



**Ντάτζης Παναγιώτης**

**ΑΜ: Π2015042**

**Ιόνιο Πανεπιστήμιο**

**Επισκόπηση οντολογιών για μετεωρολογικά δεδομένα**

**Πτυχιακή εργασία για το πτυχίο: Πληροφορικής**

**Σεπτέμβριος, 2022**

## **Επιβλέπουσα**

**Ονοματεπώνυμο:** Σωτηροπούλου Άννα, ΕΔΙΠ

Ιόνιο Πανεπιστήμιο, Τμήμα Πληροφορικής

## **Τριμελής Επιτροπή**

**Ονοματεπώνυμο:** Σωτηροπούλου Άννα, ΕΔΙΠ

Ιόνιο Πανεπιστήμιο, Τμήμα Πληροφορικής

**Ονοματεπώνυμο:** Κερμανίδου Κάτια - Λήδα, Αναπληρώτρια Καθηγήτρια

Ιόνιο Πανεπιστήμιο, Τμήμα Πληροφορικής

**Ονοματεπώνυμο:** Μυλωνάς Φοίβος, Αναπληρωτής Καθηγητής

Ιόνιο Πανεπιστήμιο, Τμήμα Πληροφορικής

## **Ευχαριστίες**

Θα ήθελα να ευχαριστήσω τους γονείς μου που στάθηκαν δίπλα μου με όλη τους την αγάπη σε αυτό το ταξίδι. Ακόμη θα ήθελα να ευχαριστήσω θερμά την καθηγήτρια Άννα Σωτηροπούλου, η οποία με εμπιστεύτηκε με το να μου προσφέρει ένα θέμα ώστε να εκπονήσω την πτυχιακή μου εργασία. Χωρίς την καθοδήγηση και τις γνώσεις το αποτέλεσμα δεν θα ήταν το ίδιο.

## Περίληψη

Με την παρούσα πτυχιακή εργασία επιχειρείται η σύγκριση ερευνών σχετικά με οντολογίες από μετεωρολογικά δεδομένα καθώς και η ανάλυση εννοιών που χρησιμοποιήθηκαν για τις οντολογίες.

Στο πλαίσιο της Οντολογίας, εξετάζει κανείς τον λόγο ύπαρξης μιας οντότητας, τον σκοπό της, αναφορικά με μια εννοιολογική μορφοποίηση σε ένα πεδίο ενδιαφέροντος (Όντα = Αυτά που έχουν υπόσταση). Συγκεκριμένα, η Οντολογία αναφέρεται σε μία κοινή και τυπική αναπαράσταση γνώσεων από ένα σύνολο δεδομένων, απλοποιώντας την εξαγωγή συμπερασμάτων, καθώς και τη δομημένη περιγραφή δεδομένων σε ένα συγκεκριμένο πεδίο.

Οι οντολογίες αποτελούν εργαλείο για πολλά πεδία για την συλλογή, απλοποίηση και οργάνωση μεγάλου όγκου δεδομένων. Τέτοια πεδία είναι ο Σηματολογικός Ιστός, η Τεχνητή Νοημοσύνη και η Βιοπληροφορική.

Το πεδίο των μετεωρολογικών δεδομένων αποτελεί χαρακτηριστικό παράδειγμα πεδίου με μεγάλο όγκο δεδομένων, αλλά και πεδίο στο οποίο διαφαίνεται ότι οι οντολογίες καθώς και τα συστήματα για την ανάλυση και μετάφραση των πληροφοριών μέσω τεχνικών και υπολογιστικών συστημάτων καθιστούν σημαντική και αναγκαία την ύπαρξή τους.

Στη σύγχρονη εποχή, οι οντολογίες διαδραματίζουν καθοριστικό ρόλο στην επιτυχημένη διεπαφή ανάμεσα σε ανθρώπους και υπολογιστικά συστήματα τα οποία φορούν την εμβάθυνση εξαγωγής, συλλογής, επεξεργασίας, αποθήκευσης και μετάφρασης τέτοιων δεδομένων.

**Λέξεις Κλειδιά:** Οντολογίες, RDF, XML, OWLTime



## Πίνακας Περιεχομένων

|                                                                                                   |    |
|---------------------------------------------------------------------------------------------------|----|
| Εισαγωγή .....                                                                                    | 10 |
| Σκοπός.....                                                                                       | 11 |
| Μέρος Α.....                                                                                      | 12 |
| 1.1 Οντολογίες .....                                                                              | 12 |
| 1.2 Resource Description Framework (RDF) .....                                                    | 13 |
| 1.3 Extensible Markup Language (XML) .....                                                        | 14 |
| 1.4 OWL Time Ontology.....                                                                        | 14 |
| 1.5 Gloze Set of Tools .....                                                                      | 15 |
| 1.6 Vortex2.....                                                                                  | 15 |
| 1.7 Πρόγραμμα Copernicus Sentinel-2.....                                                          | 15 |
| 1.8 SERVIR-Viz 3-D Visualization Tool .....                                                       | 16 |
| Μέρος Β .....                                                                                     | 17 |
| Εργασία 1 <sup>η</sup> .....                                                                      | 17 |
| Πλεονεκτήματα .....                                                                               | 18 |
| Μεθοδολογία .....                                                                                 | 19 |
| Παρόμοιες μελέτες.....                                                                            | 21 |
| Πρόταση.....                                                                                      | 22 |
| Εργασία 2 <sup>η</sup> .....                                                                      | 27 |
| Πλεονεκτήματα .....                                                                               | 27 |
| Μεθοδολογία .....                                                                                 | 28 |
| Παρόμοιες μελέτες.....                                                                            | 28 |
| Έρευνα και αποτελέσματα.....                                                                      | 30 |
| Σύγκριση 1ης και 2ης έρευνας.....                                                                 | 31 |
| Εργασία 3 <sup>η</sup> .....                                                                      | 32 |
| Μεθοδολογία δημιουργίας της οντολογίας MEMOn (Modular Environmental Monitoring Ontology)<br>..... | 33 |
| Φάση 1 <sup>η</sup> : Εξερεύνηση .....                                                            | 35 |
| Φάση 2 <sup>η</sup> : Σχεδιασμός.....                                                             | 36 |
| Φάση 3 <sup>η</sup> : Ανάπτυξη Ενότητας .....                                                     | 39 |
| Φάση 4 <sup>η</sup> : Απελευθέρωση.....                                                           | 40 |
| Παρόμοιες μελέτες.....                                                                            | 43 |
| Συμπεράσματα .....                                                                                | 44 |

|                                                                        |    |
|------------------------------------------------------------------------|----|
| Εργασία 4η .....                                                       | 45 |
| Τυπική Ανάλυση Εννοιών (Formal Concept Analysis - FCA) .....           | 46 |
| Λανθάνουσα Κατανομή Dirichlet (Latent Dirichlet Allocation - LDA)..... | 47 |
| Αλγόριθμος Apriori .....                                               | 48 |
| Δομή Κατασκευής της Οντολογίας.....                                    | 49 |
| Οντολογία τυφώνων.....                                                 | 50 |
| Υπολογισμός Σημασιολογικής Ομοιότητας .....                            | 50 |
| Εξακρίβωση συνάφειας των ιστοσελίδων .....                             | 51 |
| Παρόμοιες μελέτες.....                                                 | 52 |
| Συμπεράσματα .....                                                     | 53 |
| Σύγκριση 3 <sup>ης</sup> και 4 <sup>ης</sup> εργασίας .....            | 54 |
| Βιβλιογραφία .....                                                     | 56 |

## Ευρετήριο Εικόνων

|                                                                                                             |    |
|-------------------------------------------------------------------------------------------------------------|----|
| Εικόνα 1: Πρόγραμμα CopernicusSentinel-2[7].....                                                            | 16 |
| Εικόνα 2: Παράδειγμα συνόλου πρόβλεψης των δεδομένων του και της εγγραφής των μετα-δεδομένων του. [9] ..... | 20 |
| Εικόνα 3: Παράδειγμα αντιστοίχισης πολύπλοκων και απλών αντικειμένων στο OWL. [9] .....                     | 23 |
| Εικόνα 4: Παράδειγμα χαρτογράφησης που βασίζεται στην κατανόηση του πλαισίου. [9] .....                     | 24 |
| Εικόνα 5: Παράδειγμα χαρτογράφησης που βασίζεται στην κατανόηση του πλαισίου. [9] .....                     | 31 |
| Εικόνα 6: Δείγμα ερωτήσεων ικανότητας [6] .....                                                             | 37 |
| Εικόνα 7: Κύριες Ιστορίες που εξήγαγαν οι Masmoudi, Karray, Lamine, Zghal και Archimede. [11] .....         | 38 |
| Εικόνα 8: Μερική προβολή ενοτήτων της οντολογίας MEMOn. [11] .....                                          | 41 |
| Εικόνα 9: Το τμήμα του MEMOn που χρησιμοποιείται για τον ορισμό του τυφώνα 5 κατηγοριών. [11] .....         | 42 |
| Εικόνα 10: Εξίσωση πιθανότητας εμφάνισης λέξης σε ένα έγγραφο. [46] .....                                   | 47 |
| Εικόνα 11: Αναπαράσταση του μοντέλου LDA. [46] .....                                                        | 47 |
| Εικόνα 12: Διαδικασία κατασκευής αρχικής οντολογίας. [12] .....                                             | 50 |
| Εικόνα 13: Εξίσωση συνάφειας. [12] .....                                                                    | 51 |



## Εισαγωγή

Η έλλειψη πρόγνωσης των καιρικών φαινομένων για τον προγραμματισμό εργασιών και αναγκών (αγροτικές καλλιέργειες, πτήσεις αεροσκαφών, διαδρομές πλοίων κλπ.), αποτελεί έναν από τους πολλούς λόγους για την ανάγκη ύπαρξης οντολογιών στις μέρες μας. Συγκεκριμένα, μεγάλα και όχι τόσο ξεκάθαρα δεδομένα του μετεωρολογικού πεδίου απαιτούν την ύπαρξη οντολογίας, για την απλοποίηση, μετάφραση και μελέτη όλων αυτών των δεδομένων.

Στις αρχές του 20<sup>ου</sup> αιώνα, δεν υπήρχαν τα κατάλληλα επιστημονικά εργαλεία τα οποία θα έδιναν ακριβή δεδομένα για τη σωστή πρόβλεψη του καιρού. Γι' αυτό, η πρόγνωση ήταν μια διαδικασία η οποία ήταν αναξιόπιστη και ανακριβής. Συγκεκριμένα, ήταν ακατόρθωτο να γίνουν επιτυχημένες παρατηρήσεις καθώς αυτός που επιχειρούσε την πρόγνωση του καιρού βασιζόταν κυρίως στη διαίσθηση του παρά σε ακριβή επιστημονικά δεδομένα. Έτσι, η οργάνωση των εργασιών που γινόταν σε εξωτερικό χώρο όπως, αγροτικές καλλιέργειες, οικοδομικές εργασίες, αλιεία κ.α., ήταν πολύ δύσκολο να ολοκληρωθούν όλες αυτές οι εργασίες με επιτυχία διότι δεν υπήρχε ακριβής πρόβλεψη του καιρού ώστε να γίνουν οι εργασίες αυτές ομαλά χωρίς παρεμβολές του καιρού που θα προκαλούσαν καθυστέρηση, δυσκολία ή ακόμα και καταστροφές.

Με βάση τις ανάγκες που ανέκυψαν, έγινε κατανοητή η ανάγκη ύπαρξης κάποιας μεθόδου ώστε να υπάρχει πιο ακριβής πρόβλεψη των καιρικών συνθηκών. Αν και πλέον υπάρχουν υπολογιστικά συστήματα και εργαλεία που προβλέπουν τα μετεωρολογικά δεδομένα, πολλές φορές δεν είναι 100% ακριβή.

Η περιβαλλοντολογική κρίση, η οποία έχει πυροδοτήσει αλλαγή στη συμπεριφορά του καιρού με την εμφάνιση επικίνδυνων καιρικών φαινομένων, καθιστούν επιτακτική την ανάγκη εργαλείων τα οποία θα βοηθούν στην πρόγνωση των μετεωρολογικών εξελίξεων όσο πιο στοχευμένα γίνεται.

Η διαρκής εξέλιξη της τεχνολογίας καθώς και των υπολογιστικών συστημάτων στις μέρες μας, μπορούν να παρέχουν καλύτερα αποτελέσματα μέσω της εξαγωγής όλων των δεδομένων που συνδέονται με την πρόγνωση του καιρού. Έτσι, είναι πλέον εύκολο να γίνεται πρόβλεψη πολλών παραγόντων που επηρεάζουν τον καιρό (βροχή, ομίχλη, υγρασία, μποφόρ του ανέμου, κίνηση του ανέμου κ.α.) εβδομαδιαία και αρκετά αξιόπιστα μηνιαία. Τα τελευταία 70 χρόνια, συστήματα

όπως SSN (Semantic Sensor Network) και SOSA (Sensor Observation Sample and Actuator, διευκολύνουν τη ζωή των ανθρώπων καθώς συμβάλλουν στις μετεωρολογικές προβλέψεις με αποτέλεσμα να είναι επιτυχημένες σε πολύ μεγάλο βαθμό.

### Σκοπός

Ο σκοπός της παρούσας πτυχιακής εργασίας, είναι η σύγκριση προτάσεων που υπάρχουν για οντολογίες πάνω σε μετεωρολογικά δεδομένα καθώς και ο σχολιασμός των συμπερασμάτων μεταξύ των διαφορετικών οντολογιών που χρησιμοποιήθηκαν.

## Μέρος Α

### 1.1 Οντολογίες

Σύμφωνα με τον Gruber, οντολογίες είναι ρητές προδιαγραφές μιας έννοιας. Η λέξη οντολογία προέρχεται από τη φιλοσοφία, όπου οντολογία είναι μια συστηματική περιγραφή της ύπαρξης ενός συγκεκριμένου πεδίου. Τα συστήματα που έχουν ως βάση τη γνώση, αυτό που υφίσταται είναι ακριβώς αυτό που μπορεί να αναπαρασταθεί. Το σύνολο αντικειμένων καθώς και οι μεταξύ τους σχέσεις αντικατοπτρίζονται στο αναπαραστατικό λεξιλόγιο, το οποίο αντιπροσωπεύει τη γνώση, βασίζεται σε αυτήν. Βάση των παραπάνω, μπορούμε να περιγράψουμε την οντολογία ενός προγράμματος ορίζοντας ένα σύνολο αναπαραστατικών όρων. Οι οντολογίες χρησιμοποιούν ονόματα οντοτήτων από το σύμπαν του λόγου κάνοντας τα κατανοητά προς τον άνθρωπο [1].

Στον τομέα της πληροφορικής η οντολογία συσχετίζεται με ορισμό συμφωνημένης και κοινής μορφοποίησης, η οποία ασχολείται με ένα συγκεκριμένο πεδίο ενδιαφέροντος. Μπορούν να χρησιμοποιηθούν ως ένα σύνολο εννοιών, σχέσεων και ιδιοτήτων το οποίο μπορεί να παρέχει εξαγωγή συμπερασμάτων σε ένα δυσνόητο σύστημα.

Σύμφωνα με τους Κώτη, Βούρο και Στεργίου, οι οντολογίες έγιναν ευκάτοπτες για τη διαμόρφωση και την επεξεργασία των πληροφοριών καθώς και πώς μπορούν να διαχειριστούν αυτές με σκοπό την εξέλιξη του Σημασιολογικού Ιστού καθώς και των εφαρμογών του. Σε ένα παρόμοιο πεδίο, οι οντολογίες αναπαράγουν ένα λεξιλόγιο το οποίο είναι κοινό για τα μέλη της κοινότητας με σκοπό τη διασύνδεση, τον συνδυασμό καθώς και την επικοινωνία της γνώσης. Ακόμη, είναι πολύ πιθανό να υπάρχουν περισσότερες από μία οντολογίες για ένα συγκεκριμένο πεδίο. Σε μία τέτοια περίπτωση όπου έχουμε πολλές οντολογίες θα πρέπει να υπάρξει αποτελεσματική γεφύρωση των κενών μεταξύ τους ώστε να μην υπάρχουν κενά και συγκρούσεις. Επομένως, είναι απαραίτητος ο συντονισμός μεταξύ των οντολογιών του συγκεκριμένου πεδίου (Χαρτογράφηση, Ευθυγράμμιση και Συγχώνευση) ώστε να υπάρχει ομαλή λειτουργία μεταξύ του λογισμικού και του ανθρώπου [2].

Σύμφωνα με τους Andrew Le Clair και Ridha Khedri, υπάρχουν τρία σημαντικά προβλήματα τα οποία δημιουργήθηκαν από τον μεγάλο όγκο και τον πολλαπλασιασμό των δεδομένων.

1. Οι πηγές δεδομένων αναδιαμορφώνονται λόγω διορθωτικής συντήρησης.
2. Τα δεδομένα αποθηκεύονται σε διαφορετικές πηγές.
3. Οι πηγές δεδομένων χάνουν την ανεξαρτησία εφαρμογής τους.

Αυτά τα προβλήματα έρχονται να αντιμετωπίσουν οι οντολογίες οι οποίες δημιουργούν ένα οντολογικό επίπεδο στις πηγές δεδομένων [3].

Υπάρχουν πέντε τύποι οντολογιών:

1. Οι οντολογίες τομέα (Domain Ontologies), στις οποίες εμπεριέχεται γνώση για ένα συγκεκριμένο τομέα (π.χ Ιατρική, Μηχανική, Πληροφορική, Μετεωρολογία)
2. Οι οντολογίες μετα-δεδομένων (MetaData Ontologies), οι οποίες παρέχουν ένα λεξιλόγιο για την περιγραφή της πληροφορίας στον παγκόσμιο ιστό.
3. Οι γενικές ή ανώτατου επιπέδου οντολογίες (Genericor Top-level ontologies ή Common Sense Ontologies), οι οποίες συλλέγουν τη γενική γνώση. Οι εννοιολογήσεις που ορίζονται εδώ, έχουν ισχύ και σε άλλους τομείς.
4. Οι οντολογίες αναπαράστασης (Representational Ontologies), οι οποίες παρέχουν αναπαραστατικές οντότητες στις οποίες δεν γίνεται συγκεκριμένη αναπαράσταση. Έτσι, δεν δεσμεύονται σε κάποιο συγκεκριμένο τομέα.
5. Οι οντολογίες μεθόδων και εργασιών (Method and Task Ontologies), οι οποίες περιγράφουν όρους για συγκεκριμένες μεθόδους επίλυσης προβλημάτων. Μπορούν να χρησιμοποιηθούν για συλλογής της γνώσης σε κάποιο τομέα [4].

Παρακάτω παρουσιάζονται έννοιες οι οποίες θα χρησιμοποιηθούν για την ανάγκη αυτής της εργασίας

## 1.2 Resource Description Framework (RDF)

Σύμφωνα με τον Santosh, το RDF είναι μία γλώσσα η οποία περιγράφει πόρους. Το βασικό δομικό στοιχείο του RDF είναι μία δήλωση, μια τριάδα η οποία αποτελείται από μία οντότητα (που ονομάζεται πόρος στην ορολογία του Ιστού), μία ιδιότητα και μία τιμή (η οποία τιμή μπορεί να είναι ένας άλλος πόρος). Το RDF είναι ανεξάρτητο από τον τομέα, καθώς δεν κάνει καμία υπόθεση για έναν συγκεκριμένο τομέα χρήσης [5].

Τα RDF σχήματα δεν στοχεύουν σε θεωρητικά ζητήματα, αλλά στην επίλυση ενός μικρού αριθμού άμεσων προβλημάτων. Οι δημιουργοί του θεωρούν ότι και άλλα προβλήματα θα έχουν παρόμοια χαρακτηριστικά.

### 1.3 Extensible Markup Language (XML)

Η XML, είναι μια γλώσσα σήμανσης που ορίζει ένα σύνολο κανόνων για την κωδικοποίηση εγγράφων σε μορφή που είναι αναγνώσιμη τόσο από τον άνθρωπο, όσο και από μία μηχανή. Οι προδιαγραφές της XML είναι ένα δωρεάν ανοιχτό πρότυπο.

Οι σχεδιαστικοί στόχοι της XML δίνουν έμφαση στην απλότητα, τη γενικότητα και τη χρηστικότητα στο διαδίκτυο. Είναι μία μορφή δεδομένων κειμένου, με ισχυρή υποστήριξη κωδικοποίησης για διαφορετικές ανθρώπινες γλώσσες. Αν και ο σχεδιασμός της XML εστιάζει σε έγγραφα, η γλώσσα χρησιμοποιείται ευρέως για την αναπαράσταση αυθαίρετων δομών δεδομένων, όπως αυτές που χρησιμοποιούνται σε υπηρεσίες του διαδικτύου. Υπάρχουν πολλά συστήματα σχημάτων για τον ορισμό γλωσσών που βασίζονται σε XML, ενώ οι προγραμματιστές έχουν αναπτύξει πολλές διεπαφές προγραμματισμού εφαρμογών (API) για να βοηθήσουν στην επεξεργασία δεδομένων. Η XML έχει πολύ συγκεκριμένη και αυστηρή δομή, ώστε να γίνει αναγνώσιμη από μία μηχανή.

