασικές μετρήσεις όπως συνολικά ενεργά μέλη, ποσοστό κάλυψης προσωπικού με GSID, καθώς και μέσους δείκτες ανά ίδρυμα/τμήμα (h-index, citations, publications). Προβάλλει επίσης λίστες κορυφαίων μελών βάσει h-index και απλές κατανομές τιμών (π.χ. εύρη h-index/αναφορών/δημοσιεύσεων) με γραφήματα. Υποστηρίζονται βασικά φίλτρα (ίδρυμα, τμήμα, ύπαρξη προφίλ) και η εικόνα ενημερώνεται καθώς προστίθενται ή διορθώνονται GSID, ώστε να φαίνεται άμεσα η πρόοδος της κάλυψης και οι διαφορές ανά μονάδα.
- **Διαχείριση χρηστών:** 'GET /users'
Διαχειριστική σελίδα αποκλειστικά για ρόλο admin (επισκόπηση, δημιουργία, ενημέρωση, απενεργοποίηση χρηστών). Οι ενέργειες υποβάλλονται σε αντίστοιχα προστατευμένα endpoints.
- **Ενημέρωση ακαδημαϊκού προσωπικού:** 'POST /faculty/{faculty_id}/update'
Η υποβολή ενημέρωσης εγγραφής (π.χ. προσθήκη ή διόρθωση GSID) πραγματοποιείται μέσω POST από τις φόρμες των σελίδων /faculty και /google-scholar-stats. Με την επιτυχή αποθήκευση GSID, εκτελούνται αυτόματα δύο βοηθητικά scripts, ελαφρά προσαρμοσμένα για τη web εφαρμογή: publications.py (καταμέτρηση δημοσιεύσεων ανά έτος και σύνολα/πενταετίας) και citations.py (χρονική κατανομή αναφορών, συνολικά citations, h-index και h-index πενταετίας). Τα scripts αντλούν δεδομένα από το Google Scholar και εκτελούνται ανα GSID για να ενημερώνουν τους σχετικούς πίνακες (π.χ. dep, publications, citations), ώστε οι δείκτες να είναι άμεσα διαθέσιμοι στις σελίδες προβολής. Η ίδια ροή ισχύει ανεξάρτητα από το σημείο εκκίνησης (faculty ή google-scholar-stats), ενώ μετά την ολοκλήρωση γίνεται ανακατεύθυνση πίσω στη σελίδα προέλευσης με το ενημερωμένο περιεχόμενο.

6.8 Σελίδα Faculty (κατάλογος, αναζήτηση/φίλτρα/ταξινόμηση, σελιδοποίηση)

Faculty Members A admin

Search

University
All Universities

Department
All Departments

Rank Category
All Categories

Staff Type
All Types

Showing 20 of 1196 faculty members X Clear Filters

Name	Rank	University	Department	Staff Type	Google Scholar	Actions
Achlioptas Dimitris	Professor	Di Uoa	DI	faculty	Profile Found h: 49 C: 11205 P: 124	
Adam Maria	Professor	Dib Uth	DIB	faculty	Profile Found h: 11 C: 426 P: 65	
Adamantia Pateli	Professor	Di Ionio	Department	faculty	Profile Found h: 24 C: 3147 P: 79	
Adamidis Panagiotis	Professor	Iee Ihu	Staff Category	Faculty	Profile Found h: 9 C: 357 P: 37	
Adonis Bogris	Professor	Ice Uniwa	Staff 2	faculty	Profile Found h: 34 C: 3875 P: 213	
Agelidis Vassilios	Professor	Ee Duth	The Department	Faculty	Profile Found h: 92 C: 31790 P: 556	
Aggeliki Tsohou	Associate Professor	Di Ionio	Department	faculty	Profile Found h: 17 C: 1473 P: 72	
Aggelos Bletsas	Professor	Ece Tuc	School	faculty	Profile Found h: 34 C: 8718 P: 176	
Ahmed Mahdi	MSc Instructor	E-Ce Uth	Department	faculty	No profile found	

Σχήμα 6.5: Οθόνη διαχείρισης προσωπικού

Η /faculty λειτουργεί ως κατάλογος περιήγησης στο ακαδημαϊκό προσωπικό με server-side σελιδοποίηση, φιλτράρισμα και επιτρέποντας γρήγορη αναζήτηση. Τα φίλτρα καλύπτουν όνομα, ίδρυμα, τμήμα, κατηγορία βαθμίδας και τύπο προσωπικού, ενώ τα αποτελέσματα αποδίδονται αποδοτικά από τον διακομιστή. Για κάθε μέλος παρέχεται “quick view” με συνοπτικά στοιχεία και ένδειξη ύπαρξης προφίλ Google Scholar: όπου υπάρχει GSID, εμφανίζονται άμεσα βασικοί δείκτες που έχουν συγχρονιστεί με τη σχετική βάση. Επιπλέον, για μη αντιστοιχισμένες εγγραφές παρέχεται γρήγορη αναζήτηση στο Google Scholar με προ-συμπληρωμένα κριτήρια (όνομα + ίδρυμα) μέσω συνδέσμου, ώστε να εντοπίζεται εύκολα το σωστό προφίλ. (Σχήμα 6.5)

Πέρα από την αναζήτηση, η σελίδα υποστηρίζει γρήγορη ενημέρωση του προφίλ Google Scholar: ο χρήστης μπορεί να εισαγάγει ή να διορθώσει GSID και να υποβάλει την αλλαγή μέσω του προστατευμένου POST /faculty/{faculty_id}/update. Αν δοθεί έγκυρο GSID, ενεργοποιείται on-demand εμπλουτισμός μέσω των βοηθητικών scripts (publications.py, citations.py), τα οποία ανακτούν βιβλιομετρικά από το Google Scholar και ενημερώνουν τους αντίστοιχους πίνακες. Μετά την ολοκλήρωση, η σελίδα ανανεώνεται με τα ενημερωμένα στοιχεία, επιτρέποντας γρήγορο κύκλο εύρεσης μη αντιστοιχισμένων μελών, απόδοσης GSID και άμεσης απεικόνισης δεικτών.

6.9 Σελίδα Pending Changes (έλεγχος και έγκριση αλλαγών)

Pending Changes admin ▾

Pending Changes Review
Review and approve changes to faculty data Delete All (Debug)

Summary:

- Total Pending:** 7
- New Staff:** 0
- Rank Updates:** 7
- Staff Removals:** 0

Changes Awaiting Review

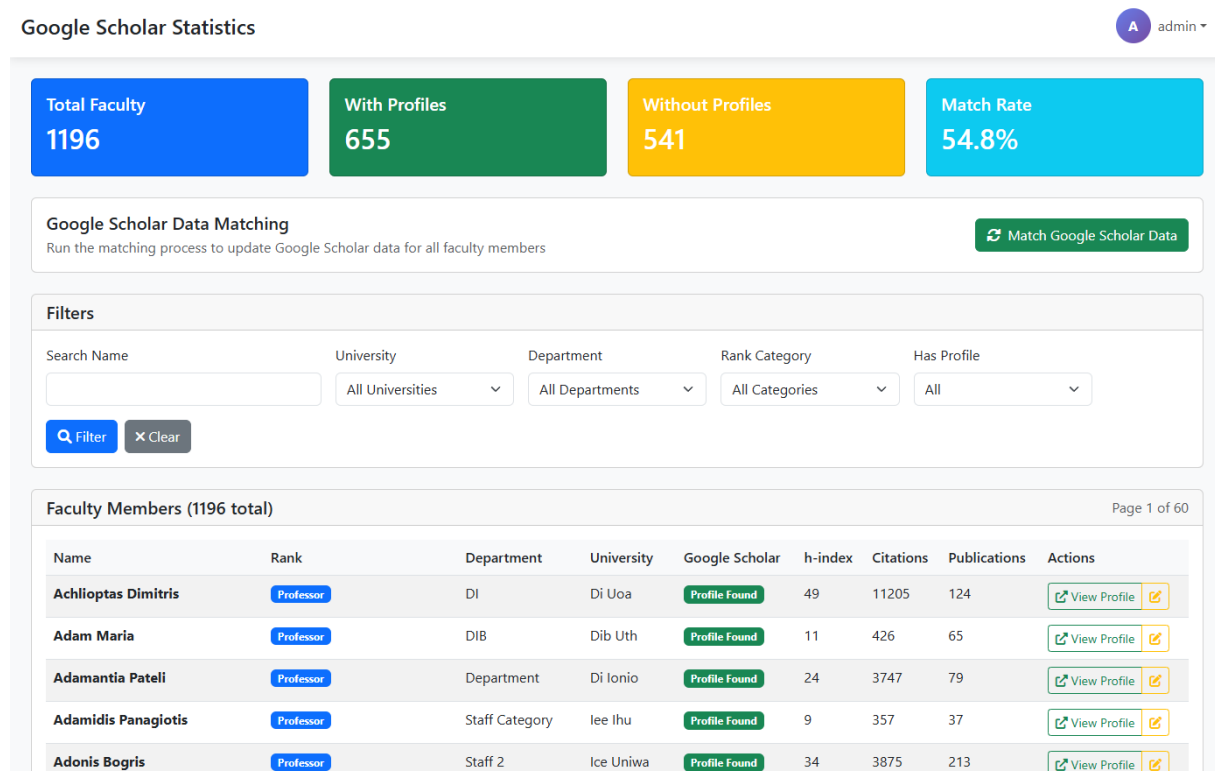
Change Type	Faculty Name	Department	Current Rank	Proposed Rank	Proposed Date	Actions
Rank Update	Sakkopoulos Evangelos	Cs Unipi → Didaskontes	Unknown	Associate Professor	2025-07-30 01:18	Accept / Decline
Rank Update	Pikrakis Aggelos	Cs Unipi → Didaskontes	Unknown	Associate Professor	2025-07-30 00:58	Accept / Decline
Rank Update	Venetis Ioannis	Cs Unipi → Didaskontes	Unknown	Assistant Professor	2025-07-30 00:52	Accept / Decline

Σχήμα 6.6: Οθόνη έγκρισης/απόρριψης προτεινόμενων αλλαγών

Η σελίδα «Εκκρεμείς αλλαγές» παρουσιάζει τις προτάσεις ενημέρωσης που προκύπτουν από τη διαδικασία σύγκρισης των τρεχόντων δεδομένων με τα αποτελέσματα νέας σάρωσης (μέσω POST /api/update-data) και εκτελώντας τον scraper για να διαβάσει το καινούργιο JSON αρχείο. Για κάθε αλλαγή αποθηκεύονται το είδος (νέα εγγραφή προσωπικού, ενημέρωση βαθμίδας/τύπου, αποχώρηση), τα πεδία που εμπλέκονται και ο προτεινών χρήστης, ώστε να υπάρχει πλήρες ιστορικό. (Σχήμα 6.6)

Η έγκριση ή απόρριψη αλλαγών γίνεται μέσω προστατευμένων ενεργειών: POST /api/pending-changes/{change_id}/accept και POST /api/pending-changes/{change_id}/decline, οι οποίες απαιτούν ρόλο approver ή admin. Σε αποδοχή, το σύστημα εφαρμόζει το κατάλληλο βήμα (π.χ. εισαγωγή νέου μέλους, ενημέρωση βαθμίδας, σήμανση ως ανενεργό) και ενημερώνει τα timestamps/μεταδεδομένα. Υπάρχει επίσης δυνατότητα εκκαθάρισης όλων των εκκρεμών με POST /api/pending-changes/delete-all (admin). Με τον τρόπο αυτό, ο κύκλος «λήψη → σύγκριση → επιθεώρηση → εφαρμογή» παραμένει ελεγχόμενος, επαναλήψιμος και συμβατός με τη διαβάθμιση ρόλων.