### 1.4 OWL Time Ontology

Οι χρονολογικές πληροφορίες είναι σημαντικές στις περισσότερες εφαρμογές του πραγματικού κόσμου. Για παράδειγμα, η ημερομηνία είναι πάντα μέρος μιας ηλεκτρονικής παραγγελίας. Τα γεγονότα στον κόσμο συμβαίνουν σε συγκεκριμένες χρονικές στιγμές και συνήθως έχουν πεπερασμένη διάρκεια. Οι συναλλαγές πραγματοποιούνται με μία σειρά, με την τρέχουσα κατάσταση ενός συστήματος να εξαρτάται από το ακριβές ιστορικό όλων των συναλλαγών. Η γνώση των χρονικών σχέσεων μεταξύ συναλλαγών, γεγονότων, ταξιδιών και παραγγελιών είναι συχνά κρίσιμη.

Σύμφωνα με τους Cox και Little, το OWL-Time αναπτύχθηκε ως απάντηση σε αυτήν την ανάγκη, για την περιγραφή των χρονικών ιδιοτήτων οποιουδήποτε πόρου που δηλώνεται χρησιμοποιώντας ένα αναγνωριστικό ιστού (URI), συμπεριλαμβανομένων ιστοσελίδων και πραγμάτων πραγματικού κόσμου, εάν είναι επιθυμητό. Εστιάζει ιδιαίτερα στις σχέσεις χρονικής διάταξης [6].

### 1.5 Gloze Set of Tools

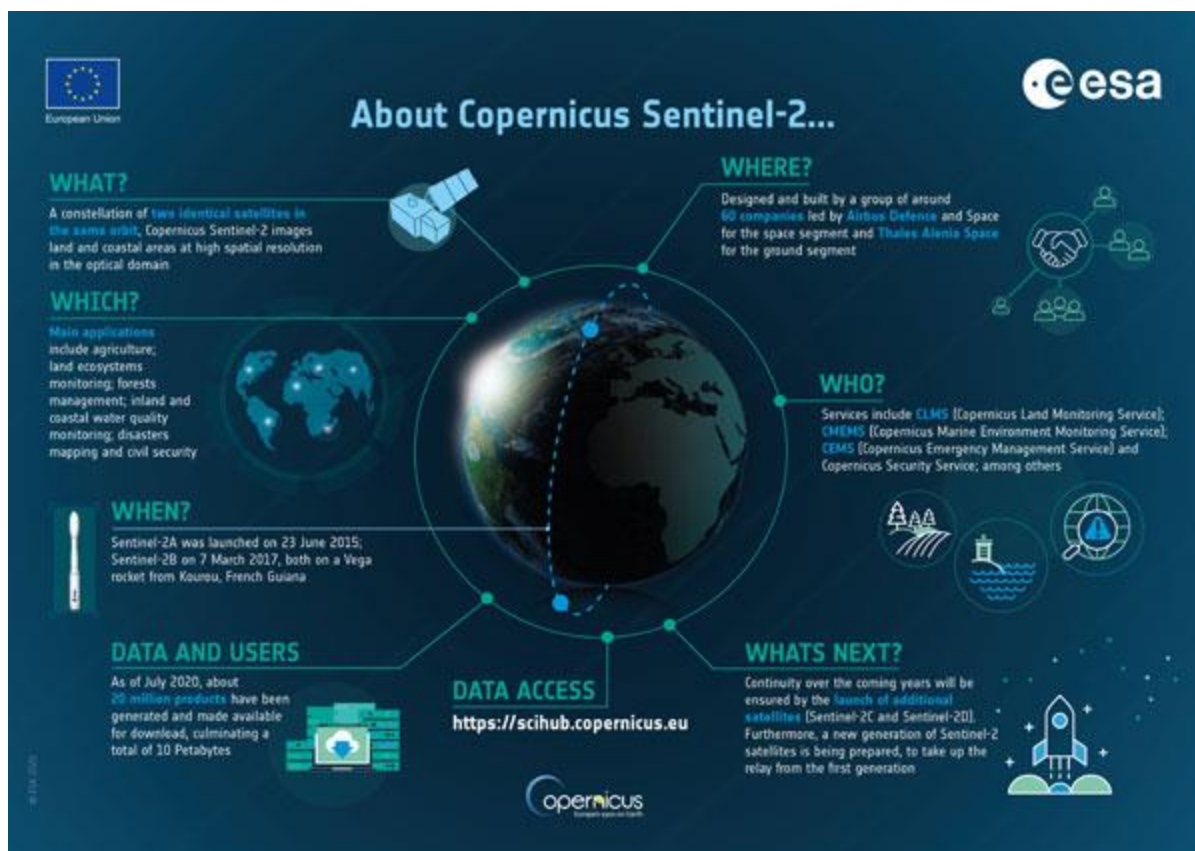
Το Gloze είναι ένα σύνολο εργαλείων τα οποία μας επιτρέπουν να ερμηνεύσουμε άμεσα ένα XML αρχείο ως μοντέλο RDF χωρίς να χρειαστεί να περάσουμε από RDF/XML. Το Gloze, χρησιμοποιεί υπάρχουσες δομές μοντελοποίησης RDF, ιδιαίτερα την πολύ κακοποιημένη ακολουθία RDF, σε συνδυασμό με μία καθορισμένη αντιστοίχιση που βασίζεται στο σχήμα XML. Επίσης, χρησιμοποιεί το σχήμα XML ως βάση για την περιγραφή του τρόπου με τον οποίο ένα XML αντιστοιχίζεται σε ένα RDF. Η συγκεκριμένη προσέγγιση έχει το πλεονέκτημα ότι τα XML είναι δηλωτικά καθώς και ουδέτερα ως προς την κατεύθυνση της αντιστοίχισης.

### 1.6 Vortex2

Το Vortex2 (Verification of the Origins of Rotation in Tornadoes Experiment 2), είναι το μεγαλύτερο ερευνητικό έργο για ανεμοστρόβιλους, το οποίο ασχολείται με το πώς, πότε και γιατί σχηματίζονται οι ανεμοστρόβιλοι. Η Εθνική Υπηρεσία Ωκεανών και Ατμόσφαιρας (National Oceanic and Atmospheric Administration – NOAA), διαθέτει περισσότερους από 100 επιστήμονες, καθώς και φοιτητές παγκόσμια, που έχουν ως κύριο αντικείμενο τη συλλογή μετρήσεων μετεωρολογικών δεδομένων. Διαθέτουν 10 κινητά ραντάρ καθώς και άλλα 70 όργανα εξοπλισμένα με τεχνολογίες επικοινωνίας.

### 1.7 Πρόγραμμα Copernicus Sentinel-2

Το πρόγραμμα Copernicus Sentinel-2, αποτελείται από έναν αστερισμό δύο δορυφόρων σε πολική τροχιά, οι οποίοι τοποθετούνται στην ίδια τροχιά, με απόσταση 180 μοιρών μεταξύ τους. Σκοπός τους είναι η παρακολούθηση των αλλαγών στις συνθήκες της επιφάνειας της γης. Παρέχει μία περιγραφή υψηλού επιπέδου των στόχων της αποστολής, την περιγραφή του δορυφόρου καθώς και το τμήμα εδάφους.



Εικόνα 1: Πρόγραμμα CopernicusSentinel-2[7]

## 1.8 SERVIR-Viz 3-D Visualization Tool

Το SERVIR-Viz, είναι μία προσαρμοσμένη έκδοση δωρεάν ανοιχτού λογισμικού τρισδιάστατου εργαλείου εξερεύνησης της γης της NASA, γνωστό και ως WorldWind. Όπως και με άλλα προγράμματα χαρτογράφησης όπως το GoogleEarth, το SERVIR-Viz, επιτρέπει στους χρήστες να κάνουν ζουμ σε οποιοδήποτε μέρος της Γης και να προσαρμόζουν τη γωνία θέασης ώστε να προσθέσουν εάν επιθυμούν, ένα τρισδιάστατο έδαφος [8].

## Μέρος Β

Στην παρούσα ενότητα θα γίνει η παρουσίαση των οντολογιών που χρησιμοποιήθηκαν σε εργασίες που έχουν παρουσιασθεί. Οι οντολογίες που θα παρουσιαστούν είναι οι εξής:

1. Η οντολογία που περιγράφεται στην εργασία των Chen και Plale του 2013, είναι μία οντολογία η οποία βασίζεται στην οντολογία μετα-δεδομένων, στην οποία αντιπροσωπεύονται σύνολα δεδομένων γεωεπιστήμης [9].
2. Η οντολογία που περιγράφεται στην εργασία των Chen και Huang του 2014, βασίζεται στην οντολογία τομέα. Συγκεκριμένα γίνεται μία αυτόματη δημιουργία σύνοψης ειδήσεων για τα μετεωρολογικά δεδομένα με βάση τις τεχνικές ασαφούς συλλογισμού και οντολογίας [10].
3. Η οντολογία που περιγράφεται στην εργασία των Masmoudi, Karray, Lamine, Zghal και Archimede του 2020, αφορά μία αρθρωτή οντολογία (MEMOn) για την παρακολούθηση του περιβάλλοντος, η οποία δημιουργήθηκε με βάση μία πρωτότυπη ευέλικτη μεθοδολογία [11].
4. Η οντολογία που περιγράφεται στην εργασία των J.Liu, Dong, Z.Liu και Chen του 2022 είναι αυτοδημιουργητή οντολογία από τους συγγραφείς, η οποία βασίζεται στην οντολογία τομέα [12].

Στη αρχή θα παρουσιαστούν και στην πορεία θα σχολιαστούν και τέλος θα συγκριθούν.

### Εργασία 1<sup>η</sup>

Στην εργασία τους οι Chen και Plale, χρησιμοποιούν τη λογική της διεξαγωγής, η οποία βασίζεται στο ότι με τον αυξανόμενο όγκο δεδομένων που παράγονται στην έρευνα της γεωεπιστήμης, καθίσταται κρίσιμο να περιγραφούν σύνολα δεδομένων με ουσιαστικούς τρόπους. Ένας μεγάλος αριθμός περιγραφόμενων συνόλων δεδομένων περιγράφεται χρησιμοποιώντας μετα-δεδομένα XML, τα οποία έχουν αποδειχθεί χρήσιμο μέσο έκφρασης χαρακτηριστικών δεδομένων. Μια οντολογική αναπαράσταση είναι ένας άλλος τρόπος αναπαράστασης συνόλων δεδομένων με το πλεονέκτημα της παροχής πλούσιας σημασιολογίας, εύκολης σύνδεσης με άλλα σύνολα δεδομένων και καλής διαλειτουργικότητας με άλλα δεδομένα.

Αυτή η εργασία, αντιπροσωπεύει σύνολα δεδομένων γεωεπιστήμης ως οντολογία που βασίζεται σε μία υπάρχουσα περιγραφή μετα-δεδομένων και στη φύση του συνόλου δεδομένων. Χρειάζεται η περίπτωση των δεδομένων Vortex2, ενός τοπικού συνόλου δεδομένων μετεωρολογικών προβλέψεων που συλλέχθηκε το καλοκαίρι του 2010, για να δείξει πώς τα δεδομένα πρόβλεψης μπορούν να αναπαρασταθούν στην οντολογία χρησιμοποιώντας τις υπάρχουσες πληροφορίες μετα-δεδομένων. Παρέχει έναν άλλο τύπο αναπαράστασης του συνόλου δεδομένων με πρόσθετη σημασιολογία και πιθανές λειτουργίες σε σύγκριση με την προηγούμενη αναπαράσταση μετα-δεδομένων.

Οι επιστημονικές δραστηριότητες δημιουργούν ένα πλήθος συνόλων δεδομένων, τα οποία δεν είναι μόνο χρήσιμα για τους δημιουργούν δεδομένων, αλλά έχουν επίσης σημαντική αξία για άλλους ερευνητές τώρα και στο μέλλον. Η περιγραφή ενός συνόλου επιστημονικών δεδομένων καθίσταται επομένως κρίσιμη για την αποτελεσματική κοινή χρήση και επαναχρησιμοποίησή του. Στις γεωεπιστήμες (επιστήμη της γης, ατμοσφαιρική επιστήμη, ωκεανογραφία κ.λπ.), τα σύνολα δεδομένων έχουν περιγραφεί από διαφορετικές γλώσσες, όπως περιγραφή επίπεδων αρχείων, εξαιρετικά μεγάλα ονόματα αρχείων καθώς και μετα-δεδομένα XML. Αυτές οι αναπαραστάσεις έχουν διαφορετικές αντιπροσωπευτικές δυνάμεις και έχουν παίξει σημαντικό ρόλο στο πλαίσιο της χρήσης τους. Τον τελευταίο καιρό, όλο και περισσότεροι πόροι, συμπεριλαμβανομένων των δεδομένων, αντιπροσωπεύονται από οντολογίες στον ιστό, χρησιμοποιώντας γλώσσες όπως η OWL και η RDF. Οι οντολογίες έχουν το πλεονέκτημα της περιγραφής πόρων σε σημασιολογικό επίπεδο και επομένως είναι ένας καλός τρόπος για περιγραφή συνόλων δεδομένων. Αυτή η μελέτη καταδεικνύει μια προσπάθεια αναπαράστασης συνόλων δεδομένων της γεωεπιστήμης χρησιμοποιώντας μια οντολογία, καθώς αυτή η πρακτική δεν είναι συνηθισμένη, ενώ έχει πολλά οφέλη.

#### Πλεονεκτήματα

Πρώτον, η αναπαράσταση οντολογίας εστιάζει στη σημασιολογία και προσφέρει πλούσια σημασιολογία για μηχανική κατανόηση. Η πλούσια σημασιολογία στην περιγραφή, π.χ. έννοιες και σχέσεις, μπορούν να βελτιώσουν την πρόσβαση στα δεδομένα, την κοινή χρήση και την επαναχρησιμοποίηση συνόλων δεδομένων.

Δεύτερον, η αναπαράσταση οντολογίας διευκολύνει την αλληλεπίδραση με άλλους πόρους στον σημασιολογικό ιστό. Για παράδειγμα, ένα σύνολο δεδομένων μπορεί εύκολα να συνδεθεί με ένα άλλο σύνολο δεδομένων, ένα άτομο, έναν οργανισμό ή μια συλλογή στον συνδεδεμένο χώρο δεδομένων. Τρίτον, οι γλώσσες οντολογίας όπως η OWL και η RDF διασφαλίζουν τη διαλειτουργικότητα μεταξύ περιγραφών διαφορετικών συνόλων δεδομένων, καθιστώντας έτσι δυνατή τη διαλειτουργικότητά τους.

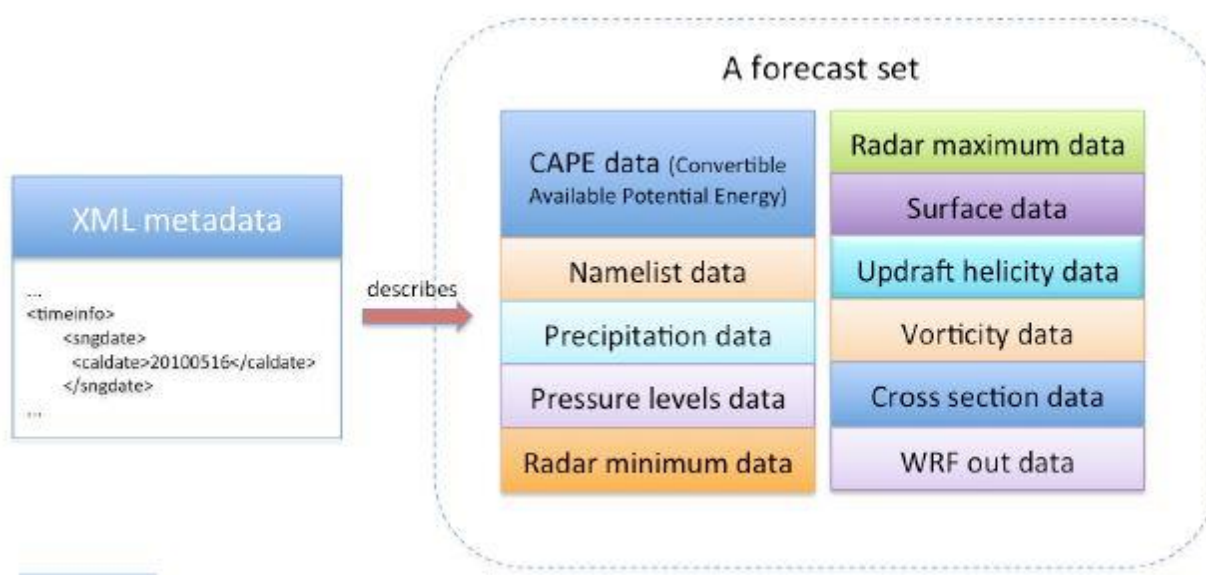
Δεδομένου ότι πολλά σύνολα δεδομένων γεωεπιστήμης αντιπροσωπεύονται ήδη από μετα-δεδομένα, σε αυτή τη μελέτη δεν ξεκινά κανείς από το μηδέν, αλλά ο ερευνητής εργάζεται για την ανάπτυξη μιας οντολογικής αναπαράστασης για σύνολα δεδομένων που βασίζονται σε υπάρχοντα μετα-δεδομένα. Προκειμένου να αναπτύξει μια γενικευμένη μεθοδολογία για τη μετατροπή των μετα-δεδομένων σε αναπαράσταση οντολογίας, επιλέγει να χρησιμοποιήσει ένα συγκεκριμένο σύνολο δεδομένων, ένα σύνολο δεδομένων πρόβλεψης καιρού με περιγραφή μετα-δεδομένων, ως δοκιμαστικό πυλώνα για την επίδειξη της διαδικασίας. Η μεθοδολογία και η εφαρμογή μπορεί να ρίξουν φως σε παρόμοια πρακτική αλλού.

### Μεθοδολογία

Το έργο Vortex2 ήταν μια επιτόπια εκστρατεία διάρκειας 6 εβδομάδων που χρηματοδοτήθηκε από την NSF που πραγματοποιήθηκε την άνοιξη του 2010 για τη συλλογή δεδομένων σχετικά με ισχυρές καταιγίδες στο μεσαίο τμήμα των Ηνωμένων Πολιτειών. Το έργο LEADII στο Πανεπιστήμιο της Ιντιάνα συνέβαλε στην επιτόπια εκστρατεία δημιουργώντας 5 βραχυπρόθεσμες περιφερειακές προβλέψεις κάθε πρωί της εκστρατείας πεδίου.

Αυτές οι βραχυπρόθεσμες προβλέψεις, διάρκειας 10-25 ωρών και εκτελούνται σε μια περιοχή περίπου στο μέγεθος μιας πολιτείας των ΗΠΑ, χρησιμοποιήθηκαν για να κατανοήσουν τα καιρικά μοτίβα της ημέρας για μια τοποθεσία. Τα αποτελέσματα του μοντέλου πρόβλεψης μετατράπηκαν σε εικόνες και ταινίες που στη συνέχεια έγιναν άμεσα διαθέσιμες σε έναν ιστότοπο για προβολή σε κινητό τηλέφωνο από ερευνητές πεδίου. Σε μια συγκεκριμένη ημέρα πρόβλεψης, το σύστημα LEADII θα ανακτούσε τον πιο πρόσφατο καιρό που παρατηρήθηκε για να χρησιμοποιήσει στην πρόγνωση, καθιστώντας τις προβλέψεις πιο ακριβείς. Κατά τη διάρκεια 6 εβδομάδων, η ομάδα LEADII δημιούργησε 175 προγνώσεις καιρού και πάνω από 9000 προϊόντα γραφικών (εικόνες

και ταινίες). Αυτά τα δεδομένα επιμελήθηκαν προσεκτικά χρησιμοποιώντας το σχήμα μετα-δεδομένων FGDC (Ομοσπονδιακή Επιτροπή Γεωγραφικών Δεδομένων). Κάθε πρόβλεψη, τα παράγωγα προϊόντα της και τα μετα-δεδομένα ομαδοποιήθηκαν σε μια ενιαία δέσμη προβλέψεων. Μια πρόβλεψη καιρού δημιουργεί πολλούς τύπους δεδομένων, 11 τύπων συνολικά, συμπεριλαμβανομένης της εξόδου του ίδιου του μοντέλου WRF (Weather Research and Forecast) και εικόνων και ταινιών συγκεκριμένων χαρακτηριστικών (π.χ. ισχυρή καταιγίδα και έχουν μεγάλη σημασία για την ανάλυση των προβλέψεων). Το σχήμα μετα-δεδομένων FGDC έχει σχεδιαστεί ειδικά για γεωχωρικά δεδομένα, αποτελούμενο από 7 γεωχωρικές πτυχές που καλύπτουν 334 στοιχεία.



Εικόνα 2: Παράδειγμα συνόλου πρόβλεψης των δεδομένων του και της εγγραφής των μετα-δεδομένων του. [9]

Η όλη προσπάθεια κατέληξε σε 175 δέσμες προβλέψεων. Κάτω από κάθε δέσμη προβλέψεων, υπάρχουν εικόνες και δεδομένα εξόδου του μοντέλου. Εδώ ονομάζονται τα δεδομένα από μία πρόβλεψη "ένα σύνολο πρόβλεψης" και τα συγκεκριμένα δεδομένα εικόνας/εξόδου εντός ενός συνόλου πρόβλεψης "υποσύνολο πρόβλεψης". Για παράδειγμα, για μια εκτέλεση πρόβλεψης στις 14:00:00 Z της ημέρας 2010/05/16, έχουμε γενικό σύνολο προβλέψεων και υποσύνολα πρόβλεψης για έξοδο μοντέλου WRF, εικόνες βροχόπτωσης και ούτω καθεξής.

Κάθε σύνολο πρόβλεψης έχει μια περιγραφή μετα-δεδομένων XML που ορίζεται από το σχήμα FGDC. Τα μετα-δεδομένα χρησιμοποιούν 3 γεωχωρικές πτυχές και 59 στοιχεία από το FGDC. Τα

σύνολα πρόβλεψης και τα αρχεία μετα-δεδομένων αντιστοιχίζονται ένα προς ένα, επομένως έχουμε 175 αρχεία μετα-δεδομένων συνολικά ως βάση για τη μετατροπή σε αναπαράσταση οντολογίας.

#### Παρόμοιες μελέτες

Υπήρξαν προηγούμενες μελέτες από τη συγκεκριμένη που παρουσίαζαν προσεγγίσεις για τη μετατροπή σχημάτων μετα-δεδομένων σε οντολογίες. Αυτές οι προσεγγίσεις είναι γενικές προσεγγίσεις που ασχολούνται με το πρόβλημα ενός δεδομένου εγγράφου μετα-δεδομένων (π.χ. αρχείο XML) πώς μπορεί να μετατραπεί αυτόματα σε αναπαράσταση οντολογίας (π.χ. αρχείο OWL ή RDF). Για παράδειγμα, ζητήματα που αντιμετωπίζονται συχνά στη μετατροπή είναι ο τρόπος με τον οποίο τα στοιχεία XML, όπως οι ακολουθίες XML, τα στοιχεία/χαρακτηριστικά, τα αναγνωριστικά, θα πρέπει να αντιπροσωπεύονται στη γλώσσα RDF. Οι γενικές προσεγγίσεις συνήθως μετατρέπουν εγγραφές μετα-δεδομένων σε συντακτικό επίπεδο, με την έννοια ότι εξετάζουν τη χαρτογράφηση μεταξύ δομικών στοιχείων XML και RDF.

Στο πλαίσιο της ίδιας έρευνας μελετώνται και προγενέστερες πάνω σε παρόμοια ζητήματα, που είτε γίνονται αποδεκτές είτε απορρίπτονται. Η γενική προσέγγιση των Ferdinand, Zirpins και Trastour 2005 [13], περιλαμβάνει δύο τύπους χαρτογράφησης: αντιστοίχιση από XML σε RDF (π.χ. δημιουργία στοιχείων/χαρακτηριστικού XML ως ιδιότητες RDF) και αντιστοίχιση από σχήμα XML σε OWL (π.χ. δημιουργία σχήματος XML σύμπλοκο τύπος ως OWL τάξη). Ο Battle στην εργασία του το 2006 [14], παρουσιάζει την εργαλειοθήκη Gloze, η οποία αντιστοιχεί το XML σε RDF και το RDF πίσω σε XML με βάση το σχήμα XML. Το επιχείρημα είναι ότι δεν υπάρχει ανάγκη να δημιουργηθεί μια γλώσσα αντιστοίχισης, αλλά αντίθετα χρησιμοποιεί το σχήμα XML ως γλώσσα για την περιγραφή σημαντικών πληροφοριών χαρτογράφησης κατά τη μετάφραση μεταξύ μορφών XML και RDF. Τα χαρακτηριστικά XML αντιστοιχίζονται σε ιδιότητες δεδομένων/αντικειμένων στο RDF και οι ακολουθίες XML αντιμετωπίζονται από μια λύση που βασίζεται σε ακολουθίες RDF. Οι Van Deursen, Poppe, Martens και Vande Walle το 2008 [15], προτείνουν μια άλλη γενική προσέγγιση, η οποία χρησιμοποιεί ένα αρχείο αντιστοίχισης για τη δημιουργία σύνδεσης μεταξύ του σχήματος XML και της οντολογίας OWL και βοηθά περαιτέρω στη μετατροπή των εγγραφών XML σε στιγμιότυπα RDF. Οι Martens, Verborgh, Poppe και Vande Walle το 2011 [16], εισάγουν μια διαδικασία για τη μετατροπή ενός

σχήματος μετα-δεδομένων που βασίζεται σε XML σε μια οντολογία OWL χρησιμοποιώντας το παράδειγμα του DIG35, ενός προτύπου μετα-δεδομένων για την περιγραφή στατικών εικόνων. Αυτοματοποιούν τη διαδικασία μετατροπής χρησιμοποιώντας ένα εργαλείο που αντιστοιχίζει το XML σε RDF με βάση ένα έγγραφο XML.

Αυτές οι σχετικές προσπάθειες επικεντρώνονται στο συντακτικό μέρος της μετατροπής, ενώ η μελέτη που εξετάζουμε ασχολείται με το σημασιολογικό μέρος. Πιστεύεται ότι η μετατροπή δεν περιλαμβάνει μόνο συντακτική αντιστοίχιση μεταξύ XML και RDF, αλλά και ρητή σημασιολογική αναπαράσταση αντικειμένων που περιέχονται σε εγγραφές XML. Επομένως, η μετατροπή εξαρτάται επίσης από το περιβάλλον, δηλαδή, η προσέγγιση μετατροπής βασίζεται στο πλαίσιο των περιγραφόμενων οντοτήτων. Οι μετατροπές των μετα-δεδομένων των γεωεπιστημών σε οντολογίες μπορούν να κάνουν χρήση αυτών των γενικών προσεγγίσεων, και εν τω μεταξύ εξετάζονται ζητήματα μετατροπής ειδικά για το πλαίσιο του συγκεκριμένου συνόλου δεδομένων.

### Πρόταση

Προτείνεται μια προσέγγιση στη μετατροπή που χρησιμοποιεί αντιστοίχιση σε δύο επίπεδα: το επίπεδο έννοιας και το επίπεδο παρουσίας. Το επίπεδο έννοιας ασχολείται με έννοιες (κλάσεις στην οντολογία) που σχετίζονται με το σύνολο δεδομένων και το επίπεδο στιγμιότυπου αντιμετωπίζει περιπτώσεις εννοιών που εμφανίζονται στα αρχεία μετα-δεδομένων. Η προαναφερθείσα γενική προσέγγιση ρίχνει φως στη χαρτογράφηση επιπέδου έννοιας, για παράδειγμα, αντιστοίχιση από στοιχεία/χαρακτηριστικά XML σε ιδιότητες OWL. Επιπλέον, το πλαίσιο των δεδομένων πρόβλεψης αναλύεται για να παρέχει σημασιολογικά σημαντικές κλάσεις και ιδιότητες στην οντολογία. Η αντιστοίχιση επιπέδου στιγμιότυπου βασίζεται στη δομή του επιπέδου έννοιας και εξάγει χρήσιμα αποσπάσματα μετα-δεδομένων και τιμές ως στιγμιότυπα στην οντολογία που προκύπτει. Παρουσιάζεται η μεθοδολογία μετατροπής σε βήματα όπως παρακάτω.

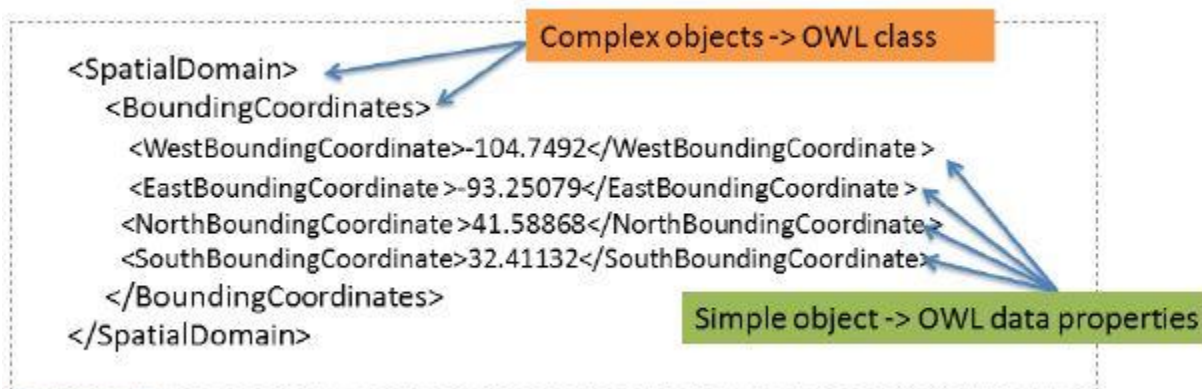
#### *1<sup>ο</sup> Προσδιορισμός για τις κύριες έννοιες/κλάσεις στο πλαίσιο*

Οι έννοιες του συνόλου πρόβλεψης εξαρτώνται από το περιβάλλον, και επομένως εξετάστηκε το περιβάλλον Vortex2 και οι εγγραφές μετα-δεδομένων XML για να βρεθούν οι κύριες έννοιες. Οι προσδιορισμένες έννοιες χρησιμεύουν ως κλάσεις στην αναπαράσταση οντολογίας. Δεδομένου

ότι το σύνολο δεδομένων Vortex2 είναι η κεντρική οντότητα στη μελέτη, προσδιορίζουν την έννοια «σύνολο δεδομένων» ως κρίσιμη έννοια για την οντολογία. Αντίστοιχα, κατασκευάζεται ως κλάση "Dataset" στην οντολογία OWL. Καθώς το σύνολο δεδομένων Vortex2 και τα υποσύνολά του (π.χ. δεδομένα βροχοπτώσεων, δεδομένα στροβιλισμού) ανήκουν στην κατηγορία "Σύνολο δεδομένων" εννοιολογικά, αποτελούν υποκατηγορίες του.

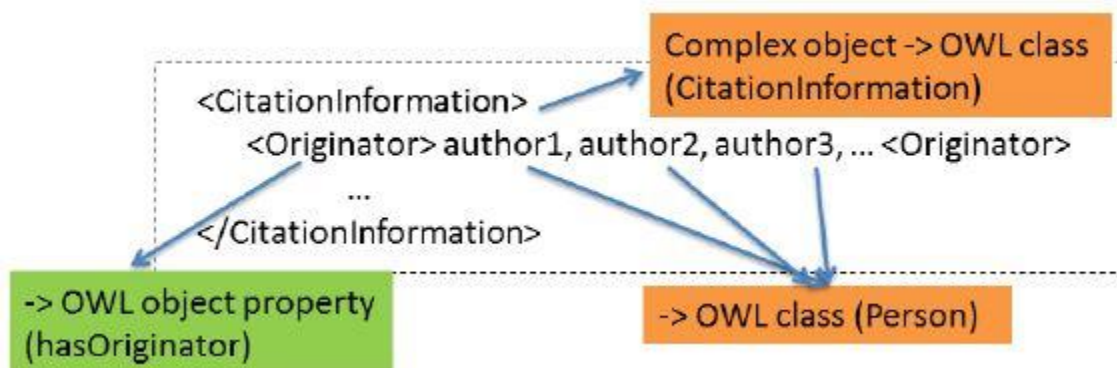
## 2<sup>ο</sup> Προσδιορισμός άλλων χρήσιμων τάξεων

Άλλες σημαντικές κλάσεις προσδιορίζονται από το σχήμα μετα-δεδομένων και το περιβάλλον Vortex2 επίσης. Αυτές οι κλάσεις διευκολύνουν κυρίως περαιτέρω την αναπαράσταση των συνόλων δεδομένων. Συνολικά, τα μετα-δεδομένα περιγράφουν δεδομένα Vortex2 από τρεις πτυχές χρησιμοποιώντας το λεξιλόγιο FGDC: πληροφορίες αναγνώρισης, πληροφορίες χαρακτηριστικών οντοτήτων και πληροφορίες μετα-δεδομένων. Αυτές οι τρεις πτυχές είναι τρία κύρια στοιχεία στα μετα-δεδομένα XML και ενσωματώνουν πολλά επίπεδα υποστοιχείων. Αυτά θεωρούνται σημαντικά αντικείμενα και γίνονται κλάσεις στην οντολογία. Οι τρεις κλάσεις χρησιμοποιούνται για την περιγραφή δεδομένων Vortex2 στην οντολογία, για παράδειγμα, έχουμε μια τέτοια δήλωση στο OWL: `Vortex2DataIsDescribedBySomeIdentificationInformation` (που σημαίνει ότι το σύνολο δεδομένων Vortex2 περιγράφεται από πληροφορίες αναγνώρισης). Ακολουθούν έναν κανόνα για να προσδιορίσουν κλάσεις από τα μετα-δεδομένα XML: εάν ένα στοιχείο XML περιέχει υποστοιχεία αντί για τιμές δεδομένων, τότε ονομάζεται το σύνθετο αντικείμενο και γίνεται ως κλάση στην οντολογία, όπως οι τρεις κλάσεις που αναφέρονται παραπάνω. Εάν ένα στοιχείο XML περιέχει μόνο τιμές δεδομένων, τότε ονομάζεται απλό αντικείμενο και γίνεται ιδιότητα δεδομένων στην οντολογία.



Εικόνα 3: Παράδειγμα αντιστοίχισης πολύπλοκων και απλών αντικειμένων στο OWL. [9]

Στη συνέχεια τελειοποιείται η οντολογία με βάση την κατανόηση των ερευνητών για το πλαίσιο. Για παράδειγμα, ένα στοιχείο είναι `<Originator>` στο σχήμα XML και είναι ένα υποστοιχείο του `<CitationInformation>`. Αυτό είναι ένα απλό αντικείμενο επειδή περιέχει μόνο τιμές δεδομένων σχετικά με τα ονόματα των δημιουργών του συνόλου δεδομένων. Ακολουθώντας τον παραπάνω κανόνα, θα κάνουμε τον "αρχηγό" μια ιδιότητα δεδομένων και τις τιμές δεδομένων ως κυριολεκτικά στην οντολογία. Ωστόσο, δεδομένου ότι η τιμή δεδομένων `<Originator>` παραθέτει στην πραγματικότητα τα ονόματα των ατόμων που είναι δημιουργοί, είναι πιο λογικό να υπάρχει μια δήλωση στο OWL που να λέει ότι το `Vortex2Data` έχει `Originator` μόνο πρόσωπο (που σημαίνει ότι το σύνολο δεδομένων `Vortex2` έχει αρχικό στοιχείο, το οποίο ανήκει στην κλάση `Person`) της απλής αντιμετώπισης των τιμών δεδομένων ως κυριολεκτικά στη γλώσσα OWL. Συνεπώς, το `<Originator>` δεν γίνεται ιδιότητα δεδομένων στην οντολογία αλλά ως ιδιότητα αντικειμένου, επειδή ακολουθείται από μια κλάση (`Person`) στη δήλωση τώρα.



Εικόνα 4: Παράδειγμα χαρτογράφησης που βασίζεται στην κατανόηση του πλαισίου. [9]

### 3<sup>ο</sup> Περιγραφή των τάξεων

Η περιγραφή των τάξεων οντολογίας κατασκευάζεται από την εκμάθηση εγγραφών μετα-δεδομένων και το πλαίσιο επίσης. Αρχικά ακολουθείται ένας κανόνας για να περιγραφούν κλάσεις: για κλάσεις που είναι αρχικά ένα σύνθετο αντικείμενο του σχήματος XML, περιγράφονται με τα υποστοιχεία τους στο σχήμα XML. Για παράδειγμα, το `Vortex2Data` έχει τρία υποστοιχεία τα οποία αντιστοιχίζονται επίσης ως κλάσεις OWL, συμπεριλαμβανομένων πληροφοριών αναγνώρισης και δύο άλλων. Σε αυτήν την περίπτωση, γράφουμε τη δήλωση OWL ως `Vortex2 DataisDescribedBySomeIdentificationInformation` (που σημαίνει ότι το σύνολο δεδομένων `Vortex2` περιγράφεται από πληροφορίες αναγνώρισης) και οι δηλώσεις OWL για τις άλλες δύο κλάσεις είναι παρόμοιες.

Όπως παραπάνω, αναλύεται το πλαίσιο για να περιγραφούν καλύτερα οι κατηγορίες OWL. Για παράδειγμα, το Vortex2Data και τα δευτερεύοντα σύνολα δεδομένων του είναι όλες υποκλάσεις της κλάσης Dataset, επιπλέον, μαθαίνουμε ότι εννοιολογικά το Vortex2Data περιέχει επίσης υποσύνολα δεδομένων. Αυτό είναι προφανές σε έναν άνθρωπο, αλλά αυτή η δήλωση δεν έχει αναπαρασταθεί ρητά στην οντολογία. Ορίζεται η σχέση μεταξύ του Vortex2Data και των δευτερευόντων συνόλων δεδομένων του ως σχέση ολόκληρου μέρους. Οι ιδιότητες προσαρτώνται στην κλάση Vortex2Data για να αναπαραστήσουν τη σχέση μέρους-ολόκληρου: π.χ. Το Vortex2Data περιέχει ορισμένα CapeData (που σημαίνει ότι το σύνολο δεδομένων Vortex2 περιέχει δεδομένα Cape ως υποσύνολο), όπου περιέχει είναι η ιδιότητα αντικειμένου που υποδεικνύει τη σχέση ολόκληρου μέρους.

#### *4<sup>ον</sup> Χρήση εξωτερικών κλάσεων και ιδιοτήτων*

Χρησιμοποιούνται στο μέγιστο οι υπάρχουσες εξωτερικές κλάσεις και ιδιότητες για να αποφευχθεί η επανεφεύρεση του τροχού και να επιτραπεί η διαλειτουργικότητα με άλλα σύνολα δεδομένων. Εάν μια κλάση ή μια ιδιότητα στην οντολογία υπάρχει σε μια εξωτερική οντολογία, κατά προτίμηση μια συχνά χρησιμοποιούμενη, χρησιμοποιείται η κλάση/ιδιότητα αυτής της οντολογίας αντί να δημιουργηθεί άλλη. Εξετάζουν τις οντολογίες DC (Dublin Core), VIVO, SWEET (Semantic Web for Earth and Environmental Terminology), FOAF (Friend Of A Friend) για να βρουν πιθανές χρήσιμες κλάσεις και ιδιότητες. Για παράδειγμα, η κλάση CitationInformation έχει την ιδιότητα "έχει ημερομηνία δημοσίευσης", που σημαίνει ότι το CitationInformation έχει πληροφορίες ημερομηνίας δημοσίευσης του συνόλου δεδομένων και διαπιστώνεται ότι οι όροι DC έχουν μια ιδιότητα "έκδοση" σχετικά με την ημερομηνία δημοσίευσης και επομένως χρησιμοποιείται αυτό το DC ιδιότητα για την περιγραφή της κλάσης CitationInformation: CitationInformation έκδοση μόνο ημερομηνία Ωρα.

#### *5<sup>ον</sup> Εξαγωγή παρουσιών από αρχεία μετα-δεδομένων*

Τα πρώτα τέσσερα βήματα αφορούν τη μετατροπή στο επίπεδο έννοιας και αυτό το τελευταίο βήμα αφορά το επίπεδο παρουσίας. Οι πληροφορίες σχετικά με τις παρουσίες εξάγονται από αρχεία μετα-δεδομένων για να προστεθούν στιγμιότυπα στην αφηρημένη οντολογία. Η οντολογία OWL αφορά τον σκελετό των συνόλων δεδομένων και περιγράφει τα χαρακτηριστικά τους, ενώ

τα στιγμιότυπα είναι δείκτες πόρων προς τα πραγματικά αντικείμενα, όπως συγκεκριμένα σύνολα δεδομένων και συγκεκριμένο άτομο. Αντιπροσωπεύουμε τις περιπτώσεις χρησιμοποιώντας το RDF, το οποίο περιγράφει τους πόρους με τον τρόπο τριπλών.

Υπάρχουν 175 προγνώσεις καιρού και προκύπτουν οπτικά προϊόντα και έτσι έχουμε 175 περιπτώσεις Vortex2Data. Κάθε στιγμιότυπο Vortex2Data περιέχει 11 παρουσίες δευτερεύοντος συνόλου δεδομένων, των οποίων οι πληροφορίες μπορούν επίσης να εξαχθούν από τα αρχεία XML. Εκτός από τα στιγμιότυπα συνόλων δεδομένων, παρουσίες άλλων κλάσεων εξάγονται επίσης αυτόματα από το αρχείο XML και οργανώνονται στην οντολογία σύμφωνα με τη σχέση κλάσης τους.