6.10 Σελίδα Google Scholar Stats (μετρικές και αντιστοιχίσεις)



Σχήμα 6.7: Οθόνη προβολής βιβλιομετρικών δεδομένων

Η σελίδα εστιάζει αποκλειστικά στα bibliometric δεδομένα του ακαδημαϊκού προσωπικού. Οπτικά μοιάζει με τη λίστα προσωπικού, αλλά το περιεχόμενο και τα φίλτρα είναι προσανατολισμένα στους δείκτες: για κάθε μέλος προβάλλονται (όπου υπάρχει GSID) h-index, συνολικές αναφορές και δημοσιεύσεις, καθώς και οι αντίστοιχες μετρήσεις πενταετίας. Υποστηρίζονται φίλτρα ανά ίδρυμα/τμήμα και φιλτράρισμα μη αντιστοιχισμένων εγγγραφών, ώστε να εντοπίζονται γρήγορα περιπτώσεις χωρίς προφίλ. Από τη σελίδα μπορεί επίσης να γίνει εισαγωγή/διόρθωση GSID· η υποβολή οδηγεί στο POST /faculty/{faculty_id}/update, το οποίο πυροδοτεί τις ρουτίνες άντλησης (publications.py, citations.py) για άμεσο συγχρονισμό των δεικτών. (Σχήμα 6.7)

Ο αρχικός μαζικός εμπλουτισμός βασίστηκε σε εσωτερικό endpoint αντιστοίχισης, το οποίο συνέδεσε τα μέλη της βάσης «faculty» με τη βάση «dep» (που διατηρείται από τον επιβλέποντα) όπου υπάρχουν αντιστοιχισμένα GSID. Η πρόκληση ήταν η ταύτιση ονομάτων: εφαρμόστηκαν κανόνες κανονικοποίησης (αφαίρεση τίτλων, πεζοποίηση, καθαρισμός σημείων στίξης) και μετρικές ομοιότητας (sequence matching, word overlap, Jaccard), με βαθμολογική ενίσχυση όταν συνέπιπτε ο κωδικός τμήματος/ιδρύματος. Όταν το σκορ υπερέβαινε το κατώφλι, γινόταν αυτόματη συσχέτιση· διαφορετικά προσφερόταν σύνδεσμος αναζήτησης στο Google Scholar. Έτσι, η σελίδα λειτουργεί τόσο ως «πίνακας δεικτών» όσο και ως πρακτικό εργαλείο ολοκλήρωσης/διόρθωσης αντιστοιχίσεων, με άμεση αντανάκλαση των αποτελεσμάτων στην απεικόνιση.

6.10.1 Ομοιότητα χαρακτήρων (SequenceMatcher)

Μετά την κανονικοποίηση ονομάτων, υπολογίζεται ομοιότητα σε επίπεδο χαρακτήρων με το 'SequenceMatcher().ratio()' της Python, που επιστρέφει τιμή στο [0,1] με βάση τα μεγαλύτερα συνεχόμενα τμήματα σύμπτωσης και κατάλληλες ευρετικές για «θόρυβο». Είναι ανθεκτική σε μικρές παραλλαγές (τυπογραφικά, διαστήματα, σημεία στίξης) και λειτουργεί συμπληρωματικά προς τις λεξικές μετρικές.

Στην εφαρμογή μας το SequenceMatcher είναι ένας από τους τρεις πυλώνες της τελικής βαθμολογίας ομοιότητας: υπολογίζεται ο χαρακτήρων-επί-χαρακτήρα δείκτης, έπειτα οι λεξικές μετρικές (Jaccard και word overlap) και τέλος λαμβάνουμε τον μέσο όρο τους. Η ενσωμάτωση χαρακτήρων-επί-χαρακτήρα με λεξικές μετρικές μειώνει τις «παγίδες» της κάθε προσέγγισης μεμονωμένα (π.χ., υψηλή Jaccard μόνο λόγω κοινού επωνύμου). Η τελική βαθμολογία χρησιμοποιείται στο ταίριασμα με Google Scholar μαζί με ένα προαιρετικό «μπόνους τμήματος» και ένα κατώφλι αποδοχής.

6.10.2 Δείκτης Jaccard

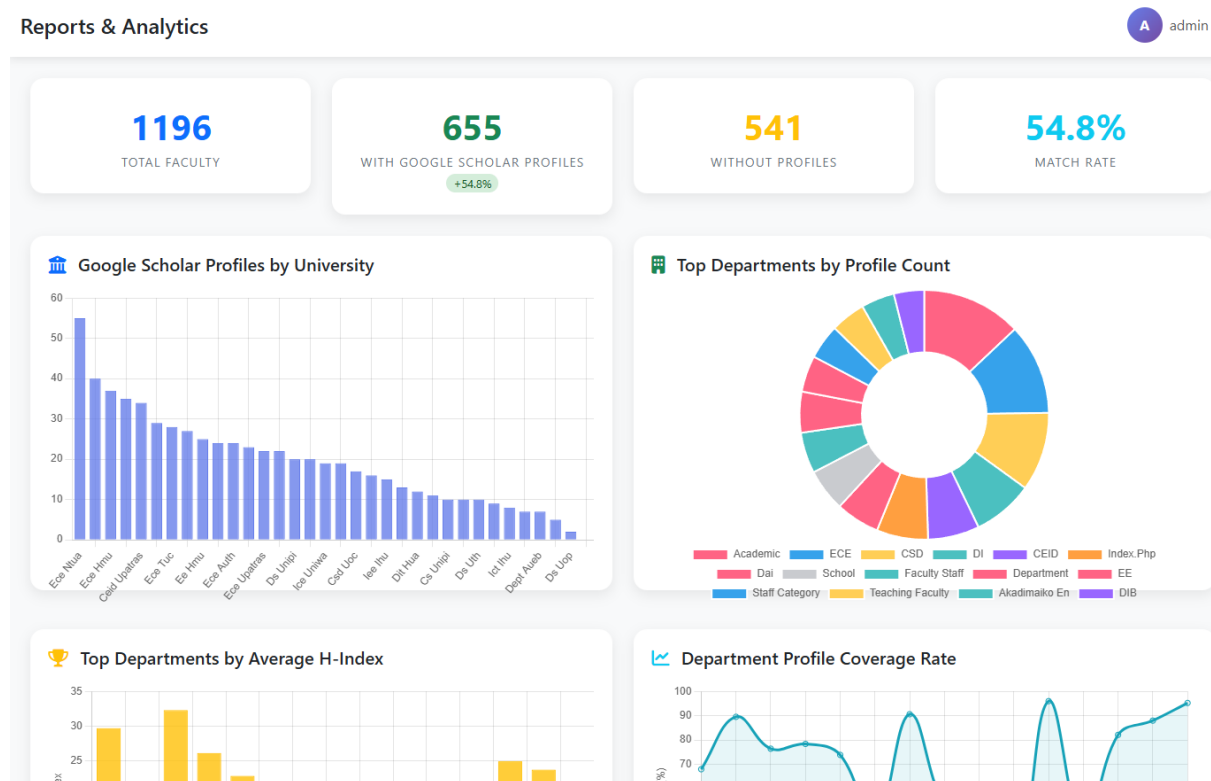
Τα ονόματα τεμαχίζονται σε σύνολα λέξεων και υπολογίζεται Jaccard: $|A \cap B| / |A \cup B|$. Η μετρική είναι αδιάφορη στη σειρά λέξεων, αποτυπώνει καθαρά την κοινή λεξική «ύλη» και λειτουργεί καλά όταν υπάρχουν συνθέσεις/μεταθέσεις (π.χ. μεσαία ονόματα, διάφορες γραφές), όπως περιγράφεται στη βιβλιοθηκονομία πληροφορίας και την ανάκτηση κειμένου.

6.10.3 Κανονικοποιημένη επικάλυψη λέξεων

Υπολογίζεται και λόγος επικάλυψης ως $|A \cap B| / \max(|A|, |B|)$, δηλαδή μια αυστηρότερη παραλλαγή της «overlap coefficient» (η κλασική εκδοχή διαιρεί με min), που τιμωρεί περισσότερο «περιττά» ή «λείποντα» tokens. Έτσι, αν ένα όνομα περιέχει έξτρα λέξεις, η βαθμολογία μειώνεται, βοηθώντας στη διάκριση οιονεί ταυτίσεων από πραγματικές αντιστοιχίσεις.

Το τελικό σκορ είναι ο μέσος όρος των τριών μετρικών και ενισχύεται με +0.3 όταν ταυτίζεται κωδικός τμήματος/ιδρύματος· η αντιστοίχιση γίνεται αποδεκτή αν total_score > 0.7.

6.11 Σελίδα Reports (αναφορές και εξαγωγές δεδομένων)



Σχήμα 6.8: Οθόνη προβολής αναφορών και αναλύσεων

Η σελίδα «Αναφορές» συγκεντρώνει και οπτικοποιεί τους βιβλιομετρικούς δείκτες, προσφέροντας κατατάξεις και συγκρίσεις ανά ίδρυμα και τμήμα. Παρουσιάζονται συνοπτικά η κάλυψη προφίλ (ποσοστό με GSID), καθώς και μέσοι δείκτες όπως h-index, citations και publications ανά ομάδα, μαζί με λίστες κορυφαίων μελών βάσει h-index. Οι κατανομές τιμών (π.χ. εύρη h-index/αναφορών/δημοσιεύσεων) απεικονίζονται σε γραφήματα για τον εντοπισμό τάσεων και διαφοροποιήσεων, ενώ τα φίλτρα (ίδρυμα/τμήμα) επιτρέπουν στοχευμένη ανάλυση. Η σελίδα ανανεώνεται με κάθε νέα αντιστοίχιση GSID ή ενημέρωση δεδομένων, λειτουργώντας ως πίνακας παρακολούθησης της προόδου και ως εργαλείο αξιολόγησης της ερευνητικής εικόνας ανά ακαδημαϊκή μονάδα. (Σχήμα 6.8)

Κεφάλαιο 7ο: Αποτελέσματα και Συμπεράσματα

7.1 Μεθοδολογία αξιολόγησης και μετρικές

Για τους σκοπούς της πειραματικής μελέτης, ως «πρόβλεψη» ορίζεται κάθε αυτόματη πρόταση αλλαγής που παράγεται από το σύστημα (pending change) σε σχέση με τα υφιστάμενα δεδομένα της βάσης (π.χ. νέα πρόσληψη, αποχώρηση, αλλαγή βαθμίδας ή αντιστοίχιση GSID). Ως «αλήθεια» θεωρήθηκε η χειροκίνητη επιβεβαίωση από τον χρήστη, κατόπιν ελέγχου των επίσημων ιστοσελίδων και συναφών πηγών. Η αξιολόγηση γίνεται με τυποποιημένες μετρικές: precision (αναλογία ορθών προτάσεων επί των προτάσεων), recall (αναλογία ορθών προτάσεων επί των πραγματικών αλλαγών) και F1 (μέσος όρος Precision/Recall) [33][34]. Για έλεγχο της ορθότητας μπορεί, όπου χρειάζεται, να χρησιμοποιηθεί και μέτρο συμφωνίας μεταξύ αξιολογητών (π.χ. Cohen's) [35]. Στο σκέλος Google Scholar δεν υπολογίζονται μετρικές ομοιότητας, χρησιμοποιούμε δείκτη κάλυψης (Match Rate), δηλαδή το ποσοστό των μελών με επιβεβαιωμένο GSID επί του συνόλου του τμήματος/ιδρύματος, το οποίο παρουσιάζεται συγκεντρωτικά ανά μονάδα.

Ενδεικτικό δείγμα/πρότυπα δεδομένων αξιολόγησης

Πίνακας 7.1: Δείγμα για name matching (σύγκριση HTML snapshot vs JSON αρχείου)

Site	Name (raw)	Name (normalized)	HTML	JSON	Result
iee.ihu.gr	Adamidis Panagiotis	adamidis panagiotis	yes	yes	TP
iee.ihu.gr	Arvanitou Elvira-Maria	arvanitou elviramaria	yes	no	FN
csd.auth.gr	Constantine Kotropoulos	constantine kotropoulos	yes	yes	TP

Πίνακας 7.2: Δείγμα για role matching (σύγκριση HTML snapshot vs JSON αρχείου)

site	name (normalized)	html_rank	json_rank (pred)	True
csd.auth.gr	angelis eleftherios	Professor	Assistant Professor	no
csd.auth.gr	dimitris vrakas	Associate Professor	Assistant Professor	no
csd.auth.gr	constantine kotropoulos	Professor	Professor	yes

Πίνακας 7.3: Name matching (Summary A)

Ιστότοπος	GT	Parsed	TP	FP	FN	Precision	Recall	F1
iee.ihu.gr	29	28	28	0	1	1.000	0.966	0.982
csd.auth.gr	43	43	43	0	0	1.000	1.000	1.000
cs.duth.gr	10	10	10	0	0	1.000	1.000	1.000
cs.unipi.gr	26	26	26	0	0	1.000	1.000	1.000
csd.uoc.gr	22	22	22	0	0	1.000	1.000	1.000

Πίνακας 7.4: Rank matching στα Matched Names (Summary B)

Ιστότοπος	support	TP	FP	FN	Precision	Recall	F1
iee.ihu.gr	0	0	0	0	1.000	1.000	1.000
csd.auth.gr	32	30	2	2	0.938	0.938	0.938
cs.duth.gr	10	9	0	1	1.000	0.900	0.947
cs.unipi.gr	2	2	0	0	1.000	1.000	1.000
csd.uoc.gr	10	8	2	2	0.800	0.800	0.800

Σημείωση: Για τους Πίνακες (Πίνακας 7.3, 7.4), οι μετρικές name matching υπολογίζονται σε σύνολα ονομάτων (μετά από κανονικοποίηση), ενώ οι μετρικές rank matching μικρο-αθροίζονται μόνο πάνω στα ονόματα που ταυτοποιήθηκαν (matched names) και για τα οποία το HTML διαθέτει μη-άγνωστη (μη-Unkown) βαθμίδα. Ως support ορίζονται οι άγνωστες βαθμίδες.