Τα πρώτα τέσσερα βήματα της μεθοδολογίας είναι η βασική εργασία, στην οποία οι κλάσεις και οι ιδιότητες οντολογίας παράγονται και γράφονται στην οντολογία. Χρησιμοποιείται η εργαλειοθήκη Protégé για να κατασκευαστεί η οντολογία OWL. Το τελευταίο βήμα, η εξαγωγή στιγμιότυπου, γίνεται αυτόματα με την ενσωμάτωση της εργαλειοθήκης Jena, μιας εργαλειοθήκης για το χειρισμό RDF, OWL, κ.λπ., επειδή πληροφορίες στιγμιότυπων λαμβάνεται από αρχεία μετα-δεδομένων και μπορούμε να εξαγάγουμε και να προσθέτουμε τις πληροφορίες τους αυτόματα όταν γνωρίζουμε τη δομή της οντολογίας.

Σε αυτή τη μελέτη, παρουσιάζεται τη μεθοδολογία που προτάθηκε για τη μετατροπή μιας περιγραφής μετα-δεδομένων XML για μια βραχυπρόθεσμη περιφερειακή πρόγνωση καιρού και σχετικά σύνολα δεδομένων σε μια οντολογική αναπαράσταση. Η μεθοδολογία αντιμετωπίζει τη μετατροπή τόσο σε επίπεδο έννοιας όσο και σε επίπεδο στιγμιότυπου και λαμβάνει υπόψη τόσο τη συντακτική όσο και τη σημασιολογική αντιστοίχιση μεταξύ του σχήματος μετα-δεδομένων και της οντολογίας. Η μεθοδολογία μπορεί να γενικευτεί και σε άλλα σύνολα δεδομένων γεωεπιστημών.

Για μελλοντικές εργασίες, το LEADII χρησιμοποίησε το FGDC για να αναπαραστήσει μια σειρά σχετικών προϊόντων που προέρχονται από μία μόνο πρόβλεψη. Η FGDC δεν είναι ισχυρή στην ικανότητά της να συσχετίζει προϊόντα μεταξύ τους. Η οντολογία θα επεκταθεί για να συσχετίσει τα προϊόντα δεδομένων μέσα σε ένα πακέτο πρόβλεψης. Επίσης, οι πληροφορίες και η

σημασιολογία μπορούν να χαθούν και να παρερμηνευθούν κατά τη μετατροπή από XML σε οντολογία, και επομένως η αξιολόγηση είναι ένας σημαντικός τρόπος εξέτασης της ποιότητας μετατροπής. Δεύτερον, θα πρέπει να διερευνήσουν βαθύτερα τη χρονική και χωρική αναπαράσταση των συνόλων δεδομένων γεωεπιστήμης στο μέλλον, καθώς αποτελούν εγγενή χαρακτηριστικά των δεδομένων γεωεπιστήμης. Τα δεδομένα Vortex2 περιέχουν χρονικές πληροφορίες όπως ημερομηνία έναρξης πρόβλεψης, ώρα έναρξης μοντέλου, διάρκεια πρόβλεψης και έχουν χωρικές πληροφορίες όπως συντεταγμένες οριοθέτησης για την περιγραφή γεωγραφικών τοποθεσιών και η αναπαράσταση αυτών των πληροφοριών πρέπει να βελτιωθεί περαιτέρω για να παρέχει καλύτερη περιγραφή δεδομένων και πρόσβαση σε τελικούς χρήστες.

## Εργασία 2<sup>η</sup>

Σε αυτή την εργασία οι Chen και Huang, παρουσιάζουν μια νέα μέθοδο για την αυτόματη δημιουργία της σύνοψης ειδήσεων για τον καιρό με βάση τις τεχνικές ασαφούς συλλογισμού και οντολογίας, όπου η οντολογία του καιρού, η οντολογία χρόνου και η οντολογία γεωγραφίας προκαθορίζονται από ειδικούς του τομέα.

## Πλεονεκτήματα

Το πλεονέκτημα της προτεινόμενης μεθόδου είναι ότι έχει καλύτερη απόδοση για τη σύνοψη ειδήσεων καιρού από τις μεθόδους που παρουσιάστηκαν. Μια νέα μέθοδος για τη δημιουργία σύνοψης ειδήσεων για τον καιρό που βασίζεται σε τεχνικές ασαφούς συλλογισμού και οντολογίας παρουσιάστηκε με ακρίβεια.

Η ακρίβεια και η ανάκληση έχουν χρησιμοποιηθεί ευρέως για την αξιολόγηση της απόδοσης ενός συστήματος περίληψης ειδήσεων. Σε αυτή την έρευνα, τα πειραματικά δεδομένα υιοθετήθηκαν από τον ιστότοπο ειδήσεων καιρού στην Ταϊβάν [17]. Η ακρίβεια και η ανάκληση ορίζονται ως εξής: όπου  $m$  υποδηλώνει τον αριθμό των προτάσεων που επιλέχθηκαν από τους ειδικούς τομέα,  $n$  υποδηλώνει τον αριθμό των προτάσεων που επιλέχθηκαν από το σύστημα σύνοψης και επιλέχθηκαν από τους ειδικούς τομέα

### Μεθοδολογία

Χωρίζονται τα αρχικά άρθρα ειδήσεων καιρού σε ένα σύνολο όρων. Στη συνέχεια, χρησιμοποιούνται δύο οντολογικά χαρακτηριστικά (δηλαδή, ο βαθμός βάθους της οντολογίας και ο βαθμός του πλάτους της οντολογίας) και ένα στατιστικό χαρακτηριστικό (δηλαδή, συχνότητα) ως είσοδοι στο σύστημα. Οι τιμές αυτών των χαρακτηριστικών αντιπροσωπεύονται από ασαφή σύνολα. Στη συνέχεια, ο αλγόριθμος ασαφούς συλλογισμού συνάγει τη βαθμολογία κάθε πρότασης. Η περίληψη αποτελείται από υποψήφιες προτάσεις που έχουν υψηλότερες βαθμολογίες, όπου τα πειραματικά δεδομένα υιοθετούνται από τον ιστότοπο ειδήσεων καιρού της Ταϊβάν. Τα πειραματικά αποτελέσματα δείχνουν ότι η προτεινόμενη μέθοδος υπερέχει των μεθόδων που παρουσιάζονται για την αυτόματη δημιουργία της περίληψης ειδήσεων καιρού.

### Παρόμοιες μελέτες

Τα τελευταία χρόνια, ορισμένοι ερευνητές εστίασαν στο αντίστοιχο ερευνητικό θέμα. Οι Abuobieda, Salim, Albaham, Kumar, και Osman το 2012 [18], παρουσίασαν μια μέθοδο επιλογής χαρακτηριστικών περίληψης κειμένου χρησιμοποιώντας το ψευδογονιδιακό μοντέλο. Οι Chandra, Gupta και Paul το 2011 [19], παρουσίασαν μια στατιστική προσέγγιση για αυτόματη σύνοψη κειμένου με εξαγωγή. Οι Chen και Verma το 2006 [20], παρουσίασαν ένα σύστημα σύνοψης κειμένου βασισμένο σε ερωτήματα χρήστη. Οι Devasena και Hemalatha το 2012 [21], παρουσίασαν μια μέθοδο για αυτόματη κατηγοριοποίηση και σύνοψη κειμένου χρησιμοποιώντας τη μείωση κανόνων. Ο Jiang το 2009 [22], παρουσίασε μια μέθοδο για αυτόματη σύνοψη κειμένου βασισμένη στην εξαγωγή λέξεων-κλειδιών. Οι Jiang, Peng, Ma και Chen το 2010 [23], παρουσίασαν μια μέθοδο αυτόματης σύνοψης κειμένου που βασίζεται σε συνδυασμένη αναγνώριση λέξεων και ομαδοποίηση παραγράφων. Ο Kaikhah το 2004 [24], παρουσίασε μια προσέγγιση μηχανικής μάθησης που χρησιμοποιεί τεχνητά νευρωνικά δίκτυα για να παράγει τις περιλήψεις αυθαίρετου μήκους άρθρων ειδήσεων. Οι Manne, Pervez και Fatima το 2011 [25], παρουσίασαν ένα νέο σύστημα αυτόματης σύνοψης κειμένου με αναγνώριση χαρακτηριστικών όρων. Ο Mohamed το 2005 [26], παρουσίασε μια μέθοδο για τη δημιουργία περιλήψεων προσδιορίζοντας σημαντικές έννοιες/οντότητες και σχέσεις στο κείμενο. Οι Nguyen, Santos και Russell το 2011 [27], μελέτησαν την επίδραση των γνωστικών στυλ χρήστη στην αξιολόγηση της σύνοψης κειμένου. Οι Thakkar, Dharaskar και Chandak το 2010 [28], παρουσίασαν αλγόριθμους βασισμένους σε γραφήματα για τη σύνοψη κειμένου. Ο Yuan και ο Sun το 2005 [29],

παρουσίασαν μια νέα μέθοδο, με το όνομα StructuredCosineSimilarity, για την αντιμετώπιση της ομαδοποίησης εγγράφων ομιλίας. Οι Yang, Dai και Cui το 2008 [30], παρουσίασαν μια νέα προσέγγιση σύνοψης κειμένου χρησιμοποιώντας εξαγωγή προτάσεων με βάση την αναγνώριση λέξεων πυρήνα. Οι Zhang και Li το 2009 [31], παρουσίασαν μια μέθοδο για αυτόματη σύνοψη κειμένων με βάση την ομαδοποίηση και την εξαγωγή προτάσεων. Οι Lee, Chen και Jian το 2003 [32], παρουσίασαν έναν μηχανισμό εξαγωγής ασαφούς γεγονότος OFEE (Ontology-based Fuzzy Event Extraction) με βάση την οντολογία για σύνοψη ηλεκτρονικών ειδήσεων, όπου ο μηχανισμός OFEE περιέχει έναν παράγοντα ανάκτησης, έναν παράγοντα επεξεργασίας εγγράφων και έναν παράγοντα ασαφών συμπερασμάτων για την εκτέλεση του συμβάντος εξαγωγή για σύνοψη ηλεκτρονικών ειδήσεων.

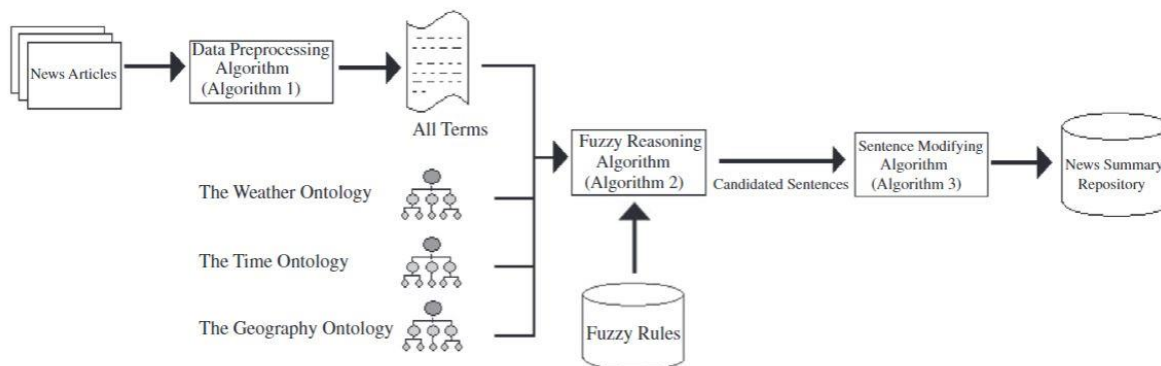
Ο μηχανισμός ανάκτησης ανακτά αυτόματα τις ηλεκτρονικές ειδήσεις του Διαδικτύου περιοδικά, τις αποθηκεύει στο χώρο αποθήκευσης ηλεκτρονικών ειδήσεων και τις στέλνει στον παράγοντα επεξεργασίας εγγράφων για επεξεργασία εγγράφων. Στη συνέχεια, ο παράγοντας επεξεργασίας εγγράφων χρησιμοποιεί την ετικέτα Part-of-Speech (POS) που παρέχεται από την ομάδα επεξεργασίας πληροφοριών γνώσης για την επεξεργασία των ανακτημένων ηλεκτρονικών ειδήσεων και φιλτράρει τον όρο που ορίζεται από ένα φίλτρο όρου. Στη συνέχεια, ο παράγοντας ασαφούς συμπεράσματος και το φίλτρο οντολογίας συμβάντων εξάγουν την οντολογία συμβάντων ηλεκτρονικών ειδήσεων με βάση το σύνολο όρων και την οντολογία τομέα. Τέλος, ο παράγοντας σύνοψης συνοψίζει τις ηλεκτρονικές ειδήσεις με την οντολογία εξαγωγής γεγονότων. Οι Lee, Jian και Huang το 2005 [33], παρουσίασαν έναν μηχανισμό ασαφούς συλλογιστικής επτά επιπέδων για τη δημιουργία της σύνοψης ειδήσεων καιρού, όπου η οντολογία τομέα με διάφορα συμβάντα ειδήσεων είναι προκαθορισμένη από ειδικούς τομέα. Ο μηχανισμός προεπεξεργασίας εγγράφων δημιουργεί όρους με νόημα με βάση το σώμα ειδήσεων και το λεξικό ειδήσεων που ορίζεται από τον ειδικό τομέα. Στη συνέχεια, αυτοί οι όροι με νόημα ταξινομούνται με βάση τα γεγονότα των ειδήσεων από τον όρο ταξινομητής. Ο μηχανισμός ασαφών συμπερασμάτων δημιουργεί τους βαθμούς συμμετοχής για κάθε ασαφή έννοια της ασαφούς οντολογίας. Ένας μηχανισμός ειδήσεων που βασίζεται στη ασαφή οντολογία αναπτύσσεται για τη σύνοψη ειδήσεων, όπου ο μηχανισμός ειδήσεων περιέχει έναν παράγοντα ανάκτησης, έναν μηχανισμό προεπεξεργασίας εγγράφων, έναν εξαγωγέα διαδρομής προτάσεων, έναν παράγοντα δημιουργίας προτάσεων και ένα φίλτρο προτάσεων για την εκτέλεση σύνοψης ειδήσεων. Ωστόσο, επειδή η

απόδοση των μεθόδων που παρουσιάζονται σε προηγούμενες έρευνες δεν είναι αρκετά καλή, ο στόχος της νέας έρευνας ήταν να εντοπιστεί μια καλύτερη μέθοδος για να ξεπεράσουν τα μειονεκτήματα των μεθόδων που παρουσιάστηκαν.

#### Έρευνα και αποτελέσματα

Σε αυτή την εργασία, παρουσιάζεται μια νέα μέθοδος για την αυτόματη δημιουργία της σύνοψης ειδήσεων για τον καιρό με βάση τις τεχνικές ασαφούς συλλογισμού και οντολογίας, όπου η οντολογία του καιρού, η οντολογία χρόνου και η οντολογία γεωγραφίας προκαθορίζονται από ειδικούς του τομέα. Χρησιμοποιούνται παραδοσιακές τεχνικές οντολογίας όπως αναφέρθηκαν στις παραπάνω έρευνες για την αναπαράσταση γνώσης, όπου οι τιμές των ιδιοτήτων των εννοιών είναι καθαρές τιμές. Χωρίζονται τα αρχικά άρθρα ειδήσεων καιρού σε ένα σύνολο όρων. Χρησιμοποιούνται δύο οντολογικά χαρακτηριστικά και ένα στατιστικό χαρακτηριστικό ως είσοδοι στο σύστημα. Στη συνέχεια, ο αλγόριθμος ασαφούς συλλογισμού συνάγει τη βαθμολογία κάθε πρότασης σε κάθε είδηση, όπου οι πράξεις Max–Min του Mamdani για ασαφή συλλογισμό χρησιμοποιούνται με βάση 27 ασαφείς κανόνες. Με βάση την οντολογία του καιρού, η χρονική οντολογία και η οντολογία γεωγραφίας, μπορούν να ληφθούν το βάθος, το πλάτος και η συχνότητα κάθε πρότασης. Στη συνέχεια, το βάθος, το πλάτος και η συχνότητα των προτάσεων χρησιμοποιούνται ως τιμές εισόδου των ασαφών κανόνων για την εκτέλεση ασαφούς συλλογισμού για να ληφθεί η βαθμολογία κάθε πρότασης. Εάν η βαθμολογία μιας πρότασης σε μια είδηση είναι μεγαλύτερη ή ίση με μια προκαθορισμένη κατώτατη τιμή που καθορίζεται από τον ειδικό τομέα, τότε είναι υποψήφια πρόταση. Αφού εκτελεσθεί η διαδικασία ασαφούς συλλογισμού, μπορούμε να πάρουμε ένα σύνολο υποψηφίων προτάσεων που έχουν υψηλότερες βαθμολογίες ως περίληψη ειδήσεων. Τα πειραματικά δεδομένα υιοθετήθηκαν από τον ιστότοπο ειδήσεων καιρού της Ταϊβάν.

Τα πειραματικά αποτελέσματα δείχνουν ότι η προτεινόμενη μέθοδος υπερέχει των μεθόδων που παρουσιάστηκαν στις προηγούμενες έρευνες την αυτόματη δημιουργία της περίληψης ειδήσεων καιρού. Η καινοτομία της προτεινόμενης μεθόδου που ξεπερνά τις μεθόδους που παρουσιάστηκαν είναι ότι βασίζεται σε τεχνικές ασαφούς συλλογισμού και οντολογίας, όπου προτείνεται ένας αλγόριθμος ασαφούς συλλογισμού και χρησιμοποιούνται 27 ασαφείς κανόνες για ασαφή συλλογιστική για να συναχθεί η βαθμολογία κάθε πρότασης σε μια είδηση.



Εικόνα 5: Παράδειγμα χαρτογράφησης που βασίζεται στην κατανόηση του πλαισίου. [9]

### Σύγκριση 1ης και 2ης έρευνας

Στο πλαίσιο της πρώτης έρευνας παρατηρούμε ότι η αναπαράσταση της οντολογίας βασίζεται κυρίως στη σημασιολογία και δίνει την δυνατότητα για πρόσβαση σε μία ανεπτυγμένη σημασιολογία που βασίζεται στη μηχανική κατανόηση. Με βάση τη σημασιολογία μπορεί να βελτιωθεί ο τρόπος πρόσβασης, τα δεδομένα, ο τρόπος με τον οποίο κοινοποιούνται αυτά και ο τρόπος με τον οποίο χρησιμοποιούνται και πάλι από νέους χρήστες.

Μία επιπλέον διευκόλυνση που μας δίνει η πρώτη έρευνα είναι στον τομέα αλληλεπίδρασης με τους υπόλοιπους πόρους που βρίσκονται στο σημασιολογικό ιστό. Έτσι διαπιστώνεται ότι τα δεδομένα συνδέονται με εύκολο τρόπο το ένα με το άλλο. Ακόμη διαπιστώνεται ότι με βάση τα αποτελέσματα της πρώτης έρευνας οι γλώσσες τις οποίες χρησιμοποιεί η οντολογία μπορούν να βοηθήσουν στην καλύτερη λειτουργικότητα ανάμεσα στα διαφορετικά σύνολα δεδομένων, κάτι που έχει ως αποτέλεσμα την πιο εύκολη και πιο κατανοητή περιγραφή τους.

Ο στόχος αυτής της μελέτης είναι να αναπτυχθεί μία οντολογική αναπαράσταση που θα αφορά τα σύνολα δεδομένων τα οποία ωστόσο θα βασίζονται στα υπάρχοντα μετα-δεδομένα προκειμένου να μετατραπούν τα μετα-δεδομένα σε ένα είδος αναπαράστασης οντολογίας. Η έρευνα επιλέγει να χρησιμοποιήσει συγκεκριμένα σύνολα δεδομένων όπως αυτό της πρόβλεψης του καιρού με τη χρήση περιγραφικών μετα-δεδομένων. Για αυτό το λόγο χρησιμοποιεί τα έργα Vortex2 και LEADII προκειμένου να δημιουργήσει βραχυπρόθεσμες προβλέψεις για τον καιρό. Στο πλαίσιο

της πρώτης μελέτης μιλάμε για τη μετατροπή μίας περιγραφής μετα-δεδομένων xml που αφορά και σε αυτή την περίπτωση μία βραχυπρόθεσμη περιφερειακή πρόγνωση του καιρού με βάση τις οντολογικές αναπαραστάσεις. Ωστόσο σε αυτή την περίπτωση κατά τη διάρκεια της μετατροπής των μετα-δεδομένων λαμβάνεται υπόψη τόσο η συντακτική όσο και η σημασιολογική αντιστοίχιση ανάμεσα στο σχήμα και την οντολογία. Όσον αφορά τη δεύτερη έρευνα αναφερόμαστε στην προσπάθεια εξαγωγής μιας μεθόδου για να αυτοματοποιηθεί ο τρόπος δημιουργίας της σύνοψης ειδήσεων για τον καιρό με βάση τις τεχνικές ασαφούς συλλογισμού και οντολογίας στο πλαίσιο των οποίων η οντολογία του χρόνου και η οντολογία της γεωγραφίας θα μπορούν να είναι προκαθορισμένες.

Η συγκεκριμένη έρευνα έλαβε υπόψη της τα οντολογικά χαρακτηριστικά, δηλαδή το βαθμό βάθους της οντολογίας και το βαθμό πλάτους της. Επιπλέον έλαβε υπόψη τη συχνότητα προκειμένου αυτά τα χαρακτηριστικά να λειτουργήσουν ως πυλώνες για την είσοδο στο σύστημα. Τέλος κάθε πρόταση βασίστηκε στον αλγόριθμο ασαφούς συλλογισμού με βάση τον οποίο εξήχθη η βαθμολογία για την καθεμία από αυτές. Έτσι οι προτάσεις που συγκέντρωσαν την υψηλότερη βαθμολογία υιοθετήθηκαν ως πειραματικά δεδομένα από τον ιστότοπο ειδήσεων καιρού της Ταϊβάν.

### Εργασία 3<sup>η</sup>

Οι Masmoudi, Karray, Lamine, Zghal και Archimede, στην εργασία τους θίγουν το μεγάλο πλέον ζήτημα των φυσικών καταστροφών. Με τις ραγδαίες κλιματικές αλλαγές, έχουν αυξηθεί κατά πολύ οι καταγίδες, οι πλημμύρες καθώς και οι τυφώνες σε όλο τον πλανήτη. Οι καταστροφές αυτές έχουν επηρεάσει τόσο το φυσικό περιβάλλον όσο και τις ανθρώπινες ζωές, είτε με καταστροφές περιουσιών είτε ακόμη και με απώλειες ζωών.

Σύμφωνα με τους συγγραφείς, υπάρχουν σημαντικά συστήματα γεωσκόπησης τα οποία έχουν σημαντικό ρόλο στην παρακολούθηση και πρόβλεψη φυσικών καταστροφών. Τα δεδομένα που εξάγονται από αυτά τα συστήματα όμως, αποθηκεύονται σε διαφορετικές μορφές καθώς και σε διαφορετικές πηγές. Αυτό δημιουργεί περιορισμό στην εκμετάλλευση αυτών των δεδομένων, καθώς και σε κάποιες περιπτώσεις τα δεδομένα δεν μπορούν να αξιοποιηθούν. Έτσι, προτείνουν μία οντολογία για τη λύση της σημασιολογικής διαλειτουργικότητας αυτών των δεδομένων. Στην

εργασία τους, παρουσιάζουν μία αρθρωτή οντολογία, για την παρακολούθηση του περιβάλλοντος, η οποία αναπτύχθηκε με βάση μια πρωτότυπη ευέλικτη μεθοδολογία, η οποία λέγεται MEMOn (Modular Environmental Monitoring Ontology). Κύριος σκοπός αυτής της οντολογίας είναι, η υποστήριξη σημασιολογικής διαλειτουργικότητας των δεδομένων καθώς και η σύνδεση ετερογενών δεδομένων τα οποία συλλέγονται μέσω τεχνικών και συστημάτων παρατήρησης.

Έτσι, οι ερευνητικές κοινότητες έδωσαν μεγάλη σημασία στην ανάπτυξη και την εφαρμογή συστημάτων παρατήρησης της γης (EO – Earth Observation) όπως αισθητήρες και δορυφορικές πλατφόρμες καθώς και προγραμμάτων παρατήρησης γης όπως το πρόγραμμα Copernicus2 καθώς και το SERVIR3 Global. Από την εξαγωγή όλων των δεδομένων μέσω των εφαρμογών και των προγραμμάτων που αναφέραμε παραπάνω, προκύπτει το εξής πρόβλημα. Το πλήθος είναι εκατοντάδες εκατομμύρια δεδομένα για το κλίμα, δεδομένα των ωκεανών των ακτών και της ξηράς, τα οποία αποθηκεύονται σε διαφορετικές μορφές όπως, βάσεις δεδομένων, αρχεία CSV (Comma Separated Values) και εικόνες Raster, και ο αριθμός των δεδομένων συνεχίζει να αυξάνεται δημιουργώντας πρόβλημα στην εκμετάλλευση των δεδομένων μεταξύ των διαφορετικών πηγών, προγραμμάτων και συστημάτων.

#### Μεθοδολογία δημιουργίας της οντολογίας MEMOn (Modular Environmental Monitoring Ontology)

Στις επόμενες παραγράφους θα γίνει παρουσίαση της μεθοδολογίας που ακολουθήθηκε από τους συγγραφείς για την κατασκευή της συγκεκριμένης οντολογίας. Σύμφωνα με τους Lopez και Corcho, η κατασκευή μίας αρθρωτής οντολογίας είναι μία σύνθετη διαδικασία, συγκεκριμένα όταν οι οντολογίες γίνονται πιο σημαντικές και περίπλοκες [34].

Ένα από τα μεγαλύτερα προβλήματα είναι, ο τρόπος επιλογής της καταλληλότερης μεθοδολογίας, λαμβάνοντας βέβαια υπόψη την πολυπλοκότητα του εκάστοτε τομέα, στον οποίο πρόκειται η οντολογία να μοντελοποιηθεί και να επαναχρησιμοποιηθεί. Στη συγκεκριμένη εργασία, οι συγγραφείς επέλεξαν τη μεθοδολογία AOM (Agile Methodology for Developing Ontology Modules).

Σύμφωνα με τον Gobin, η μεθοδολογία AOM, προτείνεται για την ανάπτυξη ενοτήτων οντολογίας οι οποίες μπορούν να επαναχρησιμοποιηθούν καθώς και να συγχωνευθούν για τη δημιουργία μιας αρθρωτής οντολογίας σε ένα συγκεκριμένο τομέα. Παρακάτω δίνεται ο ορισμός από τον Gobin:

Η μεθοδολογία AOM, είναι μία ευέλικτη μεθοδολογία η οποία χρησιμοποιείται από μηχανικούς οντολογίας για την ανάπτυξη ενοτήτων οντολογίας. Καθορίζει διαδικασίες στις οποίες ειδικοί τομέα και μηχανικοί οντολογίας συναντώνται για να δημιουργήσουν ενότητες οντολογίας, που αντιπροσωπεύονται σε μία γλώσσα οντολογίας, χρησιμοποιώντας κατάλληλα εργαλεία [35].

Η επιλογή της συγκεκριμένης μεθοδολογίας έγινε διότι, η AOM υποστηρίζει τη σταδιακή και επαναληπτική ανάπτυξη των ενοτήτων του MEMOn. Η μεθοδολογία AOM, ορίζεται από τις 4 ακόλουθες φάσεις:

1. Εξερεύνηση
2. Σχεδιασμό
3. Ανάπτυξη ενότητας
4. Απελευθέρωση

Στην συνέχεια, οι συγγραφείς προσαρμόσανε αυτές τις 4 φάσεις ανάλογα με τις ανάγκες. Ο κύκλος ζωής της ανάπτυξης ξεκινά αρχικά με τη φάση της εξερεύνησης, όπου γίνεται ο προσδιορισμός των απαιτήσεων και οι στόχοι της οντολογίας από τους ίδιους τους συγγραφείς. Αργότερα, ορίσανε ένα σύνολο ερωτήσεων ικανότητας (CQs – Competency Questions), σύμφωνα με τον τομέα που μελετήθηκε καθώς και με τις ανάγκες των ειδικών. Σύμφωνα με τους Gruninger και Fox, τα σύνολα ερωτήσεων ικανότητας, αποτελούνται από ένα σύνολο ερωτήσεων τα οποία δηλώνονται σε φυσική γλώσσα, ώστε η οντολογία να μπορεί να απαντήσει σωστά [36]. Έχοντας το συγκεκριμένο σύνολο ερωτήσεων ικανότητας, οι συγγραφείς εξήγαγαν τις ιστορίες ώστε να αναπαραστήσουν τα διαφορετικά περιβάλλοντα τα οποία αφορούν τον τομέα που εξετάζουν, το οποίο αποτελεί τη φάση του σχεδιασμού. Οι ιστορίες αυτές αποτελούν σενάρια από τα οποία λαμβάνονται γεγονότα, τα οποία συμβάλουν σημαντικά στην οικοδόμηση της οντολογίας. Οι συγκεκριμένες ιστορίες συντάχτηκαν βάση των πληροφοριών που συγκεντρώθηκαν κατά τη διάρκεια συνεντεύξεων των ειδικών του συγκεκριμένου τομέα, δηλαδή του τομέα της περιβαλλοντικής παρακολούθησης. Τα σύνολα των ερωτήσεων ικανότητας που έχουν άμεση

σχέση με μία ενότητα, χρησιμοποιήθηκαν για το σχεδιασμό μιας ημι-επίσημης ενότητας. Τέλος, οι ενότητες γίνονται επίσημες, υλοποιούνται και αξιολογούνται ώστε να συγχωνευθούν στην τελική φάση, αυτή της απελευθέρωσης.

#### Φάση 1<sup>η</sup>: Εξερεύνηση

Στη συγκεκριμένη φάση, η οποία είναι και η πρώτη του κύκλου, γίνεται η έναρξη ανάπτυξης της οντολογίας. Εδώ, ο ειδικός τομέα (Domain Expert) είναι μέρος της ομάδας ανάπτυξης και παράλληλα με το μηχανικό της γνώσης (Knowledge Engineer), ξεκινούν τη διαδικασία της ανάπτυξης. Για να υπάρξει καλύτερη κατανόηση του τομέα της περιβαλλοντικής παρακολούθησης, οι συγγραφείς, απευθύνθηκαν στον οργανισμό OSS (Sahara and Sahel Observatory) [37], ο οποίος αποτελεί τον εταίρο στο έργο τους, ο οποίος διαθέτει σημαντική εξειδίκευση στην περιβαλλοντική παρακολούθηση, τη διαχείριση φυσικών πόρων καθώς και γνώσεις και πληροφορίες για την κλιματική αλλαγή. Στη συγκεκριμένη φάση, οι μηχανικοί οντολογίας (Knowledge Engineer) καθώς και οι ειδικοί από τον OSS, συνδύασαν τις γνώσεις τους για τον τομέα ενδιαφέροντος, τους στόχους καθώς και τις απαιτήσεις της συγκεκριμένης οντολογίας. Ο στόχος της οντολογίας MEMOn, είναι να παρέχει μία επίσημη αναπαράσταση του τομέα της περιβαλλοντικής παρακολούθησης για την υποστήριξη των παρακάτω:

1. Σημασιολογική διαλειτουργικότητα
2. Ανακάλυψη και παραγωγή γνώσης
3. Ενοποίηση και σύνδεση δεδομένων
4. Συνολική προβολή δεδομένων με σκοπό την καλύτερη κατανόηση των περιβαλλοντικών φαινομένων.

Η συγκεκριμένη οντολογία αναπτύχθηκε με σκοπό της ενσωμάτωσης γνώσης η οποία περιέχεται σε πολλαπλές πηγές περιβαλλοντικών δεδομένων. Η οντολογία MEMOn έχει βασικές απαιτήσεις οι οποίες παρατίθενται ως εξής:

1. Η οντολογία λαμβάνει έννοιες οι οποίες έχουν σχέση με τον τομέα της περιβαλλοντικής παρακολούθησης.
2. Η οντολογία εφαρμόζει τις αρχές αρθρωποίησης.

Σύμφωνα με τον Doran, η αρθροποίηση οντολογίας αποτελεί μία διαδικασία καθορισμού μίας συγκεκριμένης ενότητας, η οποία αποτελεί υποσύνολο από μία μεγαλύτερη οντολογία [38]. Η αρθρώτητα της οντολογίας συμβάλει στη:

1. Βελτίωση της διαδικασίας ανάπτυξης οντολογίας με σκοπό τη μείωση της πολυπλοκότητας του σχεδιασμού, της συντήρησης, του εμπλουτισμού καθώς και της αντικατάστασης ενοτήτων.
2. Διατήρηση της σαφήνειας και της συνοχής της οντολογίας παρουσιάζοντας ενότητες της οντολογίας με τις απαραίτητες γνώσεις.
3. Προώθηση της επαναχρησιμοποίησης της κάθε ενότητας ξεχωριστά.

#### Φάση 2<sup>η</sup>: Σχεδιασμός

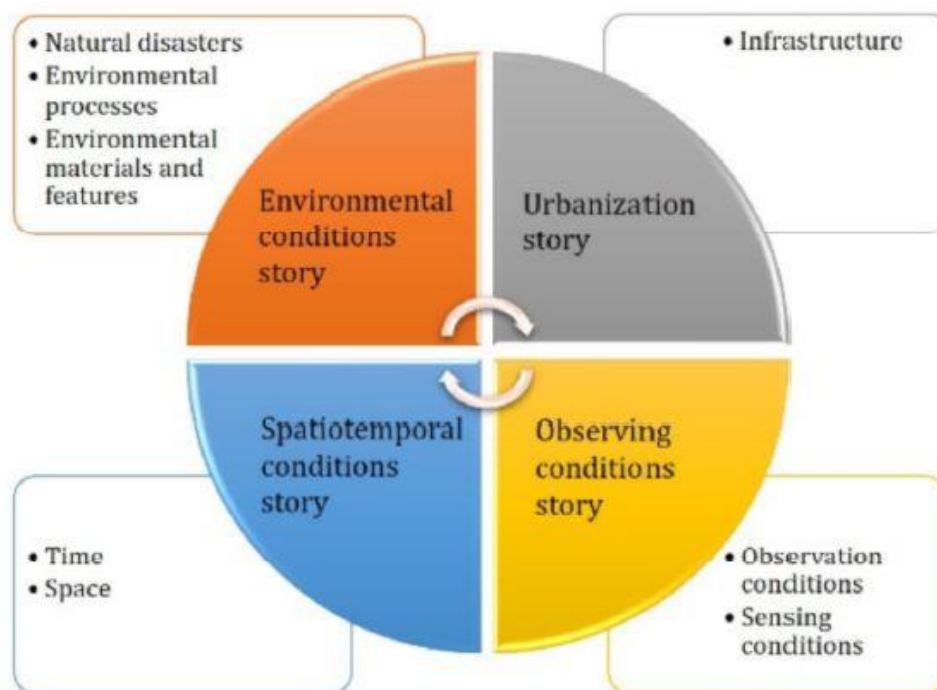
Στη συγκεκριμένη φάση, οι ερωτήσεις ικανότητας, ορίζονται από τον ειδικό τομέα (OSS). Στον τομέα ενδιαφέροντος της συγκεκριμένης έρευνας, γίνεται διερεύνηση πολλαπλών πηγών δεδομένων, όπως το OSS, το Copernicus, το EMDAT κλπ. Σε κάθε μία από αυτές τις πηγές υπάρχει ξεχωριστό σύνολο σχετικών δεδομένων τα οποία βρίσκονται σε συγκεκριμένους κλάδους. Πιο συγκεκριμένα, το OSS, παράγει δεδομένα κλιματολογικού περιεχομένου όπως δεδομένα για τη θερμοκρασία, για τη βροχόπτωση κλπ. Ενώ, το EMDAT, παράγει δεδομένα φυσικών καταστροφών. Επομένως, για να υπάρχει μεγαλύτερη κατανόηση της γνώσης που εξάγεται από διαφορετικές πηγές δεδομένων, οι συγγραφείς όρισαν τις ερωτήσεις ικανότητας, τις οποίες θα πρέπει η οντολογία να είναι σε θέση να απαντήσει.

| CQi | Competency question                                                                                                  |
|-----|----------------------------------------------------------------------------------------------------------------------|
| 1   | What is the value of a given observation in a region [X] in date [Y]?                                                |
| 2   | What are the measurements for an observed property [X] for natural disaster [Y]?                                     |
| 3   | What is the average measurement for climate parameter [X] during natural disaster [Y]?                               |
| 4   | What is the measurement unit used for a specific property [X]?                                                       |
| 5   | What are the observations associated with a specific environmental process [X]?                                      |
| 6   | What is the type of sensor used for a given observation [X]?                                                         |
| 7   | Which natural disasters may lead to natural disaster [X]?                                                            |
| 8   | What is the disaster resulting from an environmental process [X]?                                                    |
| 9   | What are the environmental processes that caused the natural disaster [X]?                                           |
| 10  | Where did the environmental process [X], which is associated with natural disaster [Y], take place?                  |
| 11  | What is the environmental process [X] that is preceded by the environmental process [Y] during natural disaster [Z]? |
| 12  | What are the environmental materials involved in the environmental process [X]?                                      |
| 13  | What environmental material is the output of the environmental process [X]?                                          |
| 14  | Where did the natural disaster [X] take place?                                                                       |
| 15  | What are the spatial zones where a specific disaster activity is the highest?                                        |
| 16  | Is the infrastructure [X] implicated in the occurrence of the natural disaster [Y]?                                  |
| 17  | When did the natural disaster [X] take place?                                                                        |
| 18  | What is the environmental process that precedes the environmental process [X]?                                       |
| 19  | Is the environmental process [X] has end location [Y]?                                                               |
| 20  | What is the environmental process that co-occurs with the environmental process [X]?                                 |

Εικόνα 6: Δείγμα ερωτήσεων ικανότητας [6]

[11]

Οι συγγραφείς, έχοντας το συγκεκριμένο σύνολο ερωτήσεων ικανότητας, εξάγονται ιστορίες ώστε να οριστεί η περίμετρος του τομέα μελέτης. Στην πορεία, εντόπισαν 4 κύριες ιστορίες οι οποίες συνέβαλαν στην περιβαλλοντική παρακολούθηση:



[11]Εικόνα 7: Κύριες Ιστορίες που εξήγαγαν οι Masmoudi, Karray, Lamine, Zghal και Archimede. [11]

### **Ιστορία 1<sup>η</sup>:** Ιστορία Περιβαλλοντικών Συνθηκών (EnvironmentalConditionsStory)

Η συγκεκριμένη ιστορία, περιλαμβάνει γεωλογικές δομές όπως τύπους βράχων και εδάφους, γεωγραφικά χαρακτηριστικά όπως δάσος και γη, υδρολογικές δομές όπως νερά, ωκεανοί, ποτάμια και λίμνες, περιβαλλοντικές διεργασίες όπως ταρακούνημα του εδάφους, καθώς και φυσικές καταστροφές.

### **Ιστορία 2<sup>η</sup>:** Ιστορία Αστικοποίησης (UrbanizationStory)

### **Ιστορία 3<sup>η</sup>:** Ιστορία Παρατήρησης Συνθηκών (ObservingConditionsStory)

Στη συγκεκριμένη ιστορία, περιλαμβάνονται συστήματα ανίχνευσης όπως αισθητήρες, δορυφόροι κλπ. καθώς και κάποιες συνθήκες μέτρησης.

### **Ιστορία 4<sup>η</sup>:** Ιστορία Χωροχρονικών Συνθηκών (SpatiotemporalConditionsStory)

Από τις 4 αυτές ιστορίες, οι συγγραφείς, εντόπισαν 8 οντολογικές ενότητες. Η πρώτη ιστορία ασχολείται με περιβαλλοντικά υλικά, περιβαλλοντικές διεργασίες και φυσικές καταστροφές, τα οποία είναι ανεξάρτητα από τον κλάδο. Έτσι, προσδιορίστηκαν ως 3 οντότητες. Στη δεύτερη ιστορία, παρατήρησαν ότι έχει να κάνει με την υποδομή. Από την τρίτη ιστορία, προσδιόρισαν 2 ενότητες, τη μονάδα παρατήρησης και μέτρησης, και τη μονάδα αισθητήρα και ανίχνευσης. Η τέταρτη ιστορία ασχολείται με τις χρονικές και τις χωρικές συνθήκες των γεγονότων ή παρατηρήσεων. Έτσι, στην τέταρτη ιστορία προσδιόρισαν τη χρονική ενότητα που περιλαμβάνει χρονικές οντότητες καθώς και τη γεωχωρική ενότητα η οποία καλύπτει τις οντότητες σχετικά με τοποθεσίες. Μία από τις απαιτήσεις της οντολογίας MEMOn, είναι η επαναχρησιμοποίηση υπαρχουσών οντολογιών, η ενότητα γεωχωρική οντολογία προκύπτει από το αποτέλεσμα του συνδυασμού της πρώτης και της τέταρτης ιστορίας.

Στην πορεία, αντιστοίχησαν τις σχέσεις μεταξύ των μονάδων στην οντολογία. Αφού χρησιμοποίησαν αξιώματα, κατάφεραν να διασφαλίσουν ότι οι περιβαλλοντικές διαδικασίες συνδέονται με ένα χωροχρονικό πλαίσιο μέσω των δύο σχέσεων Time\_R και Spatial\_R. Το Time\_R, αντιπροσωπεύει τις σχέσεις χρονισμού (π.χ., συμβαίνει\_ανα, συμβαίνει\_πριν, συμβαίνει\_μετά κλπ.) διαφορετικών καταστροφικών καιρικών/κλιματικών γεγονότων καθώς και μεταξύ τους. Το Spatial\_R, αντιπροσωπεύει τις χωρικές σχέσεις των γεγονότων που συμβαίνουν (τοποθεσία).