7.2 Αποτελέσματα: ακρίβεια, αστοχίες και παραδείγματα

Με βάση τα συγκεντρωτικά ευρήματα των Πινάκων 7.3 (Name matching) και 7.4 (Rank matching), η εικόνα είναι καθαρή: το σύστημα αναγνωρίζει με συνέπεια «ποιοι είναι στη λίστα», ενώ η ορθότητα της βαθμίδας εξαρτάται από το αν η πληροφορία που υπάρχει στο αρχείο ρυθμίσεων (JSON) για την ανάκτηση της βαθμίδας.

- Name Matching (Πίνακας 7.3)
Επιτυγχάνεται F1=1.000 για csd.auth.gr, cs.duth.gr, cs.unipi.gr και csd.uoc.gr, με μοναδική οριακή περίπτωση το iee.ihu.gr (F1=0.982) λόγω ενός FN. Η επίδοση αυτή οφείλεται σε σταθερούς selectors και στη συστηματική κανονικοποίηση ονομάτων, που απορροφά διαφοροποιήσεις γραφής.

- Rank Matching (Πίνακας 7.4)
Εκεί όπου η βαθμίδα εκτίθεται ρητά ανά μέλος (csd.auth.gr, cs.duth.gr) καταγράφονται υψηλά σκορ (rF1≈0.938 και 0.947). Στο csd.uoc.gr, όπου η πληροφορία είναι μερικώς διαθέσιμη, η επίδοση είναι μέτρια (rF1=0.800). Για iee.ihu.gr η σελίδα-λίστα δεν δίνει per-member βαθμίδες (support=0, μη εφαρμόσιμο), ενώ στο cs.unipi.gr το support είναι ελάχιστο (2 περιπτώσεις) και ο διοικητικός τίτλος («Head of Department») δεν είναι ακαδημαϊκή βαθμίδα, με αποτέλεσμα χαμηλό rF1 στον συγκριτικό πίνακα. Σημειώνεται πάντως ότι, όπου υπήρχε γνωστή βαθμίδα και έγινε χαρτογράφηση διοικητικών τίτλων σε ακαδημαϊκές βαθμίδες και αποδιπλοποίηση, οι έλεγχοι επιβεβαίωσαν ορθότητα.

Τυπικές αστοχίες που μπορούν να επιλυθούν με διαφορετικό configuration:

- Διοικητικοί τίτλοι/επικεφαλίδες ενότητας που χρειάζονται χαρτογράφηση σε ακαδημαϊκή βαθμίδα ή/και σε 'staff_type'.
- Διπλοεγγραφές του ίδιου ονόματος με διαφορετικά labels, που απαιτούν κανόνες αποδιπλοποίησης.
- Σελίδες-λίστες χωρίς per-member βαθμίδα, άρα μικρό «support» και αδυναμία δίκαιης σύγκρισης.
- Εναλλακτικά λεξιλόγια που χρειάζονται διόρθωση/εμπλουτισμό σε 'rank_mapping' και 'rank_keywords' για την μείωση των αγνώστων βαθμίδων.

Μια πιο δίκαιη και κατανοητή μέτρηση μπορεί να βασιστεί σε «λίγα αλλά καλά» παραδείγματα εκεί όπου η πληροφορία υπάρχει σίγουρα. Για ιστότοπους που δεν δείχνουν τη βαθμίδα στη σελίδα λίστα, ο έλεγχος να γίνεται σε μερικές σελίδες προφίλ (detail pages) όπου η βαθμίδα φαίνεται ρητά. Πριν μετρήσουμε, κανονικοποιούμε τα δεδομένα: ενώνουμε διπλοεγγραφές του ίδιου ονόματος (με βάση το normalized name) και, όπου έχει νόημα, χαρτογραφούμε διοικητικούς τίτλους (π.χ. Head of Department) στην αντίστοιχη ακαδημαϊκή βαθμίδα (π.χ. Professor) ή σε τύπο προσωπικού. Παράλληλα δηλώνουμε το “support” (πόσες εγγραφές έχουν πραγματικά συγκρίσιμη βαθμίδα) και, αν χρειάζεται, δίνουμε και μέσους όρους ανά κλάση (macro-average) για να φανούν τυχόν συστηματικές αστοχίες σε συγκεκριμένες βαθμίδες. Σε μικρά δείγματα, είναι χρήσιμο να αναφέρουμε και ένα εύρος αβεβαιότητας (π.χ. μέσω απλής δειγματοληψίας), ώστε τα αποτελέσματα να μην παρερμηνεύονται.

Για να αποτυπώνεται η πραγματική αξία του συστήματος, πέρα από τις στατικές συγκρίσεις, προτείνεται αξιολόγηση «γεγονότων» σε βάθος χρόνου: μετράμε precision/recall/F1 πάνω στις pending changes που προτείνει το σύστημα και εγκρίνονται/απορρίπτονται από τον χρήστη. Στο Google Scholar, εκτός από τον δείκτη κάλυψης (Match Rate), ένα μικρό χειροκίνητα επισημειωμένο δείγμα αντιστοιχίσεων αρκεί για να ελέγξουμε την ποιότητα, και όπου απαιτείται μπορούμε να εξετάσουμε πώς αλλάζουν τα αποτελέσματα αν αυξομειώσουμε το όριο αποδοχής (threshold). Με αυτά τα βήματα, η μέτρηση γίνεται πιο εύληπτη, δίκαιη και κοντά στη ρεαλιστική χρήση του συστήματος.

Εν κατακλείδι, τα αποτελέσματα δείχνουν ότι το σύστημα είναι ώριμο και αξιόπιστο στην ταυτοποίηση ονομάτων, ενώ η ποιότητα στις βαθμίδες είναι υψηλή όταν η πληροφορία παρέχεται ρητά από τον ιστότοπο. Με στοχευμένες παραμετρικές βελτιώσεις (selectors/keywords), αποδιπλοποίηση και εστιασμένη χρήση σε detail pages (containers), η εκτίμηση ακρίβειας γίνεται δικαιότερη και πιο αντιπροσωπευτική των πραγματικών δυνατοτήτων του συστήματος.

7.3 Συμπεράσματα

Ο στόχος για παραμετρικό scraper χωρίς ειδικό κώδικα ανά ιστότοπο επιτεύχθηκε σε μεγάλο βαθμό: οι περισσότεροι ιστότοποι καλύφθηκαν από τους τέσσερις τύπους αναλυτών (generic, container, table, sequential) με ελάχιστες στοχευμένες ρυθμίσεις. Η ροή «λήψη → DOM parsing → κανονικοποίηση → προτάσεις αλλαγών» αποδείχθηκε πρακτική, αναπαραγώγιμη και φιλική στη συντήρηση. Στην πράξη, η ποιότητα των προτάσεων εξαρτάται σε μεγάλο βαθμό από τη σταθερότητα της δομής των ιστοσελίδων και τη συνέπεια ονοματολογίας/βαθμίδων. Η σύζευξη κανόνων κανονικοποίησης με ευέλικτα λεξιλόγια και η προσεκτική ρύθμιση κατωφλίων ομοιότητας βελτίωσαν μετρήσιμα την απόδοση.

Σε σχέση με την αρχική προσδοκία «μία ρύθμιση για όλους», προέκυψε ότι λίγες, σαφείς διαφοροποιήσεις ανά τμήμα (selectors, λέξεις-κλειδιά) επαρκούν για υψηλή ακρίβεια, ενώ η προσθήκη μηχανισμών επαλήθευσης (manual review, pending changes) λειτουργεί συμπληρωματικά εξασφαλίζοντας ποιότητα δεδομένων.

7.4 Μελλοντικές επεκτάσεις

Ως «data layer», το αποθετήριο (σταθερό σχήμα, GSID, JSON έξοδοι) μπορεί να τροφοδοτεί άμεσα πλατφόρμες παρουσίασης/ανάλυσης και δημόσια APIs, διατηρώντας ευθυγραμμισμένο το μητρώο μελών με την τρέχουσα κατάσταση. Κεντρικός άξονας επέκτασης είναι η κάλυψη επιπλέον τμημάτων πέρα από τα προρυθμισμένα της Πληροφορικής (άλλες σχολές/τμήματα), μέσω παραμετρικών ρυθμίσεων και νέων κανόνων εξαγωγής.

Περαιτέρω, η ενσωμάτωση τεχνικών NLP μπορεί να βελτιώσει σημαντικά την ακρίβεια και την ανθεκτικότητα: αναγνώριση οντοτήτων (ονόματα προσώπων) σε ελληνικά/αγγλικά, ομογενοποίηση μεταγραφών/διακριτικών, προηγμένη αποδιπλοποίηση με συνδυασμό fuzzy matching, ανίχνευση μεταβολών από ανακοινώσεις/δελτία τύπου, καθώς και ενεργή μάθηση από τις διορθώσεις των χρηστών για συνεχή βελτίωση των ορίων αποδοχής.

7.5 Επίλογος

Το σύστημα κατέδειξε ότι η παραμετρική προσέγγιση επιτρέπει αξιόπιστη, επεκτάσιμη και οικονομική σε πόρους αυτοματοποίηση της παρακολούθησης αλλαγών ακαδημαϊκού προσωπικού. Η πειραματική αξιολόγηση ανέδειξε υψηλή ακρίβεια στις τυπικές περιπτώσεις και σαφείς άξονες βελτίωσης για αμφίσημα σενάρια (διγλωσσία, συνωνυμίες, δυναμικό DOM). Η συνδυαστική χρήση κανόνων κανονικοποίησης, μετρικών ομοιότητας και ανθρώπινου ελέγχου οδηγεί σε δεδομένα περισσότερο συνεπή και επικαιροποιημένα, δημιουργώντας μια βάση για αναλύσεις, συγκρίσεις με ορθά και κανονικοποιημένα δεδομένα.

Η συνολική πορεία του έργου κατέδειξε ότι η παραμετρική προσέγγιση επιτρέπει αξιόπιστη, επεκτάσιμη και οικονομική σε πόρους αυτοματοποίηση της παρακολούθησης αλλαγών ακαδημαϊκού προσωπικού. Η πειραματική αξιολόγηση επιβεβαίωσε υψηλή ακρίβεια στις τυπικές περιπτώσεις και ανέδειξε σαφείς άξονες βελτίωσης για αμφίσημα σενάρια (διγλωσσία/μεταγραφές, συνωνυμίες, δυναμικό DOM). Ο συνδυασμός κανονικοποίησης, μετρικών ομοιότητας και ανθρώπινου ελέγχου παρήγαγε συνεπέστερα και πιο επικαιροποιημένα δεδομένα, μειώνοντας σφάλματα και καθυστερήσεις που είναι εγγενείς στη χειροκίνητη διαδικασία.

Σε πρακτικό επίπεδο, το αποτέλεσμα είναι ένα ενιαίο, κεντρικό αποθετήριο στοιχείων ακαδημαϊκού προσωπικού, ενοποιημένο από ετερογενείς πηγές και εμπλουτισμένο με προφίλ Google Scholar για

βασικούς βιβλιομετρικούς δείκτες. Η πλατφόρμα προσφέρει συγκεντρωτικές απεικονίσεις και συγκρίσεις ανά τμήμα/ίδρυμα, καθώς και δυνατότητα παρακολούθησης της πορείας στο χρόνο, λειτουργώντας ως κοινή βάση αναφοράς για ενημέρωση, ανάλυση και διαφάνεια. Με στοχευμένες επεκτάσεις (λεξικά/κανόνες, αξιοποίηση μεταδεδομένων όπου χρειάζεται, πρόσθετες πηγές ταυτοποίησης), το σύστημα μπορεί να υποστηρίξει ακόμη πλουσιότερες αναφορές και πιο πλήρη κάλυψη του οικοσυστήματος.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1]: M. Koster, Robots Exclusion Protocol (2022). <https://www.rfc-editor.org/rfc/rfc9309>
- [2]: Microservices.io (2024). What are Microservices? <https://microservices.io/>
- [3]: Red Hat (2024). What are APIs? <https://www.redhat.com/en/topics/api/what-are-application-programming-interfaces>
- [4]: REST API Tutorial. (2024). What is REST? <https://restfulapi.net/>
- [5]: Database Normalization. (2024). Database Normalization Explained. <https://www.guru99.com/database-normalization.html>
- [6]: Redis. (2024). What is Caching? <https://redis.io/topics/caching>
- [7]: Python Software Foundation. (2024). Python Programming Language. <https://www.python.org/>
- [8]: FastAPI. (2024). FastAPI - Modern, fast (high-performance), web framework for building APIs with Python. <https://fastapi.tiangolo.com/>
- [9]: BeautifulSoup. (2024). BeautifulSoup Documentation. <https://www.crummy.com/software/BeautifulSoup/>
- [10]: Real Python. (2024). Object-Oriented Programming (OOP) in Python 3. <https://realpython.com/python3-object-oriented-programming/>
- [11]: Starlette. (2024). Starlette - ASGI toolkit. <https://www.starlette.io/>
- [12]: Pydantic. (2024). Pydantic - Data validation using Python type annotations. <https://pydantic-docs.helpmanual.io/>
- [13]: BeautifulSoup Documentation. (2024). Installing and using BeautifulSoup. <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

- [14]: Bootstrap. (2024). Bootstrap - The most popular HTML, CSS, and JS library in the world. <https://getbootstrap.com/>
- [15]: Bootstrap Documentation. (2024). Bootstrap Components and Utilities. <https://getbootstrap.com/docs/>
- [16]: MySQL. (2024). MySQL - The world's most popular open source database. <https://www.mysql.com/>
- [17]: MySQL Documentation. (2024). MySQL Reference Manual. <https://dev.mysql.com/doc/>
- [18]: Jinja2. (2024). Jinja2 - A fast and feature rich templating engine for Python. <https://jinja.palletsprojects.com/>
- [19]: Jinja2 Documentation. (2024). Jinja2 Template Designer Documentation. <https://jinja.palletsprojects.com/en/3.1.x/templates/>
- [20]: MariaDB Foundation (2024). MariaDB. <https://mariadb.org/>
- [21]: MariaDB Knowledge Base (2024). Documentation. <https://mariadb.com/kb/en/>
- [22]: Olston, C., & Najork, M. (2010). Web Crawling. Foundations and Trends in Information Retrieval, 4(3), 175–246.
- [23]: Fielding, R., Nottingham, M. (Eds.). HTTP Caching (RFC 9111). IETF, 2022. <https://www.rfc-editor.org/rfc/rfc9111>
- [24]: MDN Web Docs. HTTP caching. <https://developer.mozilla.org/en-US/docs/Web/HTTP/Caching>
- [25]: Fielding, R., Nottingham, M. (Eds.). HTTP Semantics (RFC 9110). IETF, 2022. <https://www.rfc-editor.org/rfc/rfc9110>
- [26]: Jones, M., Bradley, J., Sakimura, N. (2015). JSON Web Token (JWT). RFC 7519. <https://www.rfc-editor.org/rfc/rfc7519>
- [27]: Jones, M., Hardt, D. (2012). The OAuth 2.0 Authorization Framework: Bearer Token Usage. RFC 6750. <https://www.rfc-editor.org/rfc/rfc6750>
- [28]: OWASP (2023). JSON Web Token (JWT) Cheat Sheet. https://cheatsheetseries.owasp.org/cheatsheets/JSON_Web_Token_for_Java_Cheat_Sheet.html
- [29]: Python Software Foundation (2024). “difflib — Helpers for computing deltas.” <https://docs.python.org/3/library/difflib.html>
- [30]: Manning, C. D., Raghavan, P., Schütze, H. (2008). “Introduction to Information Retrieval.” Cambridge University Press.
- [31]: “Jaccard index.”. Wikipedia. https://en.wikipedia.org/wiki/Jaccard_index
- [32]: “Overlap coefficient (Szymkiewicz–Simpson)”. Wikipedia. https://en.wikipedia.org/wiki/Overlap_coefficient

- [33]: Manning, C. D., Raghavan, P., Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press
- [34]: Sokolova, M., Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management
- [35]: Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement
- [36]: WHATWG, HTML Standard, Link Types
<https://html.spec.whatwg.org/multipage/links.html#linkTypes>
- [37]: Chakrabarti, S. et al. “Focused Crawling: a new approach to topic-specific Web resource discovery.”
- [38]: Selenium WebDriver Documentation <https://selenium.dev>
- [39]: Playwright Documentation <https://playwright.dev>
- [40]: N.4009/2011 ΦΕΚ 195/6-9-2011
- [41]: N.4957/2022 ΦΕΚ 141/21-7-2022

ΠΑΡΑΡΤΗΜΑ ΚΩΔΙΚΑ

Κωδικός Web Scraper:

```
@dataclass
class FacultyWebsiteConfig:
    urls: List[str]
    name_selector: str
    container_class: Optional[str] = None
    staff_type: str = 'faculty'
    debug_print: bool = False
    rank_selector: Optional[str] = None
    is_table_data: bool = False
    rank_keywords: Optional[List[str]] = None
    name_container: Optional[str] = None
    info_container: Optional[str] = None
    search_parent_siblings_for_rank: bool = False
    enabled: bool = True
    use_cached: bool = True
    rank_mapping: Optional[Dict[str, str]] = None
    is_container_based_rank: bool = False
    rank_is_parent: bool = False
    table_class: Optional[str] = None
    rank_from_select: bool = False
    select_value_attr: Optional[str] = None
    select_text_attr: Optional[str] = None
    select_all_value: Optional[str] = None
    select_element: Optional[Dict[str, str]] = None
    max_siblings_to_check: int = 3
    main_content_selector: Optional[str] = None
    use_parent_rank_with_distance: bool = False
    rank_container: Optional[Dict[str, str]] = None
    rank_element: Optional[Dict[str, str]] = None
    use_sibling_rank: bool = False
    rank_sibling_selector: Optional[str] = None
    rank_navigation_steps: Optional[List[Dict[str, str]]] = None
    concat_name_elements: bool = False
    name_concat_selectors: Optional[List[str]] = None
    parser_type: str = 'generic'
    name_split_separator: Optional[str] = None
    fixed_rank: Optional[str] = None
    name_rank_split_separator: Optional[str] = None
```

```

class RobotsTxtHandler:

    def __init__(self):
        self.parser = urllib.robotparser.RobotFileParser()
        self.cache = {}
        self.user_agent = 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36'

    def can_fetch(self, url: str) -> bool:

        parsed_url = urlparse(url)
        base_url = f"{parsed_url.scheme}://{parsed_url.netloc}"

        if base_url in self.cache:
            return self.cache[base_url].can_fetch(self.user_agent, url)

        try:
            robots_url = urljoin(base_url, '/robots.txt')
            self.parser.set_url(robots_url)
            self.parser.read()
            self.cache[base_url] = self.parser
            return self.parser.can_fetch(self.user_agent, url)
        except Exception as e:
            print(f"Error reading robots.txt for {base_url}: {e}")
            return True

    def get_crawl_delay(self, url: str) -> float:

        parsed_url = urlparse(url)
        base_url = f"{parsed_url.scheme}://{parsed_url.netloc}"

        if base_url in self.cache:
            delay = self.cache[base_url].crawl_delay(self.user_agent)
            return float(delay) if delay is not None else 1.0

        try:
            robots_url = urljoin(base_url, '/robots.txt')
            self.parser.set_url(robots_url)

```

```
self.parser.read()
self.cache[base_url] = self.parser
delay = self.parser.crawl_delay(self.user_agent)
return float(delay) if delay is not None else 1.0
except Exception as e:
    print(f"Error reading robots.txt for {base_url}: {e}")
    return 1.0
```

```

def main():

    os.makedirs('faculty_data', exist_ok=True)
    os.makedirs('cached_html', exist_ok=True)

    configs = [
        FacultyWebsiteConfig(
            urls=['https://www.csd.auth.gr/en/people/faculty/'],
            name_selector='div.elementor-widget-theme-post-title p.elementor-heading-title a',
            container_class='e-loop-item',
            staff_type='faculty',
            debug_print=True,
            use_cached=True,
            rank_selector='h3.elementor-heading-title span',
            rank_keywords=['Professor', 'Associate Professor', 'Assistant Professor', 'Lecturer', 'Professor
Emeriti'],
            name_container='div.elementor-widget-container',
            info_container='div.elementor-widget-container',
            search_parent_siblings_for_rank=False,
            enabled=True,
            rank_mapping={
                'Professor': 'Professor',
                'Associate Professor': 'Associate Professor',
                'Assistant Professor': 'Assistant Professor',
                'Lecturer': 'Lecturer',
                'Professor Emeritus': 'Professor Emeritus'
            },
            is_container_based_rank=False,
            parser_type='generic'
        ),
        FacultyWebsiteConfig(

```

```

        urls=['https://www.iee.ihu.gr/en/staff_category/faculty-member/'],
        name_selector='td b a',
        container_class='table',
        staff_type='faculty',
        debug_print=True,
        is_table_data=True,
        rank_keywords=['Associate Professor', 'Assistant Professor', 'Lecturer', 'Professor'],
        search_parent_siblings_for_rank=False,
        enabled=True,
        use_cached=True,
        parser_type='table'
    ),
    FacultyWebsiteConfig(
        urls=['https://www.iee.ihu.gr/en/staff_category/edu-staff/'],
        name_selector='td b a',
        container_class='table',
        staff_type='lab_lecturer',
        debug_print=False,
        is_table_data=True,
        name_container='td',
        info_container='td',
        enabled=True,
        use_cached=True,
        parser_type='table'
    ),
    FacultyWebsiteConfig(
        urls=["https://www.cs.duth.gr/faculty.xhtml?language=en"],
        name_selector="h3.name-imagia",
        container_class="card-imagia",
        staff_type="faculty",
        debug_print=False,

```

```

is_table_data=False,
rank_selector="p.subtitle-imagia",
rank_keywords=['Associate Professor', 'Assistant Professor', 'Lecturer', 'Professor'],
search_parent_siblings_for_rank=True,
use_cached=True,
enabled=True
),
FacultyWebsiteConfig(
    urls=['https://www.ict.ihu.gr/en/staff_category/faculty_members/'],
    name_selector='td b a',
    container_class='table',
    staff_type='faculty',
    debug_print=False,
    is_table_data=True,
    rank_keywords=['Associate Professor', 'Assistant Professor', 'Lecturer', 'Professor'],
    search_parent_siblings_for_rank=False,
    enabled=True,
    use_cached=True,
    parser_type='table'
),
FacultyWebsiteConfig(
    urls=['https://www.ict.ihu.gr/en/staff_category/lab_staff/'],
    name_selector='td b a',
    container_class='table',
    staff_type='lab_lecturer',
    debug_print=False,
    is_table_data=True,
    name_container='td',
    info_container='td',
    enabled=True,
    use_cached=True,

```

```

        parser_type='table'
    ),
    FacultyWebsiteConfig(
        urls=['https://www.uom.gr/en/dai/academic-staff'],
        name_selector='td a.text-primary',
        container_class='table-custom',
        is_table_data=True,
        rank_selector='th',
        rank_keywords=['Professor', 'Associate Professor', 'Assistant Professor', 'Emeritus Professor',
        'Retired Professor', 'Former Professor', 'Honorary Doctorate'],
        rank_mapping={
            'Professors': 'Professor',
            'Associate Professors': 'Associate Professor',
            'Assistant Professors': 'Assistant Professor',
            'Emeritus Professors': 'Emeritus Professor',
            'Retired Professors': 'Retired Professor',
            'Former Professors': 'Former Professor',
            'Honorary Doctorate': 'Honorary Doctorate',
            'Dean of the School': 'Dean of the School'
        },
        debug_print=False,
        enabled=True,
        use_cached=True,
        parser_type='table',
        staff_type='faculty'
    ),
    FacultyWebsiteConfig(
        urls=['https://www.ceid.upatras.gr/en/staff/faculty/'],
        name_selector='h3.gdlr-core-personnel-list-title',
        container_class='gdlr-core-personnel-list',
        staff_type='faculty',
        debug_print=False,

```

```

rank_selector='div.gdlr-core-personnel-list-position',
rank_keywords=['Associate Professor', 'Assistant Professor', 'EDIP', 'Professor', 'Emeritus
Professor'],
search_parent_siblings_for_rank=True,
enabled=True,
use_cached=True,
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'EDIP': 'EDIP',
    'Emeritus Professor': 'Emeritus Professor'
},
name_container='div.gdlr-core-personnel-list-content-wrap',
info_container='div.gdlr-core-personnel-list-content-wrap'
),

```

```

FacultyWebsiteConfig(
    urls=['https://www.ece.upatras.gr/index.php/en/personnel.html'],
    name_selector='span.sppb-person-name a',
    container_class='sppb-addon-person',
    staff_type='faculty',
    debug_print=False,
    rank_selector='span.sppb-person-designation',
    rank_keywords=['Associate Professor', 'Assistant Professor', 'Professor'],
    name_container='div.sppb-person-information',
    info_container='div.sppb-person-information',
    search_parent_siblings_for_rank=False,
    enabled=True,
    use_cached=True,
    rank_mapping={
        'Professor': 'Professor',
        'Associate Professor': 'Associate Professor',