### Φάση 3<sup>η</sup>: Ανάπτυξη Ενότητας

Η οντολογία που προτείνουν οι συγγραφείς, αποτελείται από 8 ενότητες, οι οποίες καλύπτουν τους κλάδους της περιβαλλοντικής παρακολούθησης. Οι 8 ενότητες είναι οι εξής:

1. Παρατήρησης και μέτρησης
2. Αισθητήρα και ανίχνευσης
3. Φυσικών καταστροφών
4. Περιβαλλοντικής διαδικασίας
5. Περιβαλλοντικού υλικού
6. Υποδομής
7. Χρονική μονάδα
8. Γεωχωρική μονάδα

Αυτές οι ενότητες, αναπτύσσονται και απελευθερώνονται με 8 επαναλήψεις. Η επαναληπτική διαδικασία αρχίζει εφόσον οριστούν οι ερωτήσεις ικανότητας τις οποίες πρέπει να απαντήσει η επιλεγμένη ενότητα.

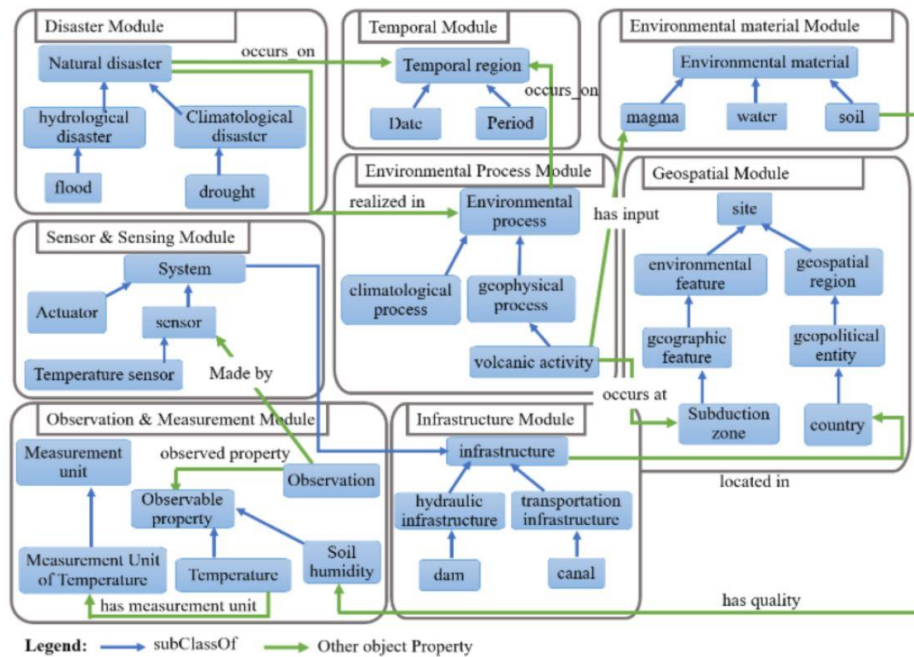
Οι συγγραφείς, χρησιμοποίησαν μία υβριδική προσέγγιση, η οποία βασίζεται στην ευθυγράμμιση από πάνω προς τα κάτω σε μία οντολογία ανώτατου επιπέδου σύμφωνα με τους Noy και McGuinness [39], καθώς και μία προσέγγιση από κάτω προς τα πάνω σε κλάσεις, οι οποίες βασίζονται σε περιβαλλοντικές πηγές δεδομένων.

#### *Πλεονεκτήματα Επαναχρησιμοποίησης της Οντολογίας*

1. Δεν είναι απαραίτητο να ξανά δημιουργηθεί η οντολογία εκ νέου.
2. Αυξάνεται η ποιότητα των νέων οντολογιών καθώς τα στοιχεία που επαναχρησιμοποιούνται, έχουν δοκιμαστεί.
3. Με την επαναχρησιμοποίηση, βελτιώνεται η αποτελεσματικότητα της συντήρησης της οντολογίας.

#### *Φάση 4<sup>η</sup>: Απελευθέρωση*

Στο πρώτο στάδιο της απελευθέρωσης, οι ανεπτυγμένες ενότητες συγχωνεύονται. Στην πορεία, η αρθρωτή οντολογία εμπλουτίζεται από σχέσεις μεταξύ των ενότητων. Έτσι προκύπτει ότι, κάθε ενότητα έχει τουλάχιστον μία σχέση με άλλες ενότητες, όπως για παράδειγμα η ενότητα της καταστροφής συνδέεται με την ενότητα περιβαλλοντικής διεργασίες μέσω της σχέσης «realizedin». Αυτό σημαίνει ότι μία καταστροφή μπορεί να δημιουργηθεί όταν γίνεται μία η περισσότερες περιβαλλοντικές διεργασίες. Ακόμη, η ενότητα περιβαλλοντικής διεργασίας, συνδέεται με τη χρονική ενότητα μέσω της σχέσης «occurson» καθώς και με τη γεωχωρική ενότητα μέσω της σχέσης «occursat». Έτσι, μέσω αυτών των συνδέσεων προσδιορίζεται ο χρόνος καθώς και η θέση μιας διεργασίας.



[11]Εικόνα 8: Μερική προβολή ενότητων της οντολογίας MEMOn. [11]

Η οντολογία σύμφωνα με τους συγγραφείς περιλαμβάνει 1246 κλάσεις, 417 ιδιότητες μαζί με 10956 αξιώματα (λογικές δηλώσεις) που διασυνδέσουν τις κλάσεις.

### Επαλήθευση Οντολογίας MEMOn

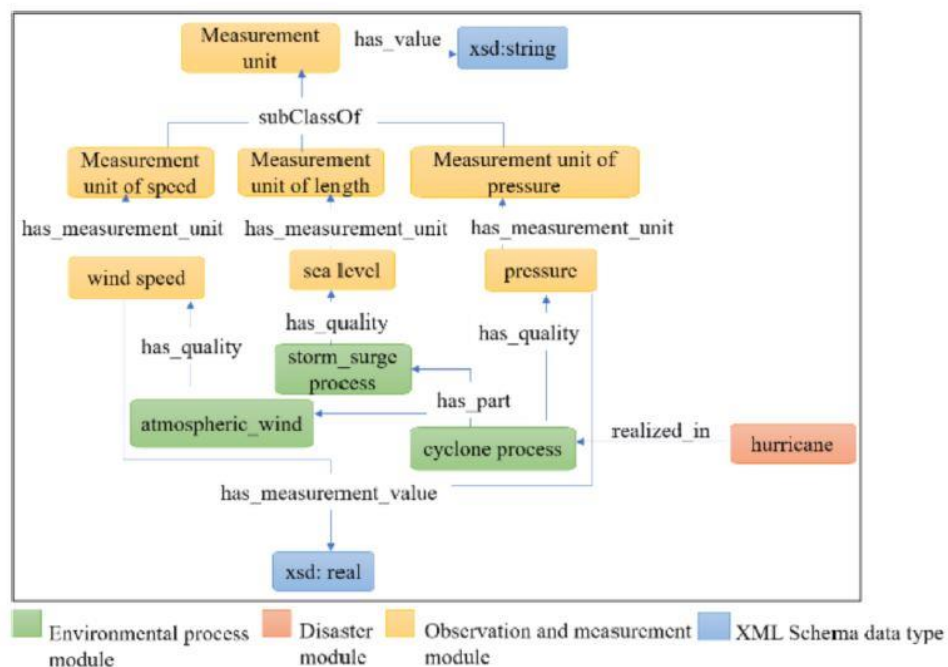
Η επαλήθευση οντολογίας είναι η αξιολόγηση οντολογίας που συγκρίνει την οντολογία σύμφωνα με τις προδιαγραφές (οντολογικές απαιτήσεις και ερωτήσεις ικανότητας, Φάση 1<sup>η</sup> και Φάση 2<sup>η</sup>).

### Επικύρωση Οντολογίας MEMOn

Για την επικύρωση της οντολογίας οι συγγραφείς επικεντρώθηκαν σε 2 μελέτες περίπτωσης. Η πρώτη αφορά την αξιοποίηση και επαναχρησιμοποίηση της περιβαλλοντικής γνώσης, και η δεύτερη αφορά την ενοποίηση των δεδομένων.

- **Μελέτη περίπτωσης:** Αξιοποίηση και επαναχρησιμοποίηση της περιβαλλοντικής γνώσης:
  - ο Εδώ οι συγγραφείς, παρουσίασαν ένα παράδειγμα με την κλάση «τυφώνα» της οντολογίας MEMOn, με την περίπτωση «IRMA» που συνέβη στην Καραϊβική το

2017. Στην πορεία, χαρτογράφησαν αυτή την περίπτωση με άλλες περιπτώσεις που περιγράφουν τον τυφώνα IRMA, όπως για παράδειγμα η βαρομετρική πίεση (914mb), την ταχύτητα του ανέμου (285 km/h) καθώς και τη στάθμη της θάλασσας (5,4m). Στη συνέχεια, χρησιμοποίησαν τη λογική Pellet ώστε να βγάλουν συμπεράσματα σχετικά με τη ρητή – άρρητη γνώση. Τέλος, χρησιμοποίησαν το ερώτημα SPARQL ώστε να εξάγουν τη λίστα των τυφώνων πέντε κατηγοριών όπου υπάρχουν στη γνωσιακή βάση της οντολογίας.



Εικόνα 9: Το τμήμα του MEMOn που χρησιμοποιείται για τον ορισμό του τυφώνα 5 κατηγοριών. [11]

Έτσι, με τη βοήθεια του κανόνα R4 ο οποίος ταξινομεί αυτόματα μία περίπτωση της κατηγορίας τυφώνα, γνωρίζοντας ότι συνέβη με ταχύτητα ανέμου η οποία ήταν μεγαλύτερη από 252 χλμ./ώρα, ως τυφώνας πέντε κατηγοριών.

- **Μελέτη περίπτωσης:** Ενοποίηση δεδομένων πολλαπλών πηγών:

- ο Στη δεύτερη μελέτη περίπτωσης, οι συγγραφείς παρουσίασαν ένα πραγματικό παράδειγμα, το οποίο εγγυάται τη σημασιολογική διαλειτουργικότητα ενσωμάτωσης και σύνδεσης παρατηρούμενων δεδομένων από πολλαπλές και διαφορετικές πηγές. Ο σκοπός αυτής της μελέτης περίπτωσης είναι ώστε να αποδειχθεί η ικανότητα της οντολογίας MEMOn. Πιο συγκεκριμένα, το παράδειγμα αυτό ασχολείται με δεδομένα βροχοπτώσεων από το OSS σε μορφή εικόνων RASTER, δεδομένα που αφορούν γεγονότα καταιγίδας από το NOAA καθώς και δεδομένα που αφορούν συμβάντα πλημμυρών από το EMDAT σε μορφή CSV (CommaSeparatedValue). Στη συνέχεια, οι συγγραφείς αντιμετώπισαν το εξής πρόβλημα, τα σχήματα των δεδομένων (μετα-δεδομένα) στις τρεις πηγές που αναφέρθηκαν πιο πάνω, είναι ετερογενή. Οι όροι, παρουσιάζονται διαφορετικά σε κάθε πηγή. Για παράδειγμα, στο EMDAT, ο όρος καταστροφή χρησιμοποιείται ώστε να περιγράψει μία φυσική καταστροφή, ενώ ο όρος γεγονός χρησιμοποιείται στο NOAA. Επομένως, ο σκοπός του MEMOn, είναι να διασφαλίσει τη σημασιολογική διαλειτουργικότητα μεταξύ των ετερογενών σχημάτων δεδομένων. Στην πορεία, ώστε να ενσωματώσουν οι συγγραφείς τα σημασιολογικά δεδομένα πολλαπλών πηγών, χρησιμοποίησαν το σύστημα WebKarma, το οποίο σύμφωνα με τον Gupta, είναι ένα πλαίσιο μοντελοποίησης και ολοκλήρωσης δεδομένων [40]. Το σύστημα Karma, παρέχει εργαλεία για την ημιαυτόματη κατασκευή ενός σημασιολογικού μοντέλου μιας πηγής δεδομένων. Έτσι, μπορεί να καθιστά δυνατή την γρήγορη αντιστοίχιση ενός συνόλου πηγών όπως XML, CSV, JSON, σε μία οντολογία τομέα.

#### Παρόμοιες μελέτες

Οι Buttiegieg, Pafilis, Lewis, Schlidhauer, Walls και Mungall [41], το 2016 ανέπτυξαν την ήδη υπάρχουσα οντολογία ENVO (Environmental Ontology) επάνω στη γλώσσα οντολογίας ιστού (OWL) και χρησιμοποίησαν μεθόδους προτύπων για να επιταχύνουν τη δημιουργία κλάσεων. Η συγκεκριμένη οντολογία σχολιάστηκε στο παρελθόν το 2013 από τους Buttiegieg, Morrison, Smith, Mungall και Lewis [42]. Σε σχέση με το 2013, το πεδίο εφαρμογής της οντολογίας ENVO έχει βελτιωθεί καθώς επίσης έχει βελτιωθεί και η ίδια εφαρμογή του ENVO. Επιπροσθέτως, ένα μεγάλο μέρος του υπάρχοντος περιεχομένου του έχει αναθεωρηθεί για ακόμη πιο βελτιωμένη

σημασιολογική αναπαράσταση. Οι Corominas, Garrido-Baserba, Villez, Olsson, Corte και Poch το 2018 [43], στην εργασία τους, το κύριο αντικείμενο ήταν η περιγραφή των σύγχρονων τεχνικών οι οποίες βασίζονται σε υπολογιστικά συστήματα για την ανάλυση δεδομένων με σκοπό τη βελτίωση της λειτουργίας των μονάδων επεξεργασίας λυμάτων. Συγκεκριμένα, ασχολήθηκαν με την διαχείριση γνώσης με τη χρήση της ασαφούς λογικής (από πληροφορίες σε γνώση) και έφτασαν στο αποτέλεσμα ότι έχουν αναπτυχθεί διαφορετικές τεχνικές/εφαρμογές σε διαφορετικά επίπεδα διαχείρισης πληροφορίας (π.χ από απλούς βρόχους ελέγχου έως συστήματα υποστήριξης αποφάσεων για περιβαλλοντολογικά δεδομένα). Οι Dahlek και Fox το 2016 [44], στην εργασία τους ασχολήθηκαν την περιβαλλοντική απόδοση μιας πόλης. Χρησιμοποίησαν 3 οντολογίες για την απάντηση ερωτήσεων που αφορούν ερωτήματα, όπως μόλυνση, αισθητήρες και τα είδη. Αυτό τους επέτρεψε να έχουν μία αυτοματοποιημένη ανάλυση απόδοσης μιας πόλης.

#### Συμπεράσματα

Στην παρούσα εργασία, προτείνεται από τους συγγραφείς την αρθρωτή οντολογία περιβαλλοντικής παρακολούθησης MEMOn, η οποία υποστηρίζει τη σημασιολογική διαλειτουργικότητα, την ενοποίηση και τη σύνδεση δεδομένων ώστε να υπάρχει μία συνολική παρουσίαση δεδομένων. Η μεθοδολογία που χρησιμοποιήθηκε σε αυτή την εργασία για την ανάπτυξη οντολογίας, είχε σημαντικό πλεονέκτημα τη στήριξη σε πραγματικά περιβαλλοντικά δεδομένα, τόσο για τη συλλογή του λεξιλογίου, όσο και για τη δοκιμή της οντολογίας. Οι ίδιοι, πιστεύουν πως, συμμορφώνονται με το λεξιλόγιο το οποίο μπορεί να εγγυηθεί ένα παγκόσμιο όραμα δεδομένων από πολλαπλές και διαφορετικές πηγές. Αυτό συμπεραίνεται από το γεγονός ότι η συνολική προβολή δεδομένων διασφαλίζεται από το γράφημα γνώσης το οποίο βασίζεται στην οντολογία. Ένα από τα σημαντικά πλεονεκτήματα ενός γραφήματος γνώσης, είναι η ευκολία που παρέχει στις διασυνδέσεις δεδομένων. Ακόμη, οι ειδικοί μπορούν με ευκολία να εξάγουν νέα γνώση από αυτά τα δεδομένα και επίσης να ανακαλύψουν τις σχέσεις μεταξύ των δεδομένων στην παρακολούθηση του περιβάλλοντος. Παρακάτω παρατίθενται οι κύριες συνεισφορές της συγκεκριμένης εργασίας:

1. Σημασιολογική αναπαράσταση του τομέα περιβαλλοντικής παρακολούθησης που επιτρέπει την εκφραστικότητα διαφορετικών περιβαλλοντικών κλάδων με επίσημο και δομημένο τρόπο.

2. Επαναχρησιμοποίηση οντολογιών ανωτέρου, μεσαίου επιπέδου τομέα για τη βελτίωση της διαλειτουργικότητας του MEMOn με τις υπάρχουσες οντολογίες.
3. Σπονδυλωτή αναπαράσταση του τομέα με την προδιαγραφή διαφορετικών ενοτήτων.

Σύμφωνα με τα συμπεράσματα των συγγραφέων, η οντολογία MEMOn, χάρη στην αρθρωτότητα της, προσφέρει την επαναχρησιμοποίηση της ως ξεχωριστές μονάδες. Ενώ ο κύριος σκοπός κατασκευής της οντολογίας MEMOn, ήταν να αντιπροσωπεύει πληροφορίες σχετικά με την περιβαλλοντική παρακολούθηση, πλέον η οντολογία μπορεί να χρησιμοποιηθεί και σε άλλους τομείς, καθώς και να χρησιμοποιηθεί σε άλλα πλαίσια.

Ακόμη, η συγκεκριμένη οντολογία μπορεί να επεκταθεί μελλοντικά καθώς η οντολογία MEMOn μπορεί να εξελιχθεί αν προστεθούν κλάσεις και ιδιότητες καθώς και να αλλάξουν οι κλάσεις και οι ιδιότητες. Επιπλέον, λόγω του ότι το MEMOn επαναχρησιμοποιεί οντολογίες αναφοράς, μπορούν άλλες οντολογίες να ενσωματωθούν και να εμπλουτίσουν το MEMOn.

#### Εργασία 4η

Οι F.Liu, Dong, Z.Liu και Chen, βασίζουν το άρθρο τους στις μετεωρολογικές καταστροφές καθώς και τα προβλήματα που δημιουργούνται από αυτές. Έτσι, θέλοντας να μειώσουν ή να αυξήσουν την πρόληψη ή ακόμη και να αποφύγουν στο μέγιστο βαθμό τα προβλήματα που αφήνουν πίσω μετεωρολογικές καταστροφές (Καταστροφές περιουσιών, Τραυματισμοί ανθρώπων, Απώλειες ζώων) αλλά και προκειμένου να ξεπεράσουν τα όρια του γνωστικού αποθέματος και της υποκειμενικότητας των κατασκευαστών στη διαδικασία κατασκευής μίας τεχνητής οντολογίας προτείνουν σε αυτό το άρθρο έναν εστιασμένο ανιχνευτή ώστε να αντλούν δεδομένα και πληροφορίες, στον οποίο χρησιμοποιήθηκε μία νέα οντολογία τομέα η οποία βασίζεται στην τεχνολογία εκμάθησης οντολογίας.

Λόγω του μεγάλου όγκου πληροφοριών στο διαδίκτυο σχετικά με την ενημέρωση για τις μετεωρολογικές προβλέψεις καθώς και τις μετεωρολογικές καταστροφές, δημιουργήθηκε η ανάγκη από τους συγγραφείς αυτού του άρθρου να δημιουργήσουν μία νέα οντολογία ανιχνευτή ώστε να αυξήσουν τη σωστή ταξινόμηση και εξαγωγή της πληροφορίας που αφορούν μετεωρολογικά δεδομένα, αποφεύγοντας τα ανακριβή ή ακόμη και λανθασμένα δεδομένα. Λόγω

της έλλειψης οντολογιών οι οποίες μπορούν να επαναχρησιμοποιηθούν στον πεδίο των μετεωρολογικών καταστροφών, οι συγγραφείς κατασκεύασαν μία νέα οντολογία. Η συγκεκριμένη οντολογία εστιάζει στην αξιολόγηση της συνάφειας μεταξύ μιας υπερ-σύνδεσης και ενός συγκεκριμένου θέματος. Για τη λειτουργία της συγκεκριμένης οντολογίας, οι συγγραφείς, πρότειναν ένα μοντέλο βελτιστοποίησης πολλαπλών στόχων για την αξιολόγηση υπερ-συνδέσμων καθώς και τον αλγόριθμο Multi-Objective Ant Colony Optimization (MOACO), ο οποίος αποτελεί μία μέθοδο της πλησιέστερης υποψήφιας λύσης. Στον αλγόριθμο αυτό, συνδυάζεται η γρήγορη ταξινόμηση χωρίς κυριαρχία με σκοπό την επιλογή ενός συνόλου βέλτιστων υπερ-συνδέσμων για τη σωστή καθοδήγηση αναζήτησης των ανιχνευτών. Από τα πειραματικά αποτελέσματα που έκαναν, κατέληξαν στο γεγονός ότι η εστιασμένη ανίχνευση των καταστροφών από τυφώνα και βροχοπτώσεων, απέδειξαν ότι η ικανότητα των προτεινόμενων εστιασμένων ανιχνευτών, μπορούν να ανακτούν σχετικές ιστοσελίδες με τις καταστροφές.