```

```

        'Assistant Professor': 'Assistant Professor'
    },
    parser_type='container',
    name_split_separator=' '
),
FacultyWebsiteConfig(
    urls=['https://www.ece.upatras.gr/index.php/en/laboratory-teaching-staff.html'],
    name_selector='span.sppb-person-name a',
    container_class='sppb-addon-person',
    staff_type='lab_lecturer',
    debug_print=True,
    rank_selector=None,
    rank_keywords=[],
    name_container='div.sppb-person-information',
    info_container='div.sppb-person-information',
    search_parent_siblings_for_rank=False,
    enabled=True,
    use_cached=True,
    rank_mapping={},
    parser_type='container',
    name_split_separator=' ',
    fixed_rank='Laboratory Teaching Staff'
),
FacultyWebsiteConfig(
    urls=['https://www.dept.aueb.gr/en/cs_depen'],
    name_selector='div.profs a',
    container_class='quicktabs-tabpage',
    staff_type='faculty',
    debug_print=False,
    rank_keywords=['Professor Emeritus', 'Associate Professor', 'Assistant Professor', 'Professor'],
    name_container='div.profs',

```

```

info_container='div.profs',
enabled=True,
use_cached=True,
rank_selector='div.quicktabs-tabpage',
search_parent_siblings_for_rank=False,
rank_mapping={
    'quicktabs-tabpage-view__json_cs_faculty__page_1-0': 'Professor',
    'quicktabs-tabpage-view__json_cs_faculty__page_1-1': 'Professor Emeritus',
    'quicktabs-tabpage-view__json_cs_faculty__page_1-2': 'Associate Professor',
    'quicktabs-tabpage-view__json_cs_faculty__page_1-3': 'Assistant Professor'
},
is_container_based_rank=True,
rank_is_parent=True,
rank_from_select=False,
parser_type='container'
),
FacultyWebsiteConfig(
    urls=['https://di.ionio.gr/en/department/staff/professors/'],
    name_selector='td.tt-big-full a',
    staff_type='faculty',
    debug_print=False,
    use_cached=True,
    enabled=True,
    is_table_data=True,
    rank_selector='p.staff-label',
    rank_keywords=['Professor', 'Associate Professor', 'Assistant Professor', 'Emeritus Professor'],
    rank_mapping={
        'Professors': 'Professor',
        'Associate Professors': 'Associate Professor',
        'Assistant Professors': 'Assistant Professor',
        'Professors Emeriti': 'Emeritus Professor'
    }
)

```

```

    },
    parser_type='table',
    use_parent_rank_with_distance=True,
    max_siblings_to_check=3,
    search_parent_siblings_for_rank=True,
    rank_is_parent=True,
    rank_navigation_steps=[
        {'action': 'parent'},
        {'action': 'parent'},
        {'action': 'parent'},
        {'action': 'prev_sibling'},
        {'action': 'find', 'selector': 'p.staff-label'}
    ]
),
FacultyWebsiteConfig(
    urls=['https://cs.unipi.gr/en/didaskontes/'],
    name_selector='div.wpb_text_column.color-blue.font-type-5.family-1.text-justify p a',
    container_class='profile-box',
    staff_type='faculty',
    debug_print=False,
    use_cached=True,
    rank_selector='div.profile-section-title',
    rank_keywords=['Professor', 'Associate Professor', 'Assistant Professor', 'Emeritus Professor',
'Head of Department', 'Lecturer'],
    search_parent_siblings_for_rank=False,
    enabled=True,
    rank_is_parent=False,
    is_container_based_rank=True,
    name_container='div.wpb_text_column.color-blue.font-type-5.family-1.text-justify',
    info_container='div.profile-text',
    main_content_selector='div.wpb_wrapper',
    rank_mapping={

```

```

        'Professors': 'Professor',
        'Associate Professors': 'Associate Professor',
        'Assistant Professors': 'Assistant Professor',
        'Emeritus Professors': 'Emeritus Professor',
        'Head of Department': 'Head of Department',
        'Lecturers': 'Lecturer',
        'Professor': 'Professor',
        'Associate Professor': 'Associate Professor',
        'Assistant Professor': 'Assistant Professor',
        'Emeritus Professor': 'Emeritus Professor',
        'Lecturer': 'Lecturer'
    },
    parser_type='container'
),
FacultyWebsiteConfig(
    urls=['https://www.ds.unipi.gr/en/faculty/'],
    name_selector='a.teacher-name th',
    container_class='tab-content',
    staff_type='faculty',
    debug_print=False,
    use_cached=True,
    is_table_data=False,
    rank_selector='h2',
    rank_keywords=[
        'Professor', 'Associate Professor', 'Assistant Professor',
        'Emeritus Professors', 'Retired Faculty'
    ],
    search_parent_siblings_for_rank=False,
    enabled=True,
    rank_is_parent=False,
    use_parent_rank_with_distance=False,

```

```

max_siblings_to_check=1,
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Emeritus Professors': 'Emeritus Professor',
    'Retired Faculty': 'Retired Professor'
},
main_content_selector=None,
is_container_based_rank=True,
rank_container={
    'tag': 'div',
    'class': 'tab-content'
},
rank_element={
    'tag': 'h2'
},
parser_type='container'
),
FacultyWebsiteConfig(

```

```

urls=['https://www.csd.uoc.gr/CSD/index.jsp?content=academic_staff&openmenu=demoAcc2&lang=en'],

```

```

    name_selector='.small_title',
    container_class='position-group',
    rank_container={'element': '.position-group', 'attr': 'group-name'},
    rank_mapping={
        'Professors': 'Professor',
        'Associate Professors': 'Associate Professor',
        'Assistant Professors': 'Assistant Professor'
    },
    parser_type='container',

```

```

        debug_print=False,
        enabled=True
    ),
    FacultyWebsiteConfig(
        urls=['https://www.cs.uoi.gr/staff/teaching-faculty/?cat=155&lang=en'],
        name_selector='h4 a',
        container_class='article.plainbackground',
        staff_type='faculty',
        debug_print=False,
        use_cached=True,
        rank_selector='h3',
        rank_keywords=[
            'Professor', 'Associate Professor', 'Assistant Professor'
        ],
        search_parent_siblings_for_rank=False,
        enabled=True,
        rank_mapping={
            'Professor': 'Professor',
            'Associate Professor': 'Associate Professor',
            'Assistant Professor': 'Assistant Professor',
            'Associate Professor': 'Associate Professor',
            'Assisant Professor': 'Assistant Professor'
        },
        parser_type='container'
    ),
    FacultyWebsiteConfig(
        urls=['https://www.dit.uoi.gr/staff/'],
        name_selector='td:nth-child(1) h4',
        debug_print=False,
        staff_type='faculty',
        use_cached=True,

```

```

enabled=True,
rank_navigation_steps=[
    {'action': 'parent'},
    {'action': 'parent'},
    {'action': 'next_sibling'}
],
rank_keywords=[
    'ΚΑΘΗΓΗΤΗΣ',
    'ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ',
    'ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ',
    'ΛΕΚΤΟΡΑΣ',
    'ΔΕΠ',
    'ΕΔΙΠ',
    'ΕΔ',
    'ΥΑΔΕ'
],
rank_mapping={

    'ΔΕΠΚΑΘΗΓΗΤΗΣ': 'Professor',
    'ΔΕΠΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ': 'Associate Professor',
    'ΔΕΠΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ': 'Assistant Professor',
    'ΔΕΠΛΕΚΤΟΡΑΣ': 'Lecturer',
    'ΔΕΠΚΑΘΗΓΗΤΗΣ-': 'Professor',
    'ΔΕΠΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ-': 'Associate Professor',
    'ΔΕΠΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ-': 'Assistant Professor',
    'ΔΕΠΛΕΚΤΟΡΑΣ-': 'Lecturer',
    'ΔΕΠΚΑΘΗΓΗΤΗΣΑ': 'Professor',
    'ΔΕΠΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣΑ': 'Associate Professor',
    'ΔΕΠΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣΑ': 'Assistant Professor',
    'ΔΕΠΛΕΚΤΟΡΑΣΑ': 'Lecturer',

```

'ΕΔΙΠ-A': 'Laboratory Teaching Staff',
'ΕΔΙΠ-': 'Laboratory Teaching Staff',
'ΕΔΙΠ--': 'Laboratory Teaching Staff',

'ΕΔ-': 'Contracted Teacher',
'ΕΔ': 'Contracted Teacher',
'ΕΔ--': 'Contracted Teacher',

'ΥΑΔΕ-': 'Teaching Experience Fellow',
'ΥΑΔΕ': 'Teaching Experience Fellow',
'ΥΑΔΕ--': 'Teaching Experience Fellow',

'ΚΑΘΗΓΗΤΗΣ': 'Professor',
'ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ': 'Associate Professor',
'ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ': 'Assistant Professor',
'ΛΕΚΤΟΡΑΣ': 'Lecturer'

```
},  
parser_type='generic'  
)  
FacultyWebsiteConfig(  
    urls=['https://cs.uowm.gr/en/home-page/members-of-the-staff/'],  
    name_selector='div.tmm_names',  
    container_class='tmm_textblock',  
    staff_type='faculty',  
    debug_print=False,  
    use_cached=True,  
    enabled=True,  
    rank_selector='div.tmm_job',  
    rank_keywords=[  
        'Professor',  
        'Associate Professor',
```

```

        'Assistant Professor',
        'Lecturer',
        'Special Teaching Staff'
    ],
    rank_mapping={
        'Professor': 'Professor',
        'Associate Professor': 'Associate Professor',
        'Assistant Professor': 'Assistant Professor',
        'Lecturer': 'Lecturer',
        'Special Teaching Staff': 'Special Teaching Staff',
        'Καθηγητής': 'Professor',
        'Αναπληρωτής Καθηγητής': 'Associate Professor',
        'Επίκουρος Καθηγητής': 'Assistant Professor',
        'Λέκτορας': 'Lecturer',
        'Ειδικό Διδακτικό Προσωπικό': 'Special Teaching Staff'
    },
    parser_type='container',
    name_concat_selectors=['span.tmm_fname', 'span.tmm_lname']
),
FacultyWebsiteConfig(
    urls=['https://ece.uowm.gr/en/faculty/'],
    name_selector='h3.wpr-member-name',
    container_class='wpr-team-member',
    staff_type='faculty',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector='div.wpr-member-job',
    rank_keywords=[
        'Professor',
        'Associate Professor',

```

```

    'Assistant Professor',
    'Lecturer',
    'Emeritus Professor',
    'Non-tenure Teaching Staff'
],
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Emeritus Professor': 'Emeritus Professor',
    'Non-tenure Teaching Staff': 'Non-tenure Teaching Staff',
    'Καθηγητής': 'Professor',
    'Αναπληρωτής Καθηγητής': 'Associate Professor',
    'Επίκουρος Καθηγητής': 'Assistant Professor',
    'Λέκτορας': 'Lecturer',
    'Ομότιμος Καθηγητής': 'Emeritus Professor'
},
parser_type='container',
search_parent_siblings_for_rank=False,
name_container='div.wpr-member-content',
info_container='div.wpr-member-content'
),
FacultyWebsiteConfig(
    urls=['https://dit.uth.gr/?page_id=35105&lang=en'],
    name_selector='td.ninja_column_0',
    staff_type='faculty',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_keywords=[

```

```

'Professor',
'Associate Professor',
'Assistant Professor',
'Lecturer',
'Emeritus Professor'
],
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Emeritus Professor': 'Emeritus Professor',
    'Καθηγητής': 'Professor',
    'Αναπληρωτής Καθηγητής': 'Associate Professor',
    'Επίκουρος Καθηγητής': 'Assistant Professor',
    'Λέκτορας': 'Lecturer',
    'Ομότιμος Καθηγητής': 'Emeritus Professor'
},
parser_type='generic',
rank_navigation_steps=[
    {'action': 'parent'},
    {'action': 'find', 'selector': 'td.ninja_column_1'}
],
is_table_data=False,
search_parent_siblings_for_rank=False
),
FacultyWebsiteConfig(
    urls=['https://dib.uth.gr/?page_id=8000&lang=en'],
    name_selector='h3.gdlr-core-personnel-list-title a',
    container_class='gdlr-core-personnel-list-column',
    staff_type='faculty',

```

```

debug_print=True,
use_cached=True,
enabled=True,
rank_selector='div.gdlr-core-personnel-list-position',
rank_keywords=[
    'Professor',
    'Associate Professor',
    'Assistant Professor',
    'Lecturer',
    'Emeritus Professor'
],
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Emeritus Professor': 'Emeritus Professor',
    'Καθηγητής': 'Professor',
    'Αναπληρωτής Καθηγητής': 'Associate Professor',
    'Επίκουρος Καθηγητής': 'Assistant Professor',
    'Λέκτορας': 'Lecturer',
    'Ομότιμος Καθηγητής': 'Emeritus Professor'
},
parser_type='container',
search_parent_siblings_for_rank=False,
name_container='div.gdlr-core-personnel-list-content-wrap',
info_container='div.gdlr-core-personnel-list-content-wrap'
),
FacultyWebsiteConfig(
    urls=['https://ds.uth.gr/en/staff-en/faculty-en/'],
    name_selector='h4.et_pb_module_header a',

```

```

container_class='et_pb_blurb_content',
staff_type='faculty',
debug_print=False,
use_cached=True,
enabled=True,
rank_selector='div.et_pb_blurb_description p',
rank_keywords=[
    'Professor',
    'Associate Professor',
    'Assistant Professor',
    'Lecturer',
    'Emeritus Professor',
    'Head of Department',
    'LTeS',
    'Ph.D. (LTeS)',
    '(LTeS)',
    'M.Sc.(STLS)',
    'Postdoc Researcher',
    'Adjunct Lecturer'
],
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Emeritus Professor': 'Emeritus Professor',
    'Head of Department': 'Head of Department',
    'LTeS': 'Lab Lecturer',
    'Ph.D. (LTeS)': 'Lab Lecturer',
    '(LTeS)': 'Lab Lecturer',
    'M.Sc.(STLS)': 'Special Lab Lecturer',

```

```

        'Postdoc Researcher': 'Postdoc Researcher',
        'Adjunct Lecturer': 'Adjunct Lecturer',
        'Καθηγητής': 'Professor',
        'Αναπληρωτής Καθηγητής': 'Associate Professor',
        'Επίκουρος Καθηγητής': 'Assistant Professor',
        'Λέκτορας': 'Lecturer',
        'Ομότιμος Καθηγητής': 'Emeritus Professor'
    },
    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='div.et_pb_blurb_container',
    info_container='div.et_pb_blurb_container'
),
FacultyWebsiteConfig(
    urls=['https://www.e-ce.uth.gr/department/faculty/?lang=en'],
    name_selector='h3.trigger a',
    container_class='toggle-wrap',
    staff_type='faculty',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector='div.heading h2',
    rank_keywords=[
        'Professor',
        'Associate Professor',
        'Assistant Professor',
        'Lecturer',
        'Emeritus Professor',
        'Laboratory Teaching Staff',
        'Adjunct Professor',
        'Collaborating Instructor',
    ]
)

```

```

'MSc Instructor',
'Scientific Collaborator',
'Academic Teaching Experience',
'Retired Professor',
'Former Professor',
'Professor on Sabbatical'
],
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Emeritus Professor': 'Emeritus Professor',
    'Laboratory Teaching Staff': 'Laboratory Teaching Staff',
    'Adjunct Professor': 'Adjunct Professor',
    'Collaborating Instructor': 'Collaborating Instructor',
    'MSc Studies Instructors': 'MSc Instructor',
    'MSc Instructors': 'MSc Instructor',
    'MSC Studies Instructors': 'MSc Instructor',
    'Scientific Collaborator': 'Scientific Collaborator',
    'Academic Teaching Experience': 'Academic Teaching Experience',
    'Retired Professor': 'Retired Professor',
    'Former Professor': 'Former Professor',
    'Professor on Sabbatical': 'Professor on Sabbatical',
    'Καθηγητής': 'Professor',
    'Αναπληρωτής Καθηγητής': 'Associate Professor',
    'Επίκουρος Καθηγητής': 'Assistant Professor',
    'Λέκτορας': 'Lecturer',
    'Ομότιμος Καθηγητής': 'Emeritus Professor'
},
parser_type='sequential',

```

```

main_content_selector='div.entry',
search_parent_siblings_for_rank=False,
name_container='div.toggle-wrap',
info_container='div.toggle_container'
),
FacultyWebsiteConfig(
    urls=['https://ece.auth.gr/en/academic/'],
    name_selector='div.uc_post_title a div',
    container_class='uc_post_grid_style_one_item',
    staff_type='faculty',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector='div.ue-grid-item-meta-data',
    rank_keywords=[
        'Professor',
        'Associate Professor',
        'Assistant Professor',
        'Lecturer',
        'Emeritus Professor',
        'Laboratory Teaching Staff'
    ],
    rank_mapping={
        'Professor': 'Professor',
        'Associate Professor': 'Associate Professor',
        'Assistant Professor': 'Assistant Professor',
        'Lecturer': 'Lecturer',
        'Emeritus Professor': 'Emeritus Professor',
        'Laboratory Teaching Staff': 'Laboratory Teaching Staff',
        'Καθηγητής': 'Professor',
        'Αναπληρωτής Καθηγητής': 'Associate Professor',
    }
)

```

```

        'Επίκουρος Καθηγητής': 'Assistant Professor',
        'Λέκτορας': 'Lecturer',
        'Ομότιμος Καθηγητής': 'Emeritus Professor'
    },
    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='div.uc_content',
    info_container='div.uc_content'
),
FacultyWebsiteConfig(
    urls=['https://ece.auth.gr/en/ergasthriako/'],
    name_selector='div.uc_post_title a div',
    container_class='uc_post_grid_style_one_item',
    staff_type='lab_lecturer',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector='div.ue-grid-item-meta-data',
    rank_keywords=[
        'Ε.Τ.Ε.Π.',
        'Ε.ΔΙ.Π.',
        'Επιστημονικοί Συνεργάτες',
        'Ε.Ε.Π.'
    ],
    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='div.uc_content',
    info_container='div.uc_content'
),
FacultyWebsiteConfig(
    urls=['https://www.di.uoa.gr/en/staff'],

```

```

name_selector='div.views-field-title h6 a',
container_class='col col-xs-12 col-sm-12 col-md-6 col-lg-4',
staff_type='faculty',
debug_print=True,
use_cached=True,
enabled=True,
rank_selector='div.views-field-field-rank-for-professors .field-content',
rank_keywords=[
    'Professor',
    'Associate Professor',
    'Assistant Professor',
    'Lecturer',
    'Emeritus Professor',
    'Retired Professor',
    'Laboratory Teaching Staff',
    'Ε.Τ.Ε.Π.',
    'Ε.ΔΙ.Π.',
    'Επιστημονικοί Συνεργάτες',
    'Ε.Ε.Π.'
],
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Emeritus Professor': 'Emeritus Professor',
    'Retired Professor': 'Retired Professor',
    'Laboratory Teaching Staff': 'Laboratory Teaching Staff',
    'Ε.Τ.Ε.Π.': 'Laboratory Teaching Staff',
    'Ε.ΔΙ.Π.': 'Laboratory Teaching Staff',
    'Επιστημονικοί Συνεργάτες': 'Scientific Collaborator',

```

```

'E.E.Π.': 'Special Teaching Staff',
'Καθηγητής': 'Professor',
'Αναπληρωτής Καθηγητής': 'Associate Professor',
'Επίκουρος Καθηγητής': 'Assistant Professor',
'Λέκτορας': 'Lecturer',
'Ομότιμος Καθηγητής': 'Emeritus Professor'
},
parser_type='container',
search_parent_siblings_for_rank=False,
name_container='div.ep-instructor-details',
info_container='div.ep-instructor-details'
),
FacultyWebsiteConfig(
    urls=['https://www.di.uoa.gr/en/staff?field_staff_specialty_target_id=8'],
    name_selector='div.views-field-title h6 a',
    container_class='col col-xs-12 col-sm-12 col-md-6 col-lg-4',
    staff_type='leb_lecturer',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector='div.views-field-field-staff-specialty .field-content',
    rank_keywords=[
        'Laboratory Teaching Staff',
        'Ε.Τ.Ε.Π.',
        'Ε.ΔΙ.Π.',
        'Επιστημονικοί Συνεργάτες',
        'Ε.Ε.Π.'
    ],
    rank_mapping={
        'Laboratory Teaching Staff': 'Laboratory Teaching Staff',
        'Ε.Τ.Ε.Π.': 'Laboratory Teaching Staff',

```

```

'E.ΔΙ.Π.': 'Laboratory Teaching Staff',
'Επιστημονικοί Συνεργάτες': 'Scientific Collaborator',
'E.Ε.Π.': 'Special Teaching Staff'
},
parser_type='container',
search_parent_siblings_for_rank=False,
name_container='div.ep-instructor-details',
info_container='div.ep-instructor-details'
),
FacultyWebsiteConfig(
    urls=['https://dit.hua.gr/index.php/en/department/faculty-members'],
    name_selector='h3',
    container_class='row',
    staff_type='faculty',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_keywords=[
        'Professor',
        'Associate Professor',
        'Assistant Professor',
        'Lecturer',
        'Emeritus Professor',
        'Professor Emeritus',
        'Retired Professor',
        'Adjunct Professor',
        'Visiting Professor',
        'Research Professor',
        'Clinical Professor',
        'Distinguished Professor',
        'Honorary Professor',
    ]
)

```

```

        'Teaching Associate'
    ],
    rank_mapping={
        'Professor': 'Professor',
        'Associate Professor': 'Associate Professor',
        'Assistant Professor': 'Assistant Professor',
        'Lecturer': 'Lecturer',
        'Emeritus Professor': 'Emeritus Professor',
        'Professor Emeritus': 'Emeritus Professor',
        'Retired Professor': 'Retired Professor',
        'Adjunct Professor': 'Adjunct Professor',
        'Visiting Professor': 'Visiting Professor',
        'Research Professor': 'Research Professor',
        'Clinical Professor': 'Clinical Professor',
        'Distinguished Professor': 'Distinguished Professor',
        'Honorary Professor': 'Honorary Professor',
        'Teaching Associate': 'Teaching Associate'
    },
    parser_type='container',
    name_rank_split_separator=', ',
    search_parent_siblings_for_rank=False,
    name_container='div.col-md-8',
    info_container='div.col-md-8'
),
FacultyWebsiteConfig(
    urls=['https://dit.uop.gr/en/members_depen'],
    name_selector='span.tbold a',
    rank_selector='div.views-field-value-4 span.field-content',
    container_class='grid',
    staff_type='faculty',
    debug_print=True,

```

```

use_cached=True,
enabled=True,
rank_keywords=[
    'PROFESSOR',
    'ASSOCIATE PROFESSOR',
    'ASSISTANT PROFESSOR',
    'LECTURER',
    'EMERITUS PROFESSOR',
    'PROFESSOR EMERITUS',
    'RETIRED PROFESSOR',
    'ADJUNCT PROFESSOR',
    'VISITING PROFESSOR',
    'RESEARCH PROFESSOR',
    'CLINICAL PROFESSOR',
    'DISTINGUISHED PROFESSOR',
    'HONORARY PROFESSOR',
    'TEACHING ASSOCIATE'
],
rank_mapping={
    'PROFESSOR': 'Professor',
    'ASSOCIATE PROFESSOR': 'Associate Professor',
    'ASSISTANT PROFESSOR': 'Assistant Professor',
    'LECTURER': 'Lecturer',
    'EMERITUS PROFESSOR': 'Emeritus Professor',
    'PROFESSOR EMERITUS': 'Emeritus Professor',
    'RETIRED PROFESSOR': 'Retired Professor',
    'ADJUNCT PROFESSOR': 'Adjunct Professor',
    'VISITING PROFESSOR': 'Visiting Professor',
    'RESEARCH PROFESSOR': 'Research Professor',
    'CLINICAL PROFESSOR': 'Clinical Professor',
    'DISTINGUISHED PROFESSOR': 'Distinguished Professor',

```

```

        'HONORARY PROFESSOR': 'Honorary Professor',
        'TEACHING ASSOCIATE': 'Teaching Associate'
    },
    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='li.grid',
    info_container='li.grid'
),
FacultyWebsiteConfig(
    urls=['https://dit.uop.gr/en/members_edipen'],
    name_selector='strong.views-field-value-1 a',
    rank_selector='div.views-field-value-4 span.field-content',
    container_class='grid',
    staff_type='lab_lecturer',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_keywords=[
        'LABORATORY TEACHING STAFF',
        'Ε.Τ.Ε.Π.',
        'Ε.ΔΙ.Π.',
        'Επιστημονικοί Συνεργάτες',
        'Ε.Ε.Π.',
        'SPECIAL TEACHING STAFF'
    ],
    rank_mapping={
        'LABORATORY TEACHING STAFF': 'Laboratory Teaching Staff',
        'Ε.Τ.Ε.Π.': 'Laboratory Teaching Staff',
        'Ε.ΔΙ.Π.': 'Laboratory Teaching Staff',
        'Επιστημονικοί Συνεργάτες': 'Scientific Collaborator',
        'Ε.Ε.Π.': 'Special Teaching Staff',
    }
)