Για την κατασκευή της οντολογίας που χρησιμοποίησαν στο άρθρο τους, οι συγγραφείς αξιοποίησαν την τυπική ανάλυση εννοιών (Formal Concept Analysis – FCA) όπου βασίζεται ο σκελετός της, και για να εμπλουτίσουν τις έννοιες οντολογίας χρησιμοποίησαν τη λανθάνουσα κατανομή Dirichlet (LDA) και τον αλγόριθμο του Apriori. Για τη δημιουργία του σκελετού, χρησιμοποίησαν τον Κινέζικο θησαυρό ταξινόμησης συμβόλων με την αγγλική έκδοση και στην πορεία το σύγκριναν με τις έννοιες του WordNet.

Στις επόμενες παραγράφους θα γίνει μία επεξήγηση των ορισμών FCA, LDA και αλγορίθμου Apriori ώστε να υπάρχει μία πιο σαφή εικόνα για την ολοκληρωτική εικόνα της οντολογίας που κατασκεύασαν οι συγγραφείς.

#### Τυπική Ανάλυση Εννοιών (Formal Concept Analysis - FCA)

Σύμφωνα με τους Peng, Du, Hai, Chen και Gao, η FCA, είναι μία μαθηματική μέθοδος η οποία βασίζεται στην τυπική έκφραση της γνώσης και την ανάλυση δεδομένων [45]. Η μέθοδος αυτή, αναλύει δεδομένα από το επίσημο πλαίσιο, εξάγει τους κανόνες μεταξύ εννοιών και πραγματοποιεί την ταξινόμηση, την εμφάνιση και την ανακάλυψη εννοιών. Η κύρια δομή του είναι ένα μοντέλο γραφήματος το οποίο αποτελείται από πολλαπλούς κόμβους εννοιών. Ο κάθε κόμβος, αποτελείται από δύο μέρη, την επέκταση και τον συνειρμό. Στη μέθοδο FCA, η συλλογή

όλων των αντικειμένων σε μία έννοια, ονομάζεται επέκταση της έννοιας. Το σύνολο των χαρακτηριστικών που μοιράζονται τα αντικείμενα που περιέχονται σε μία έννοια, ονομάζεται συνειρμός της έννοιας.

Παρακάτω δίνεται ο ορισμός της μεθόδου FCA:

*Το επίσημο πλαίσιο μπορεί να οριστεί ως μια τριάδα  $F = (O, A, I)$ , όπου  $O = \{O_1, O_2, \dots, O_b\}$  και  $A = \{A_1, A_2, \dots, A_c\}$  είναι σύνολα και το  $I \subseteq O \times A$  είναι μία δυαδική σχέση. Τα στοιχεία του  $O$ , ονομάζονται αντικείμενα και τα στοιχεία του  $A$  ονομάζονται ιδιότητες. Εάν  $(O_i, A_j) \in I$  ( $1 \leq i \leq b$ ,  $1 \leq j \leq c$ ), το αντικείμενο  $O_i$  έχει χαρακτηριστικό  $A_j$ .*

Ένα επίσημο πλαίσιο, είναι ένας απλός τρόπος προσδιορισμού ποια αντικείμενα έχουν ποιες ιδιότητες. Είναι ένας πίνακας δεδομένων όπου οι γραμμές αντιστοιχούν σε αντικείμενα και οι στήλες σε χαρακτηριστικά.

#### Λανθάνουσα Κατανομή Dirichlet (Latent Dirichlet Allocation - LDA)

Σύμφωνα με τους Jelodar, Wang, Yuan, Feng, Jiang, Li και Zhao, η λανθάνουσα κατανομή Dirichlet, είναι ένα πιθανολογικό μοντέλο για τη μοντελοποίηση διακριτών συνόλων δεδομένων. Στο LDA, για ένα έγγραφο, κάθε λέξη στο έγγραφο λαμβάνεται μέσω μιας διαδικασίας επιλογής ενός συγκεκριμένου θέματος με μια συγκεκριμένη πιθανότητα και επιλογής μιας συγκεκριμένης λέξης από αυτό το θέμα, με μία συγκεκριμένη πιθανότητα [46].

Η πιθανότητα κάθε λέξης να εμφανίζεται σε ένα έγγραφο είναι:

$$P(\text{word}|\text{document}) = \sum_{\text{topic}} P(\text{word}|\text{topic}) \times P(\text{topic}|\text{document})$$

Εικόνα 10: Εξίσωση πιθανότητας εμφάνισης λέξης σε ένα έγγραφο. [46]

Η παραπάνω εξίσωση μπορεί να αναπαρασταθεί καλύτερα από το παρακάτω γράφημα:

$$\begin{array}{ccccc} \text{document} & & \text{topic} & & \text{document} \\ \text{word} & \boxed{\mathbf{C}} & = \text{word} & \boxed{\boldsymbol{\theta}} & \times \text{topic} & \boxed{\boldsymbol{\phi}} \end{array}$$

Εικόνα 11: Αναπαράσταση του μοντέλου LDA. [46]

Στην παραπάνω αναπαράσταση, το πρώτο σκέλος, «έγγραφο-λέξη» (document-word), αντιπροσωπεύεται η πιθανότητα εμφάνισης κάθε λέξης σε ένα έγγραφο. Στο δεύτερο σκέλος, «θέμα-λέξη» (topic-word), αντιπροσωπεύεται η πιθανότητα εμφάνισης κάθε λέξης για κάθε θέμα. Στο τρίτο σκέλος, «έγγραφο-θέμα» (document-topic), αντιπροσωπεύεται η πιθανότητα εμφάνισης κάθε θέματος σε κάθε έγγραφο.

Στο μοντέλο LDA, γίνεται υπολογισμός της συχνότητας κάθε λέξης σε κάθε έγγραφο από ένα σύνολο εγγράφων που έχουν αποκτηθεί για να ληφθεί μία σχέση «έγγραφο-λέξη». Μέσω αυτής της εκπαίδευσης σχέσης, λαμβάνεται η σχέση «θέμα-λέξη» καθώς και η σχέση «έγγραφο-θέμα». Στο LDA, είναι δύσκολο να προσδιοριστεί ο αριθμός των θεμάτων. Οι συγγραφείς σε αυτό το άρθρο, όρισαν πρώτα μια τιμή  $Q$  για τον αρχικό αριθμό θεμάτων και απόκτησαν το αρχικό μοντέλο μέσω της εκπαίδευσης. Σύμφωνα με το αρχικό μοντέλο, υπολογίζεται η ομοιότητα μεταξύ των θεμάτων. Στην πορεία, λαμβάνεται η μέση ομοιότητα αυτών των θεμάτων. Ανάλογα με την αύξηση ή μείωση της τιμής  $Q$  που όρισαν οι συγγραφείς, το μοντέλο επανεκπαιδεύεται για τον υπολογισμό της ομοιότητας μεταξύ των θεμάτων και της μέσης ομοιότητας τους. Αργότερα, συνέκριναν όλες τις μέσες ομοιότητες και βρήκαν την ελάχιστη μέση ομοιότητα. Τέλος, η τιμή  $Q$  που αντιστοιχεί στην ελάχιστη μέση ομοιότητα, θεωρείται ο αριθμός των θεμάτων στην λανθάνουσα κατανομή του Dirichlet.

#### Αλγόριθμος Apriori

Σύμφωνα με τον Sornalakshmi, η βασική ιδέα του αλγορίθμου Apriori, είναι να υπολογίσει την υποστήριξη του συνόλου στοιχείων σαρώνοντας το σύνολο δεδομένων πολλές φορές και να βρει όλα τα συχνά στοιχεία του συνόλου, ώστε να δημιουργηθούν οι κανόνες συσχέτισης [47].

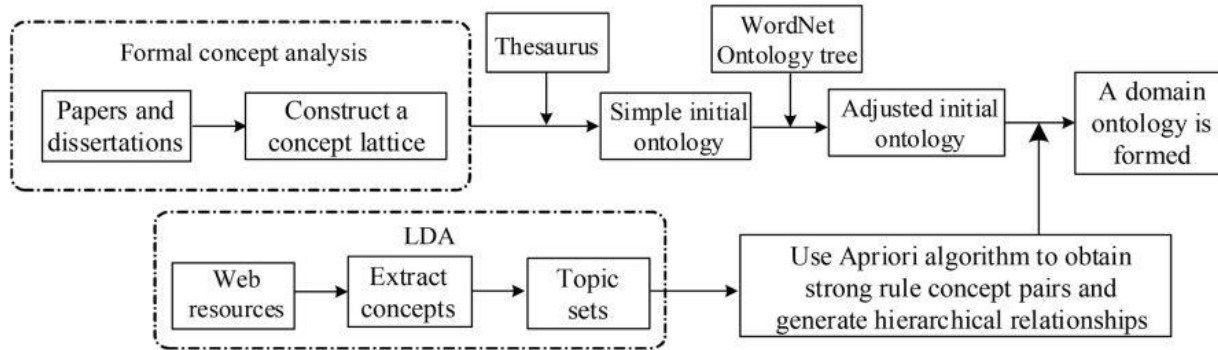
Στον αλγόριθμο Apriori, εάν υπάρχει συχνότητα εμφάνισης ενός συνόλου στοιχείων, όλα τα υποσύνολα στοιχείων του συνόλου, έχουν την ίδια συχνότητα εμφάνισης. Επίσης, εάν δεν υπάρχει συχνότητα εμφάνισης ενός συνόλου στοιχείων, τότε όλα τα υπερσύνολα στοιχείων του συνόλου δεν έχουν ούτε αυτά συχνότητα εμφάνισης.

### Δομή Κατασκευής της Οντολογίας

Η δομή κατασκευής της οντολογίας που ακολούθησαν αποτελείται από τα παρακάτω βήματα:

1. Αναζήτησαν περίπου 20 εργασίες από το China National Knowledge Internet (CNKI) οι οποίες είχαν στη βάση δεδομένων το συγκεκριμένο θέμα. Το CNKI ψηφιοποιεί το 80% των πληροφοριών καθώς και τους πόρους γνώσης της Κίνας, μέσα στις οποίες περιλαμβάνονται διατριβές και εργασίες. Από αυτές τις εργασίες εξάγονται ως σύνολο όρων οι τίτλοι, οι περιλήψεις καθώς και λέξεις-κλειδιά. Στην πορεία γίνεται τμηματοποίηση με τη χρήση του δείκτη IK-analysis και γίνεται η μέτρηση συχνότητας λέξεων.
2. Χρησιμοποίησαν την τυπική ανάλυση εννοιών (FCA) με βάση κάποιους όρους, καθώς και τη συχνότητα των λέξεων ώστε να κατασκευάσουν ένα πλέγμα εννοιών. Σε αυτό το πλέγμα, ένας όρος στο πλέγμα, λαμβάνεται ως χαρακτηριστικό και κατασκευάζεται μία απλή αρχική οντολογία χρησιμοποιώντας το πλέγμα, καθώς και τον κινέζικο θησαυρό ταξινόμησης (με την αγγλική του έκδοση).
3. Σύγκριναν τις έννοιες στην αρχική οντολογία με τις έννοιες από το WordNet ώστε να δουν τη συχνότητα εμφάνισης εννοιών. Εάν η έννοια A εμφανίζεται τόσο στην αρχική οντολογία, όσο και στο WordNet, βρήκαν τους κόμβους έννοιας κοντά στην έννοια A στο WordNet και χρησιμοποίησαν τη συνάρτηση μη γραμμικού συνδυασμού. Στην πορεία, οι συγγραφείς όρισαν ένα όριο  $\gamma$ . Εάν η συσχέτιση της έννοιας B από το WordNet είναι μεγαλύτερη από το  $\gamma$ , τότε η έννοια B προστίθεται στην αρχική οντολογία.
4. Αναζήτησαν λέξεις-κλειδιά μέσω των δημοφιλών μηχανών αναζήτησης όπως Google, Bing κ.λπ. και επέλεξαν τις 30 πρώτες ιστοσελίδες στην κατάταξη.
5. Ανέλυσαν και τμηματοποίησαν το κείμενο κάθε ιστοσελίδας χρησιμοποιώντας τον δείκτη ανάλυσης IKAnalyzer ώστε να επιλεχθούν τα ουσιαστικά ως το σύνολο υποψήφιων εννοιών.
6. Χρησιμοποίησαν το μοντέλο LDA για να ανακτήσουν 20 σύνολα θεμάτων από το σύνολο.
7. Τέλος, όρισαν το ελάχιστο όριο υποστήριξης Q και χρησιμοποίησαν τον κανόνα συσχέτισης του αλγορίθμου Apriori ώστε να αποκτήσουν ζεύγη εννοιών μεγαλύτερα από Q, από το σύνολο των 20 θεμάτων. Εφόσον τα ζεύγη αυτά επεκταθούν στην αρχική οντολογία, σχηματίζεται μία οντολογία τομέα.

Στο παρακάτω γράφημα απεικονίζεται η διαδικασία κατασκευής της αρχικής οντολογίας



Εικόνα 12: Διαδικασία κατασκευής αρχικής οντολογίας. [12]

Έτσι, σύμφωνα με τη διαδικασία κατασκευής της οντολογίας που αναφέρεται παραπάνω, οι συγγραφείς κατασκεύασαν δύο οντολογίες τομέα. Μία που αφορά τις καταστροφές από τυφώνες και άλλη μία που αφορά τις καταστροφές από τις καταιγίδες.

#### Οντολογία τυφώνων

Για να εξακριβώσουν τη συνάφεια της κάθε ιστοσελίδας, εξήγαγαν όλες τις έννοιες από την κατασκευασμένη οντολογία και δημιούργησαν ένα διάνυσμα  $TK = \{tk1, tk2, \dots, tkh\}$ , όπου  $h$  είναι ο αριθμός των λέξεων. Στη συνέχεια, υπολόγισαν τη σημασιολογική ομοιότητα μεταξύ των εννοιών από την ιστοσελίδα με αυτές από το διάνυσμα με βάση την οντολογία. Σύμφωνα με τα αποτελέσματα από το διάνυσμα εννοιών, δίνεται η αντίστοιχη προτεραιότητα επίσκεψης ή μη επίσκεψης του υπερ-συνδέσμου, ανάλογα με την συνάφεια.

#### Υπολογισμός Σημασιολογικής Ομοιότητας

Στην πορεία, οι συγγραφείς ώστε να υπολογίσουν ορθά τη σημασιολογική ομοιότητα, εξέτασαν τους παράγοντες επιρροής της ομοιότητας οι οποίοι είναι:

1. Σημασιολογική απόσταση (IFDis)
2. Πυκνότητα εννοιών (IFDen)
3. Βάθος εννοιών (IFDep)
4. Βαθμός σύμπτωσης εννοιών (IFCoi)
5. Σημασιολογική σχέση εννοιών (IFRel)

Παρακάτω δίνεται ο ορισμός για τη σημασιολογική σχέση εννοιών:

Στις οντολογίες τομέα, θεωρούνται πέντε σημασιολογικές σχέσεις οι οποίες είναι, «Συνώνυμα» (Synonym), «είναι-α» (is-a), «μέρος-του» (part-of) και «συντεταγμένοι-όροι-του» (coordinate-term-of).

#### Εξακρίβωση συνάφειας των ιστοσελίδων

Οι περισσότερες ιστοσελίδες στον παγκόσμιο ιστό χρησιμοποιούν τη γλώσσα HTML (Hypertext Markup Language). Είναι γλώσσα σήμανσης υπερκειμένου, η οποία είναι πολύ απλή και ανεξάρτητη. Για την εμφάνιση του περιεχομένου στην ιστοσελίδα, η HTML χρησιμοποιεί ετικέτες. Έτσι, οι συγγραφείς χώρισαν τις κύριες ετικέτες σε 5 διαφορετικές ομάδες. Στην G1 = (Τίτλος, λέξη-κλειδί, περιγραφή, τίτλος πρώτου επιπέδου), στην G2 = (Τίτλος δευτέρου επιπέδου, τίτλος τρίτου επιπέδου), στην G3 = (Τίτλος τέταρτου επιπέδου, τίτλος πέμπτου επιπέδου, τίτλος έκτου επιπέδου, έντονο κείμενο), στην G4 = (Πληροφορίες σώματος) και G5 = (Πληροφορίες εκτός σώματος). Ακόμη, όρισαν διαφορετικά βάρη σε κάθε ομάδα  $W = (2, 1.5, 1.2, 1.0$  και  $0.2)$ . Στην πορεία, αντιστοίχησαν το κείμενο μιας ιστοσελίδας σε ένα διάνυσμα χαρακτηριστικών  $DK = (dk_1, dk_2, \dots, dk_h)$ , καθώς και σε ένα διάνυσμα βάρους χαρακτηριστικών  $WDK = (w_{dk1}, w_{dk2}, \dots, w_{dk1}, \dots, w_{dkh})$  το οποίο το υπολόγισαν σύμφωνα με την παρακάτω εξίσωση:

$$w_{dk_i} = \sum_{l=1}^L tf_{i,l} \times W_l = \sum_{l=1}^L \left( \frac{f_{i,l}}{\max f_{i,l}} \times W_l \right)$$

Εικόνα 13: Εξίσωση συνάφειας. [12]

**tf<sub>i,l</sub>**: Είναι ο όρος συχνότητας (TermFrequency - TF) μετά την κανονικοποίηση της i-ης θεματικής λέξης στην l-η θέση του κειμένου της ιστοσελίδας.

**f<sub>i,j</sub>**: Είναι ο όρος συχνότητας της λέξης θέματος στην l-η θέση του κειμένου της ιστοσελίδας.

**maxf<sub>i,j</sub>**: Είναι ο μέγιστος όρος συχνότητας των ομάδων στις οποίες εμφανίζεται η i-η λέξη θέματος σε όλες τις εμφανίσεις του κειμένου της ιστοσελίδας.

**L**: Είναι ο αριθμός των ομάδων ετικετών, όπου στην προκειμένη περίπτωση είναι  $L = 5$ .

**W<sub>i</sub>**: Είναι το βάρος της i-ης ετικέτας.

Στην πορεία, χρησιμοποίησαν το μοντέλο διανυσματικού χώρου (Vector Space Model – VSM), ώστε να υπολογίσουν τη συνάφεια θέματος του κειμένου της ιστοσελίδας, δηλαδή να υπολογίσουν το συνημίτονο του θέματος λέξης-βάρους διάνυσμα WTK και το χαρακτηριστικό κείμενο-βάρους διάνυσμα WDK.

### **Διαδικασία κατασκευής εστιασμένου ανιχνευτή με Οντολογία και τον αλγόριθμο MOACO**

Οι συγγραφείς, πρότειναν έναν εστιασμένο ανιχνευτή ο οποίος χρησιμοποιεί την εκμάθηση οντολογίας που προαναφέραμε παραπάνω, καθώς και τις στρατηγικές MOACO (Multi-Objective Ant Colony Optimization Algorithm) για τη τις γνώσεις των μετεωρολογικών καταστροφών. Στην αρχή, χρησιμοποίησαν λέξεις κλειδιά οι οποίες έχουν σχέση με το συγκεκριμένο θέμα, ώστε να αναζητήσουν τις 100 πιο δημοφιλείς ιστοσελίδες με φυλλομετρητή το Google, Firefox και το Bing. Στη συνέχεια, διάλεξαν τις 30 από τις 100 ιστοσελίδες, οι οποίες ορίστηκαν ως οι αρχικοί υπερ-σύνδεσμοι Link\_Init. Πραγματοποίησαν αναζήτηση σε ιστοσελίδες που αφορούν τους τυφώνες χρησιμοποιώντας ένα σύνολο λέξεων κλειδιά όπως «τυφώνας, κύμα καταιγίδας, άνεμος, τυφώνας» (typhoon, storm surge, wind, hurricane). Παράλληλα, χρησιμοποίησαν λέξεις κλειδιά όπως «βροχόπτωση, καταιγίδα, πλημμύρα» (rainfall, rainstorm, flood), για την αναζήτηση σε ιστοσελίδες που σχετίζονται με τις καταιγίδες, που αφορά τη δεύτερη οντολογία για καταιγίδες.