```

```

        'SPECIAL TEACHING STAFF': 'Special Teaching Staff'
    },
    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='li.grid',
    info_container='li.grid'
),
FacultyWebsiteConfig(
    urls=['https://ds.uop.gr/en/members_depen'],
    name_selector='span.tbld a',
    rank_selector='div.views-field-value-4 span.field-content',
    container_class='grid',
    staff_type='faculty',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_keywords=[
        'PROFESSOR',
        'ASSOCIATE PROFESSOR',
        'ASSISTANT PROFESSOR',
        'LECTURER',
        'EMERITUS PROFESSOR',
        'PROFESSOR EMERITUS',
        'RETIRED PROFESSOR',
        'ADJUNCT PROFESSOR',
        'VISITING PROFESSOR',
        'RESEARCH PROFESSOR',
        'CLINICAL PROFESSOR',
        'DISTINGUISHED PROFESSOR',
        'HONORARY PROFESSOR',
        'TEACHING ASSOCIATE'
    ]
)

```

```

],
rank_mapping={
    'PROFESSOR': 'Professor',
    'ASSOCIATE PROFESSOR': 'Associate Professor',
    'ASSISTANT PROFESSOR': 'Assistant Professor',
    'LECTURER': 'Lecturer',
    'EMERITUS PROFESSOR': 'Emeritus Professor',
    'PROFESSOR EMERITUS': 'Emeritus Professor',
    'RETIRED PROFESSOR': 'Retired Professor',
    'ADJUNCT PROFESSOR': 'Adjunct Professor',
    'VISITING PROFESSOR': 'Visiting Professor',
    'RESEARCH PROFESSOR': 'Research Professor',
    'CLINICAL PROFESSOR': 'Clinical Professor',
    'DISTINGUISHED PROFESSOR': 'Distinguished Professor',
    'HONORARY PROFESSOR': 'Honorary Professor',
    'TEACHING ASSOCIATE': 'Teaching Associate'
},
parser_type='container',
search_parent_siblings_for_rank=False,
name_container='li.grid',
info_container='li.grid'
),
FacultyWebsiteConfig(
    urls=['https://ds.uop.gr/index.php/en/members_eidikoen'],
    name_selector='strong.views-field-value-1 a',
    container_class='grid',
    staff_type='lab_lecturer',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector=None,

```

```

rank_keywords=[],
rank_mapping={},
parser_type='container',
search_parent_siblings_for_rank=False,
name_container='li.grid',
info_container='li.grid',
fixed_rank='Laboratory Teaching Staff'
),
FacultyWebsiteConfig(
    urls=['https://www.ece.uop.gr/en/our_team_category/faculty_members/'],
    name_selector='h4.title a',
    container_class='our-team-item',
    staff_type='faculty',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector='div.regency',
    rank_keywords=[
        'Professor',
        'Associate Professor',
        'Assistant Professor',
        'Lecturer',
        'Emeritus Professor',
        'Professor Emeritus',
        'Retired Professor',
        'Adjunct Professor',
        'Visiting Professor',
        'Research Professor',
        'Clinical Professor',
        'Distinguished Professor',
        'Honorary Professor',
    ]
)

```

```

        'Teaching Associate'
    ],
    rank_mapping={
        'Professor': 'Professor',
        'Associate Professor': 'Associate Professor',
        'Assistant Professor': 'Assistant Professor',
        'Lecturer': 'Lecturer',
        'Emeritus Professor': 'Emeritus Professor',
        'Professor Emeritus': 'Emeritus Professor',
        'Retired Professor': 'Retired Professor',
        'Adjunct Professor': 'Adjunct Professor',
        'Visiting Professor': 'Visiting Professor',
        'Research Professor': 'Research Professor',
        'Clinical Professor': 'Clinical Professor',
        'Distinguished Professor': 'Distinguished Professor',
        'Honorary Professor': 'Honorary Professor',
        'Teaching Associate': 'Teaching Associate'
    },
    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='div.content-team',
    info_container='div.content-team'
),
FacultyWebsiteConfig(
    urls=['https://www.ece.uop.gr/en/our_team_category/etep-en/'],
    name_selector='h4.title a',
    container_class='our-team-item',
    staff_type='lab_lecturer',
    debug_print=True,
    use_cached=True,
    enabled=True,

```

```

rank_selector='div.regency',
rank_keywords=[
    'Special Technical Laboratory Staff (ETEP)',
    'ETEP',
    'Special Technical Laboratory Staff',
    'Laboratory Teaching Staff',
    'Ε.Τ.Ε.Π.',
    'Ε.ΔΙ.Π.',
    'Επιστημονικοί Συνεργάτες',
    'Ε.Ε.Π.',
    'Special Teaching Staff'
],
rank_mapping={
    'Special Technical Laboratory Staff (ETEP)': 'Special Technical Laboratory Staff (ETEP)',
    'ETEP': 'Special Technical Laboratory Staff (ETEP)',
    'Special Technical Laboratory Staff': 'Special Technical Laboratory Staff (ETEP)',
    'Laboratory Teaching Staff': 'Laboratory Teaching Staff',
    'Ε.Τ.Ε.Π.': 'Laboratory Teaching Staff',
    'Ε.ΔΙ.Π.': 'Laboratory Teaching Staff',
    'Επιστημονικοί Συνεργάτες': 'Scientific Collaborator',
    'Ε.Ε.Π.': 'Special Teaching Staff',
    'Special Teaching Staff': 'Special Teaching Staff'
},
parser_type='container',
search_parent_siblings_for_rank=False,
name_container='div.content-team',
info_container='div.content-team'
),
FacultyWebsiteConfig(
    urls=['https://ece.hmu.gr/en/faculty/'],
    name_selector='h3',

```

```

container_class='et_pb_row',
staff_type='faculty',
debug_print=True,
use_cached=True,
enabled=True,
rank_selector='h4',
rank_keywords=[
    'Professor',
    'Associate Professor',
    'Assistant Professor',
    'Lecturer',
    'Emeritus Professor',
    'Professor Emeritus',
    'Retired Professor',
    'Adjunct Professor',
    'Visiting Professor',
    'Research Professor',
    'Clinical Professor',
    'Distinguished Professor',
    'Honorary Professor',
    'Teaching Associate',
    'EMERITUS PROFESSOR',
    'ASSISTANT PROFESSOR'
],
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Emeritus Professor': 'Emeritus Professor',
    'Professor Emeritus': 'Emeritus Professor',

```

```

'Retired Professor': 'Retired Professor',
'Adjunct Professor': 'Adjunct Professor',
'Visiting Professor': 'Visiting Professor',
'Research Professor': 'Research Professor',
'Clinical Professor': 'Clinical Professor',
'Distinguished Professor': 'Distinguished Professor',
'Honorary Professor': 'Honorary Professor',
'Teaching Associate': 'Teaching Associate',
'EMERITUS PROFESSOR': 'Emeritus Professor',
'ASSISTANT PROFESSOR': 'Assistant Professor'
},
parser_type='container',
search_parent_siblings_for_rank=False,
name_container='div.et_pb_text_inner',
info_container='div.et_pb_text_inner',
rank_navigation_steps=[
    {'action': 'parent', 'selector': 'div.et_pb_text_inner'},
    {'action': 'find', 'selector': 'h4'}
]
),
FacultyWebsiteConfig(
    urls=['https://ece.hmu.gr/en/teaching-lab-staff/'],
    name_selector='p strong',
    container_class='et_pb_text_inner',
    staff_type='lab_lecturer',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector=None,
    rank_keywords=[],
    rank_mapping={},

```

```

parser_type='container',
search_parent_siblings_for_rank=False,
name_container='div.et_pb_text_inner',
info_container='div.et_pb_text_inner',
fixed_rank='Teaching Laboratory Staff'
),
FacultyWebsiteConfig(
    urls=['https://ece.hmu.gr/en/special-technical-laboratory-staff/'],
    name_selector='p span',
    container_class='et_pb_row',
    staff_type='lab_lecturer2',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector=None,
    rank_keywords=[],
    rank_mapping={},
    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='div.et_pb_text_inner',
    info_container='div.et_pb_text_inner',
    fixed_rank='Special Technical Laboratory Staff'
),
FacultyWebsiteConfig(
    urls=['https://ee.hmu.gr/en/faculty/'],
    name_selector='h3',
    container_class='et_pb_row',
    main_content_selector='div.et_pb_post_content_0_tb_body',
    staff_type='faculty',
    debug_print=True,
    use_cached=True,

```

```

enabled=True,
rank_selector='h4',
rank_keywords=[
    'Professor',
    'Associate Professor',
    'Assistant Professor',
    'Lecturer',
    'Emeritus Professor',
    'Professor Emeritus',
    'Retired Professor',
    'Adjunct Professor',
    'Visiting Professor',
    'Research Professor',
    'Clinical Professor',
    'Distinguished Professor',
    'Honorary Professor',
    'Teaching Associate',
    'EMERITUS PROFESSOR',
    'ASSISTANT PROFESSOR',
    'PROFESSOR',
    'PROFESSOR | HEAD OF DEPARTMENT',
    'PROFESSOR | VICE-HEAD OF DEPARTMENT'
],
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Emeritus Professor': 'Emeritus Professor',
    'Professor Emeritus': 'Emeritus Professor',
    'Retired Professor': 'Retired Professor',

```

```

'Adjunct Professor': 'Adjunct Professor',
'Visiting Professor': 'Visiting Professor',
'Research Professor': 'Research Professor',
'Clinical Professor': 'Clinical Professor',
'Distinguished Professor': 'Distinguished Professor',
'Honorary Professor': 'Honorary Professor',
'Teaching Associate': 'Teaching Associate',
'EMERITUS PROFESSOR': 'Emeritus Professor',
'ASSISTANT PROFESSOR': 'Assistant Professor',
'PROFESSOR': 'Professor',
'PROFESSOR | HEAD OF DEPARTMENT': 'Professor',
'PROFESSOR | VICE-HEAD OF DEPARTMENT': 'Professor'
},
parser_type='container',
search_parent_siblings_for_rank=False,
name_container='div.et_pb_text_inner',
info_container='div.et_pb_text_inner'
),
FacultyWebsiteConfig(
    urls=['https://ee.hmu.gr/en/department/staff/melh-e-di-p/'],
    name_selector='h3 strong',
    container_class='et_pb_row_0',
    staff_type='lab_lecturer',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector=None,
    rank_keywords=[],
    rank_mapping={},
    parser_type='container',
    search_parent_siblings_for_rank=False,

```

```

name_container='div.et_pb_text_inner',
info_container='div.et_pb_text_inner',
fixed_rank='Laboratory Faculty',
main_content_selector='div.et_pb_post_content_0_tb_body'
),
FacultyWebsiteConfig(
    urls=['https://ee.hmu.gr/en/department/staff/special-technical-laboratory-staff/'],
    name_selector='p strong',
    container_class='et_pb_post_content_0_tb_body',
    staff_type='lab_lecturer2',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector=None,
    rank_keywords=[],
    rank_mapping={},
    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='div.et_pb_text_inner',
    info_container='div.et_pb_text_inner',
    fixed_rank='Special Technical Laboratory Staff',
    main_content_selector='div.et_pb_post_content_0_tb_body'
),
FacultyWebsiteConfig(
    urls=['https://www.ece.ntua.gr/en/staff/academic'],
    name_selector='div.name',
    container_class='div.list-item.member',
    staff_type='faculty',
    debug_print=True,
    use_cached=True,
    enabled=True,

```

```

rank_selector='div.job-title',
rank_keywords=[
    'Professor',
    'Associate Professor',
    'Assistant Professor',
    'Lecturer',
    'Emeritus Professor',
    'Professor Emeritus',
    'Retired Professor',
    'Adjunct Professor',
    'Visiting Professor',
    'Research Professor',
    'Clinical Professor',
    'Distinguished Professor',
    'Honorary Professor',
    'Teaching Associate'
],
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Emeritus Professor': 'Emeritus Professor',
    'Professor Emeritus': 'Emeritus Professor',
    'Retired Professor': 'Retired Professor',
    'Adjunct Professor': 'Adjunct Professor',
    'Visiting Professor': 'Visiting Professor',
    'Research Professor': 'Research Professor',
    'Clinical Professor': 'Clinical Professor',
    'Distinguished Professor': 'Distinguished Professor',
    'Honorary Professor': 'Honorary Professor',