### **Παρόμοιες μελέτες**

Οι Naing, Lim και Liam το 2002 [48], ασχολήθηκαν με την αντιμετώπιση του προβλήματος της ετερογένειας. Ετερογενείς και αυτόνομοι ιστότοποι περιέχουν πληροφορίες οι οποίες δεν μπορούν εύκολα να υποστούν επεξεργασία από αυτοματοποιημένα εργαλεία. Για την αντιμετώπιση αυτού του προβλήματος πρότειναν ένα σχήμα σχολιασμού το οποίο βασίζεται στην οντολογία OWA (Ontology-based Web Annotation). Οι Asano, Tezuka και Nishizeki το 2008 [49], πρότειναν στην εργασία τους αρκετούς τρόπους βελτίωσης του ήδη υπάρχοντος αλγορίθμου HITS (Hyperlinked Induced Topic Search), ο οποίος έχει σκοπό να προσφέρει στο χρήστη πιο στοχευμένη αναζήτηση, η οποία γίνεται μέσω του υψηλού σκορ που δίνει ο αλγόριθμος σε κάθε σελίδα κάνοντας αναζήτηση των λέξεων κλειδιών μέσα στις εκάστοτε σελίδες. Έτσι εμφανίζει πρώτα αυτές με το υψηλότερο σκορ. Έτσι, πρότειναν 3 δικούς τους αλγορίθμους οι οποίοι βελτιώνουν τον HITS κάνοντας την αναζήτηση ακόμη πιο στοχευμένη ακόμη και σε σελίδες οι οποίες δεν μπορούν να βρεθούν εύκολα από τη γνωστή μηχανή αναζήτησης της Google.

### Συμπεράσματα

Για να ελέγξουν τα αποτελέσματα των πειραμάτων του ανιχνευτή, οι συγγραφείς, έτρεξαν το πείραμα σε δύο διαφορετικές περιπτώσεις. Η μία αφορά τις καταστροφές από τυφώνα και η άλλη αφορά τις καταστροφές από τις βροχοπτώσεις. Όλοι οι αλγόριθμοι που χρησιμοποιήθηκαν, είναι στη γλώσσα Java και ο υπολογιστής που χρησιμοποιήθηκε είχε επεξεργαστή Intel Core i7-7700HQ σε 2.80GHz συχνότητα, καθώς και 8.0GB μνήμης RAM.

Οι γενικοί ανιχνευτές υστερούν σε σχέση με τους εστιασμένους ανιχνευτές διότι δεν μπορούν να παρέχουν ακριβή αποτελέσματα αναζήτησης για συγκεκριμένες γνώσεις τομέα. Σε αντίθεση με τους γενικούς ανιχνευτές, οι εστιασμένοι ανιχνευτές είναι προσαρμοσμένοι σε συγκεκριμένα θέματα. Με σκοπό να αποφευχθούν τα προβλήματα που ήδη αντιμετωπίζουν οι γενικοί ανιχνευτές, οι συγγραφείς υπολόγισαν τη συνάφεια των θεμάτων που βασίζονται στη σημασιολογία, κάνοντας τους εστιασμένους ανιχνευτές να έχουν στρατηγική ανίχνευσης.

Αρχικά κατασκεύασαν δύο οντολογίες τομέα, μία που αφορά τις καταστροφές από τυφώνα και άλλη μία που αφορά τις καταστροφές από τις βροχοπτώσεις, οι οποίες έχουν ως βάση την εκμάθηση οντολογίας. Στη συνέχεια χρησιμοποίησαν την οντολογία, ως μοντέλο αναφοράς των θεμάτων. Στη συνέχεια, αξιολόγησαν τη συνάφεια των υπερ-συνδέσμων με το συγκεκριμένο θέμα (τυφώνες – βροχοπτώσεις) και διαπίστωσαν ότι ο παραδοσιακός τρόπος σταθμισμένου αθροίσματος έχει το μειονέκτημα ότι, ο προσδιορισμός των βέλτιστων βαρών είναι αρκετά δύσκολο. Έτσι, για την αποφυγή των προβλημάτων προσδιορισμού, οι συγγραφείς εισήγαγαν το μοντέλο βελτιστοποίησης πολλαπλών στόχων ώστε να γίνεται πιο στοχευμένα η αξιολόγηση του ανιχνευτή.

Αν και η μέθοδος των συγγραφέων είναι ανώτερη από τις υπόλοιπες που υπάρχουν αυτή τη στιγμή, ακόμη και αυτή έχει κάποια ελαττώματα. Συγκεκριμένα, τα ελαττώματα αφορούν κυρίως το χρόνο εκτέλεσης. Επειδή είναι εφικτό μία άσχετη ιστοσελίδα να συνδέσει μία σχετική ιστοσελίδα, απαιτείται να γίνεται προσπέλαση και άσχετων ιστοσελίδων από το θέμα των τυφώνων και των βροχοπτώσεων, ώστε να καταλήξουν στις σχετικές ιστοσελίδες.

### Σύγκριση 3<sup>ης</sup> και 4<sup>ης</sup> εργασίας

Στην 3<sup>η</sup> εργασία, δηλαδή αυτή των Masmoudi, Karray, Lamine, Zghal και Archimede [11], παρουσιάζεται μία οντολογία (MEMOn) η οποία έχει ως τομέα την παρακολούθηση του περιβάλλοντος βάση των δεδομένων που συλλέγονται. Η συγκεκριμένη οντολογία έχει ως κύριο σκοπό τη σημασιολογική διαλειτουργικότητα μετεωρολογικών δεδομένων, τα οποία δεδομένα είναι πάρα πολλά και συνεχίζουν να επεκτείνονται καθώς επίσης, αυτά τα δεδομένα αποθηκεύονται σε πολλές και διαφορετικές πηγές. Ακόμη, στην οντολογία MEMOn, γίνεται σύνδεση ετερογενών δεδομένων τα οποία συλλέγονται μέσω τεχνικών και συστημάτων παρατήρησης. Τα αποτελέσματά τους ήταν ότι, η συγκεκριμένη οντολογία μπορεί να επαναχρησιμοποιηθεί, πράγμα που είναι πολύπλοκο ή ακόμη και αδύνατο για πολλές οντολογίες που ήδη υπάρχουν, σε άλλους τομείς πέραν αυτού, δηλαδή της μετεωρολογίας. Επίσης, κατέληξαν στο γεγονός ότι, η συγκεκριμένη οντολογία παρέχει ευκολία στις διασυνδέσεις μεταξύ δεδομένων διαφορετικών πηγών, κάνοντας έτσι τη ζωή των ειδικών, ακόμη πιο εύκολη ώστε να καταλήξουν σε πιο ακριβή και στοχευμένα συμπεράσματα στην παρακολούθηση του περιβάλλοντος. Τέλος, κατέληξαν στο γεγονός, ότι παρόλο που η οντολογία MEMOn έχει μεγάλη πλέον σημασία στην παρακολούθηση του περιβάλλοντος, θα μπορούσε μελλοντικά να επεκταθεί και να εξελιχθεί ακόμη και σε άλλους τομείς.

Από την άλλη, στην 4<sup>η</sup> εργασία, αυτή των J.Liu, Dong, Z.Liu και Chen [12], κατασκεύασαν μία οντολογία ανιχνευτή η οποία έχει ως κύριο σκοπό την αύξηση σωστής εξαγωγής και ταξινόμησης της πληροφορίας που αφορούν μετεωρολογικά δεδομένα, με σκοπό να αποφύγουν ανακριβείς ή και ακόμη, λανθασμένες πληροφορίες. Η οντολογία αυτή, εστιάζει στην αξιολόγηση της συνάφειας ενός υπερ-συνδέσμου και ενός συγκεκριμένου θέματος. Τα αποτελέσματά τους ήταν ότι, οι γενικοί ανιχνευτές που ήδη υπήρχαν υστερούσαν σε σχέση με τους εστιασμένους ανιχνευτές καθώς δεν αποδίδουν σωστά και ακριβή αποτελέσματα για συγκεκριμένες γνώσεις τομέα. Οι συγγραφείς, υπολόγισαν τη συνάφεια των θεμάτων που βασίζονται στη σημασιολογία, καθώς φρόντισαν οι εστιασμένοι ανιχνευτές να έχουν συγκεκριμένη στρατηγική ανίχνευσης. Για να αποφύγουν τα προβλήματα προσδιορισμού, οι συγγραφείς παρουσίασαν το μοντέλο βελτιστοποίησης πολλαπλών στόχων με αποτέλεσμα, η αξιολόγηση του ανιχνευτή να είναι περισσότερο στοχευμένη. Τέλος, παρά το γεγονός ότι οι γενικοί ανιχνευτές υστερούν κατά πολύ

με τους εστιασμένους ανιχνευτές της οντολογίας που κατασκεύασαν οι συγγραφείς, παρατήρησαν κάποια ελαττώματα τα οποία θα μπορούν να μην υπάρχουν στο μέλλον. Ένα από αυτά ήταν ο χρόνος εκτέλεσης για την εύρεση του σωστού υπερ-συνδέσμου. Άλλο ελάττωμα που παρατήρησαν ήταν πρέπει να γίνει προσπάθεια και άσχετων ιστοσελίδων με αποτέλεσμα να παρουσιάσει ο ανιχνευτής τις σχετικές ιστοσελίδες που αφορούν τις μετεωρολογικές αναζητήσεις. Παρατηρούμε ότι στην 3<sup>η</sup> εργασία με την οντολογία MEMOn, οι συγγραφείς είχαν την ανάγκη να κατασκευάσουν τη συγκεκριμένη οντολογία λόγω των εκατοντάδων εκατομμυρίων δεδομένων σχετικά με το κλίμα όπως προκύπτει από τις εφαρμογές των συστημάτων παρατήρησης της γης. Έτσι, παρουσίασαν την οντολογία MEMOn η οποία έχει ως σκοπό την απλοποίηση και διασύνδεση των δεδομένων. Από την άλλη, βλέπουμε ότι στην 4<sup>η</sup> εργασία, οι συγγραφείς έκαναν κάτι παρόμοιο, δηλαδή την απλοποίηση και στοχευμένη εξαγωγή πληροφοριών, συγκεκριμένα από το διαδίκτυο.

## Βιβλιογραφία

- [1] T. R. Gruber, “A Translation Approach to Portable Ontology Specifications,” 1993.
- [2] K. Kotis, G. A. Vouros, and K. Stergiou, “Towards Automatic Merging of Domain Ontologies: The HCONE-merge approach.”
- [3] A. LeClair and R. Khedri, “Conto: A Protégé Plugin for Configuring Ontologies,” in *Procedia Computer Science*, 2016, vol. 83, pp. 179–186. doi: 10.1016/j.procs.2016.04.114.
- [4] “Type of Ontologies.” <https://web.fe.up.pt/~eol/SOCRATES/Palzer/ontologies.htm> (accessed Sep. 24, 2022).
- [5] D. T. Santosh, “Keyword and URI: An Analysis in Semantic Content Retrieval.” [Online]. Available: <https://www.researchgate.net/publication/257413463>
- [6] Cox Simon and Little Chris, “Sensor Web Enablement,” Jul. 2020. Accessed: Sep. 21, 2022. [Online]. Available: <https://www.w3.org/TR/vocab-owl-time-rel/>
- [7] “Sentinel-2-infographic”, Accessed: Sep. 22, 2022. [Online]. Available: <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2>
- [8] “SERVIR-Viz 3-D Visualization Tool,” 2008. <https://servirglobal.net/Global/Articles/Article/460/servir-viz-3-d-visualization-tool> (accessed Sep. 24, 2022).
- [9] M. Chen and B. Plale, “From Metadata to Ontology Representation: A Case of Converting Severe Weather Forecast Metadata to an Ontology.” [Online]. Available: <http://newsinfo.iu.edu/news/page/normal/14369.html>
- [10] S. M. Chen and M. H. Huang, “Automatically generating the weather news summary based on fuzzy reasoning and ontology techniques,” *Inf Sci (N Y)*, vol. 279, pp. 746–763, Sep. 2014, doi: 10.1016/j.ins.2014.04.027.
- [11] M. Masmoudi, M. H. Karray, S. ben Abdallah Ben Lamine, H. B. Zghal, and B. Archimede, “MEMOn: Modular Environmental Monitoring Ontology to link heterogeneous Earth observed data,” *Environmental Modelling and Software*, vol. 124, Feb. 2020, doi: 10.1016/j.envsoft.2019.104581.
- [12] J. Liu, Y. Dong, Z. Liu, and D. Chen, “Applying ontology learning and multi-objective ant colony optimization method for focused crawling to meteorological disasters domain knowledge,” *Expert Syst Appl*, vol. 198, Jul. 2022, doi: 10.1016/j.eswa.2022.116741.
- [13] M. Ferdinand, C. Zirpins, and D. Trastour, “Lifting XML Schema to OWL.”
- [14] S. Battle, “Gloze: XML to RDF and back again,” 2006. [Online]. Available: [http://www.hpl.hp.com/personal/steve\\_battle](http://www.hpl.hp.com/personal/steve_battle)
- [15] D. van Deursen, C. Poppe, G. Martens, E. Mannens, and R. van de Walle, “XML to RDF Conversion: a Generic Approach.” [Online]. Available: <http://multimedialab.elis.ugent.be>

- [16] Martens Gaetan, Verborgh Ruben, Poppe Chris, and van de Walle Rik, "Lifting a Metadata Model to the Semantic Multimedia World," 2011.
- [17] "ChinaTimes." [http://tol.chinatimes.com/CT\\_NS/CTSearch.aspx](http://tol.chinatimes.com/CT_NS/CTSearch.aspx) (accessed Sep. 24, 2022).
- [18] A. Abuobieda, N. Salim, A. T. Albaham, A. H. Osman, and Y. J. Kumar, "Text summarization features selection method using pseudo genetic-based model," in *Proceedings - 2012 International Conference on Information Retrieval and Knowledge Management, CAMP'12*, 2012, pp. 193–197. doi: 10.1109/InfRKM.2012.6204980.
- [19] M. Chandra, V. Gupta, and S. K. Paul, "A statistical approach for automatic text summarization by extraction," in *Proceedings - 2011 International Conference on Communication Systems and Network Technologies, CSNT 2011*, 2011, pp. 268–271. doi: 10.1109/CSNT.2011.65.
- [20] P. Chen and R. Verma, "A query-based medical information summarization system using ontology knowledge," in *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 2006, vol. 2006, pp. 37–42. doi: 10.1109/CBMS.2006.25.
- [21] Ieee, *2012 International Conference on Advances in Engineering, Science and Management*. IEEE, 2012.
- [22] X. Y. Jiang, "Chinese automatic text summarization based on keyword extraction," in *Proceedings - 2009 1st International Workshop on Database Technology and Applications, DBTA 2009*, 2009, pp. 225–228. doi: 10.1109/DBTA.2009.9.
- [23] C. J. Jiang, H. Peng, Q. L. Ma, and J. C. Chen, "Automatic summarization for Chinese text based on combined words recognition and paragraph clustering," in *3rd International Symposium on Intelligent Information Technology and Security Informatics, IITSI 2010*, 2010, pp. 591–594. doi: 10.1109/IITSI.2010.15.
- [24] K. Kaikhah, "Automatic Text Summarization with Neural Networks," 2004.
- [25] S. Manne, S. M. Z. Pervez, and S. S. Fatima, "A novel automatic text summarization system with feature terms identification," 2011. doi: 10.1109/INDCON.2011.6139386.
- [26] A. A. Mohamed, "Generating user-focused, content-based summaries for multi-documents using document graphs," in *Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology*, 2005, vol. 2005, pp. 675–679. doi: 10.1109/ISSPIT.2005.1577178.
- [27] H. Nguyen, E. Santos, and J. Russell, "Evaluation of the impact of user-cognitive styles on the assessment of text summarization," *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, vol. 41, no. 6, pp. 1038–1051, Nov. 2011, doi: 10.1109/TSMCA.2011.2116001.
- [28] K. S. Thakkar, R. v. Dharaskar, and M. B. Chandak, "Graph-based algorithms for Text Summarization," in *Proceedings - 3rd International Conference on Emerging Trends in Engineering and Technology, ICETET 2010*, 2010, pp. 516–519. doi: 10.1109/ICETET.2010.104.
- [29] S. T. Yuan and J. Sun, "Ontology-based structured cosine similarity in document summarization: With applications to mobile audio-based knowledge management," *IEEE Transactions on*

- Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 35, no. 5, pp. 1028–1040, Oct. 2005, doi: 10.1109/TSMCB.2005.850153.
- [30] W. Yang, R. Dai, and X. Cui, “A novel Chinese text summarization approach using sentence extraction based on kernel words recognition,” in *Proceedings - 5th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008*, 2008, vol. 4, pp. 134–139. doi: 10.1109/FSKD.2008.20.
  - [31] P. Y. Zhang and C. H. Li, “Automatic text summarization based on sentences clustering and extraction,” in *Proceedings - 2009 2nd IEEE International Conference on Computer Science and Information Technology, ICCSIT 2009*, 2009, pp. 167–170. doi: 10.1109/ICCSIT.2009.5234971.
  - [32] C. S. Lee, Y. J. Chen, and Z. W. Jian, “Ontology-based fuzzy event extraction agent for Chinese e-news summarization,” *Expert Syst Appl*, vol. 25, no. 3, pp. 431–447, Oct. 2003, doi: 10.1016/S0957-4174(03)00062-9.
  - [33] C. S. Lee, Z. W. Jian, and L. K. Huang, “A fuzzy ontology and its application to news summarization,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 35, no. 5, pp. 859–880, Oct. 2005, doi: 10.1109/TSMCB.2005.845032.
  - [34] M. Fernández-López, “Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web Linkset Quality View project NeOn-Lifecycle Support for Networked Ontologies View project,” 2004. [Online]. Available: <https://www.researchgate.net/publication/230771114>
  - [35] *An Agile Methodology for developing ontology modules which can be used to develop modular ontologies*. 2013. [Online]. Available: <https://www.researchgate.net/publication/307638746>
  - [36] M. Grüninger and M. S. Fox, “The Role of Competency Questions in Enterprise Engineering,” 1995, pp. 22–31. doi: 10.1007/978-0-387-34847-6\_3.
  - [37] “Sahara and Sahel Observatory.” [oss-online.org](http://oss-online.org) (accessed Sep. 24, 2022).
  - [38] P. Doran, “Ontology Modularization: Principles and Practice,” 2009.
  - [39] N. F. Noy and D. L. McGuinness, “Ontology Development 101: A Guide to Creating Your First Ontology.” [Online]. Available: [http://translate.googleusercontent.com/translate\\_c?anno=2&hl=...opment/ontology101.pdf&usg=ALkJrhjVKWkRnOxyPTCMqp98iRSKKnZ9HQ](http://translate.googleusercontent.com/translate_c?anno=2&hl=...opment/ontology101.pdf&usg=ALkJrhjVKWkRnOxyPTCMqp98iRSKKnZ9HQ)
  - [40] G. Goos *et al.*, “Lecture Notes in Computer Science 7540.” [Online]. Available: <http://www.springer.com/series/7409>
  - [41] P. L. Buttigieg, E. Pafilis, S. E. Lewis, M. P. Schildhauer, R. L. Walls, and C. J. Mungall, “The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperability,” *J Biomed Semantics*, vol. 7, no. 1, Sep. 2016, doi: 10.1186/s13326-016-0097-6.
  - [42] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, and S. E. Lewis, “The environment ontology: contextualising biological and biomedical entities,” 2013. [Online]. Available: [www.environmentontology.org](http://www.environmentontology.org)

- [43] L. Corominas, M. Garrido-Baserba, K. Villez, G. Olsson, U. Cortés, and M. Poch, "Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques," *Environmental Modelling and Software*, vol. 106, pp. 89–103, Aug. 2018, doi: 10.1016/j.envsoft.2017.11.023.
- [44] D. Dahleh, M. S. Fox, and M. S. Fox, "An Environmental Ontology for Global City Indicators (ISO 37120) PolisGnosis Project: Representing and Analysing City Indicators View project Analysis Methods for Urban Infrastructure Systems View project An Environmental Ontology for Global City Indicators (ISO 37120)," 2016, doi: 10.13140/RG.2.2.10776.21769.
- [45] Q. Peng, Y. Du, Y. Hai, S. Chen, and Z. Gao, "Topic-specific crawling on the web with concept context graph based on FCA," 2009. doi: 10.1109/ICMSS.2009.5302301.
- [46] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimed Tools Appl*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019, doi: 10.1007/s11042-018-6894-4.
- [47] M. Sornalakshmi *et al.*, "Hybrid method for mining rules based on enhanced Apriori algorithm with sequential minimal optimization in healthcare industry," *Neural Comput Appl*, Jul. 2020, doi: 10.1007/s00521-020-04862-2.
- [48] M. M. Naing, E. P. Lim, and D. G. Hoe-Lian, "Ontology-based web annotation framework for hyperlink structures," in *WISE 2002 - Proceedings of the 3rd International Conference on Web Information Systems Engineering Workshops*, 2002, pp. 184–193. doi: 10.1109/WISEW.2002.1177862.
- [49] Y. Asano, Y. Tezuka, and T. Nishizeki, "Improvements of HITS algorithms for spam links," *IEICE Trans Inf Syst*, vol. E91-D, no. 2, pp. 200–208, 2008, doi: 10.1093/ietisy/e91-d.2.200.