```

```

        'Teaching Associate': 'Teaching Associate'
    },
    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='div.list-item-inner',
    info_container='div.list-item-inner'
),
FacultyWebsiteConfig(
    urls=['https://www.ece.ntua.gr/en/staff/laboratory'],
    name_selector='div.name',
    container_class='div.list-item.member',
    staff_type='lab_lecturer',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector=None,
    rank_keywords=[],
    rank_mapping={},
    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='div.list-item-inner',
    info_container='div.list-item-inner',
    fixed_rank='Laboratory Staff'
),
FacultyWebsiteConfig(
    urls=['https://www.ece.tuc.gr/en/school/faculty-and-staff/faculty'],
    name_selector='div.datadivcontainer strong',
    container_class='div.personitem.twocolumns',
    staff_type='faculty',
    debug_print=True,
    use_cached=True,

```

```

enabled=True,
rank_selector='div.datadivcontainer div:nth-child(2)',
rank_keywords=[
    'Professor',
    'Associate Professor',
    'Assistant Professor',
    'Lecturer',
    'Emeritus Professor',
    'Professor Emeritus',
    'Retired Professor',
    'Adjunct Professor',
    'Visiting Professor',
    'Research Professor',
    'Clinical Professor',
    'Distinguished Professor',
    'Honorary Professor',
    'Teaching Associate'
],
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Emeritus Professor': 'Emeritus Professor',
    'Professor Emeritus': 'Emeritus Professor',
    'Retired Professor': 'Retired Professor',
    'Adjunct Professor': 'Adjunct Professor',
    'Visiting Professor': 'Visiting Professor',
    'Research Professor': 'Research Professor',
    'Clinical Professor': 'Clinical Professor',
    'Distinguished Professor': 'Distinguished Professor',

```

```

        'Honorary Professor': 'Honorary Professor',
        'Teaching Associate': 'Teaching Associate'
    },
    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='div.datadivcontainer',
    info_container='div.datadivcontainer'
),
FacultyWebsiteConfig(
    urls=['https://www.ece.tuc.gr/en/school/faculty-and-staff/laboratory-teaching-staff'],
    name_selector='div.datadivcontainer strong',
    container_class='div.personitem.twocolumns',
    staff_type='lab_lecturer',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector='div.datadivcontainer div:nth-child(2)',
    rank_keywords=[
        'EDIP',
        'ΕΔΙΠ',
        'Laboratory Teaching Staff',
        'Special Teaching Staff',
        'Ειδικό Διδακτικό Προσωπικό'
    ],
    rank_mapping={
        'EDIP': 'Laboratory Teaching Staff',
        'ΕΔΙΠ': 'Laboratory Teaching Staff',
        'Laboratory Teaching Staff': 'Laboratory Teaching Staff',
        'Special Teaching Staff': 'Special Teaching Staff',
        'Ειδικό Διδακτικό Προσωπικό': 'Special Teaching Staff'
    },

```

```

    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='div.datadivcontainer',
    info_container='div.datadivcontainer'
),
FacultyWebsiteConfig(
    urls=['https://www.ee.duth.gr/en/the-department/personnel/faculty-members/'],
    name_selector='td a',
    container_class='table-1',
    staff_type='faculty',
    debug_print=True,
    use_cached=True,
    enabled=True,
    is_table_data=False,
    rank_selector='td:nth-child(2)',
    rank_keywords=[
        'Professor',
        'Associate Professor',
        'Assistant Professor',
        'Lecturer',
        'Emeritus Professor',
        'Professor Emeritus',
        'Retired Professor',
        'Adjunct Professor',
        'Visiting Professor',
        'Research Professor',
        'Clinical Professor',
        'Distinguished Professor',
        'Honorary Professor',
        'Teaching Associate'
    ],

```

```

rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Emeritus Professor': 'Emeritus Professor',
    'Professor Emeritus': 'Emeritus Professor',
    'Retired Professor': 'Retired Professor',
    'Adjunct Professor': 'Adjunct Professor',
    'Visiting Professor': 'Visiting Professor',
    'Research Professor': 'Research Professor',
    'Clinical Professor': 'Clinical Professor',
    'Distinguished Professor': 'Distinguished Professor',
    'Honorary Professor': 'Honorary Professor',
    'Teaching Associate': 'Teaching Associate'
},
parser_type='generic',
rank_navigation_steps=[
    {'action': 'parent'},
    {'action': 'parent'},
    {'action': 'find', 'selector': 'td:nth-child(2)'}
],
search_parent_siblings_for_rank=False,
name_container='td',
info_container='td'
),
FacultyWebsiteConfig(
    urls=['https://eee.uniwa.gr/en/faculty-staff/edu-staff'],
    name_selector='div.personfield.index0.fieldname span',
    container_class='div.person',
    staff_type='faculty',

```

```

debug_print=True,
use_cached=True,
enabled=True,
rank_selector='div.personfield.index1.fieldposition span.fieldvalue',
rank_keywords=[
    'Professor',
    'Associate Professor',
    'Assistant Professor',
    'Lecturer',
    'Emeritus Professor',
    'Professor Emeritus',
    'Retired Professor',
    'Adjunct Professor',
    'Visiting Professor',
    'Research Professor',
    'Clinical Professor',
    'Distinguished Professor',
    'Honorary Professor',
    'Teaching Associate'
],
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Emeritus Professor': 'Emeritus Professor',
    'Professor Emeritus': 'Emeritus Professor',
    'Retired Professor': 'Retired Professor',
    'Adjunct Professor': 'Adjunct Professor',
    'Visiting Professor': 'Visiting Professor',
    'Research Professor': 'Research Professor',

```

```

        'Clinical Professor': 'Clinical Professor',
        'Distinguished Professor': 'Distinguished Professor',
        'Honorary Professor': 'Honorary Professor',
        'Teaching Associate': 'Teaching Associate'
    },
    parser_type='container',
    search_parent_siblings_for_rank=False,
    name_container='div.personinfo',
    info_container='div.personinfo'
),
FacultyWebsiteConfig(
    urls=['https://eee.uniwa.gr/en/faculty-staff/etep-en'],
    name_selector='div.personfield.index0.fieldname span',
    container_class='div.person',
    staff_type='lab_lecturer',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector='div.personfield.index1.fieldposition span.fieldvalue',
    rank_keywords=[
        'Technical Laboratory Staff',
        'Laboratory Staff',
        'Technical Staff',
        'Lab Staff',
        'Laboratory Technician',
        'Technical Assistant',
        'Lab Assistant',
        'Research Assistant',
        'Laboratory Assistant',
        'Technical Support Staff'
    ],

```

```

rank_mapping={
    'Technical Laboratory Staff': 'Technical Laboratory Staff',
    'Laboratory Staff': 'Laboratory Staff',
    'Technical Staff': 'Technical Staff',
    'Lab Staff': 'Lab Staff',
    'Laboratory Technician': 'Laboratory Technician',
    'Technical Assistant': 'Technical Assistant',
    'Lab Assistant': 'Lab Assistant',
    'Research Assistant': 'Research Assistant',
    'Laboratory Assistant': 'Laboratory Assistant',
    'Technical Support Staff': 'Technical Support Staff'
},
parser_type='container',
search_parent_siblings_for_rank=False,
name_container='div.personinfo',
info_container='div.personinfo'
),
FacultyWebsiteConfig(
    urls=['https://ice.uniwa.gr/en/staff-2/faculty/'],
    name_selector='h2.name',
    container_class='container',
    staff_type='faculty',
    debug_print=True,
    use_cached=True,
    enabled=True,
    rank_selector='div.job-title',
    rank_keywords=[
        'Professor',
        'Associate Professor',
        'Assistant Professor',
        'Lecturer',

```

```

    'Applications Lecturer',
    'Laboratory Teaching Staff',
    'Laboratory Technical Staff',
    'Administrative Staff',
    'Retired',
    'Emeritus Professor'
],
rank_mapping={
    'Professor': 'Professor',
    'Associate Professor': 'Associate Professor',
    'Assistant Professor': 'Assistant Professor',
    'Lecturer': 'Lecturer',
    'Applications Lecturer': 'Applications Lecturer',
    'Laboratory Teaching Staff': 'Laboratory Teaching Staff',
    'Laboratory Technical Staff': 'Laboratory Technical Staff',
    'Administrative Staff': 'Administrative Staff',
    'Retired': 'Retired',
    'Emeritus Professor': 'Emeritus Professor'
},
parser_type='container',
search_parent_siblings_for_rank=False,
name_container='div.container',
info_container='div.container'
),
FacultyWebsiteConfig(
    urls=['https://ice.uniwa.gr/en/staff-2/lab-teaching/'],
    name_selector='h2.name',
    container_class='container',
    staff_type='edip_staff',
    debug_print=True,
    use_cached=True,

```

```

        enabled=True,
        rank_selector=None,
        rank_keywords=[],
        rank_mapping={},
        parser_type='container',
        search_parent_siblings_for_rank=False,
        name_container='div.container',
        info_container='div.container',
        fixed_rank='Lab Teaching Staff'
    ),
    FacultyWebsiteConfig(
        urls=['https://ice.uniwa.gr/en/staff-2/lab-technical/'],
        name_selector='h2.name',
        container_class='container',
        staff_type='etep_staff',
        debug_print=True,
        use_cached=True,
        enabled=True,
        rank_selector=None,
        rank_keywords=[],
        rank_mapping={},
        parser_type='container',
        search_parent_siblings_for_rank=False,
        name_container='div.container',
        info_container='div.container',
        fixed_rank='Lab Technical Staff'
    ),
    FacultyWebsiteConfig(
        urls=['https://www.icsd.aegean.gr/akadimaiko_en'],
        name_selector='p.name',
        container_class='element-item',

```

```

        staff_type='faculty',
        debug_print=True,
        use_cached=True,
        enabled=True,
        rank_container={'attr': 'data-category'},
        rank_mapping={
            'epikouros': 'Assistant Professor',
            'anaplirotis': 'Associate Professor',
            'kathigitis': 'Professor'
        },
        parser_type='generic'
    )
]

```

```

robots_handler = RobotsTxtHandler()

```

```

all_faculty_data = []

```

```

enabled_configs = [config for config in configs if config.enabled]

```

```

print(f"Starting scraping of {len(enabled_configs)} enabled configurations...")

```

```

for i, config in enumerate(enabled_configs, 1):

```

```

    print(f"\n[{i}/{len(enabled_configs)}] Processing: {config.urls[0]}")

```

```

    try:

```

```

        faculty_data = scrape_faculty_with_validation(config)

```

```

university = get_university_name(config.urls[0])

if ' ' in university and len(university.split()) >= 2:
    department = extract_department_from_university_name(university)
else:
    department = extract_department_from_url(config.urls[0])

for member in faculty_data:
    member['university'] = university
    member['department'] = department
    member['source_url'] = config.urls[0]
    member['staff_type'] = config.staff_type

    filename = f'faculty_data/{urlparse(config.urls[0]).netloc.replace('www.',
    '')}_{config.staff_type}.json'
    try:
        with open(filename, 'w', encoding='utf-8') as f:
            json.dump(faculty_data, f, indent=4, ensure_ascii=False)
            print(f'Saved {len(faculty_data)} members to {filename}')
    except Exception as e:
        print(f'Error saving to {filename}: {str(e)}')

all_faculty_data.extend(faculty_data)
print(f'Found {len(faculty_data)} faculty members from {university} - {department}')

print(f'\nDeduplicating faculty data...')
original_count = len(all_faculty_data)

```

```

all_faculty_data = deduplicate_faculty_data(all_faculty_data)
deduplicated_count = len(all_faculty_data)
print(f"Deduplication complete: {original_count} -> {deduplicated_count} (removed
{original_count - deduplicated_count} duplicates)")

```

```

except Exception as e:
    print(f"Error processing {config.urls[0]}: {str(e)}")
    continue

```

```

output_file = 'all_faculty_data.json'
try:
    with open(output_file, 'w', encoding='utf-8') as f:
        json.dump(all_faculty_data, f, ensure_ascii=False, indent=2)
    print(f"\nSuccessfully saved {len(all_faculty_data)} faculty members to {output_file}")
except Exception as e:
    print(f"Error saving to {output_file}: {str(e)}")
    return 1

```

```

print(f"\n=== SCRAPING SUMMARY ===")
print(f"Total faculty members found: {len(all_faculty_data)}")

```

```

universities = {}
for member in all_faculty_data:
    uni = member.get('university', 'Unknown')
    universities[uni] = universities.get(uni, 0) + 1

```

```

print(f"\nBy University:")
for uni, count in sorted(universities.items()):
    print(f" {uni}: {count}")

```

```
ranks = {}  
  
for member in all_faculty_data:  
    rank = member.get('rank', 'Unknown')  
    ranks[rank] = ranks.get(rank, 0) + 1  
  
print(f"\nBy Rank:")  
for rank, count in sorted(ranks.items()):  
    print(f" {rank}: {count}")  
  
print(f"\nScraping completed successfully!")  
return 0
```