



**ΜΙΑ ΑΝΟΙΧΤΗ ΚΑΙ ΟΛΟΚΛΗΡΩΜΕΝΗ  
ΠΛΑΤΦΟΡΜΑ ΣΥΝΕΡΓΑΤΙΚΗΣ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ**  
**AN OPEN AND INTEGRATED COLLABORATIVE  
BIOINFORMATICS PLATFORM**

**ΔΡΑΣΗ ΕΘΝΙΚΗΣ ΕΜΒΕΛΕΙΑΣ: «ΕΡΕΥΝΩ-ΔΗΜΙΟΥΡΓΩ-ΚΑΙΝΟΤΟΜΩ»**  
**ΚΩΔΙΚΟΣ ΕΡΓΟΥ : Τ1ΕΔΚ-05275**

**ΠΑΡΑΔΟΤΕΟ Π1**

**ΜΕΛΕΤΗ ΑΠΑΙΤΗΣΕΩΝ ΧΡΗΣΤΩΝ –  
ΑΡΧΙΚΗ ΠΡΩΤΟΤΥΠΗ ΥΛΟΠΟΙΗΣΗ**

ΕΕ1: Μελέτη απαιτήσεων χρηστών και καθορισμός προδιαγραφών

|                       |              |
|-----------------------|--------------|
| Είδος Παραδοτέου:     | Έκθεση       |
| Υπεύθυνος Φορέας:     | ΙΤΕ-ΙΠ       |
| Ημερομηνία Παράδοσης: | Μάρτιος 2019 |

<sup>2</sup> Κ: Κοινοπραξία έργου, ΕΕ: ΕΥΔΕ-ΕΤΑΚ, Δ: Δημόσιο έγγραφο (ελεύθερο, π.χ., Web-Site έργου)

## ΠΕΡΙΛΗΨΗ

Η επιστημονική έρευνα στη μεταγονιδιωματική βιοϊατρική και στην ανάλυση μεγάλου όγκου ετερογενών βιο-δεδομένων κινείται σε ένα άκρως ανταγωνιστικό περιβάλλον όπου η ταχύτητα, η ακρίβεια και η ποιότητα των ερευνητικών αποτελεσμάτων είναι υψίστης σημασίας. Ταυτόχρονα, κυρίως λόγω της μεγάλης πίεσης για αξιοποίηση των επιστημονικών ευρημάτων αποτελεσμάτων στη κλινική πρακτική, έχει υποτιμηθεί ένας σημαντικός παράγοντας επιστημονικής προόδου: την **αναπααραγωγιμότητα** (reproducibility) των αποτελεσμάτων καθώς, τον “ανοιχτό” χαρακτήρα και τη “διαφάνεια” των μεθόδων που χρησιμοποιούνται για τη παραγωγή τους. Σε αυτό το προβληματικό και κατακερματισμένο ερευνητικό περιβάλλον καθίσταται αναγκαία η **σχεδίαση και ανάπτυξη υπολογιστικών υποδομών και υπηρεσιών** οι οποίες επιτρέπουν στους επιστήμονες μεταγονιδιωματικής και βιοπληροφορικής την απρόσκοπτη πρόσβαση σε κατανεμημένα βιο-δεδομένα, στο λογισμικό καθώς και στους αναγκαίους κατανεμημένους υπολογιστικούς πόρους επεξεργασίας δεδομένων, μέσω ολοκληρωμένων περιβαλλόντων σύνθεσης, εκτέλεσης και διαμοιρασμού βιοπληροφορικών ροών εργασίας. Επιπλέον, η ανάγκη για **συνεργασία και συνέργεια** μεταξύ των διαφορετικών κοινοτήτων βιοϊατρικής και μεταγονιδιωματικής έρευνας, απαιτεί την ανάπτυξη **συνεργατικών** υπηρεσιών, όχι μόνο κατά τις πειραματικές διαδικασίες αλλά και κατά τη διάχυση, τη διασπορά και την ερμηνεία των σχετικών επιστημονικών ευρημάτων. Τα προαναφερθέντα αποτελούν βασικές στοχεύσεις και αντικείμενα έρευνας και ανάπτυξης του έργου OpenBio-C. ➡ Στο παρόν παραδοτέο παρουσιάζουμε τις βασικές απαιτήσεις ενός ολοκληρωμένου διαδικτυακού περιβάλλοντος σύνθεσης και εκτέλεσης βιοπληροφορικών ροών εργασίας με βάση **εκτενή βιβλιογραφική έρευνα** και εκτεταμένη αναφορά και ανάλυση των εμπλεκόμενων **τεχνολογιών αιχμής**. Παρουσιάζονται επίσης τα βασικά συμπεράσματα από σχετικό ερωτηματολόγιο του OpenBio-C και τις απαιτήσεις οι οποίες κατεγράφησαν με βάση απαντήσεις από ερευνητές της περιοχής και βιοπληροφορικούς. Τέλος, παρουσιάζονται τα **χαρακτηριστικά και η λειτουργικότητα της αρχικής υλοποίησης βασικών συστατικών λογισμικού της υποδομής OpenBio-C**.

## ABSTRACT

Scientific research in post-genomic biomedicine and the analysis of a large volumes of heterogeneous bio-data is moving in a highly competitive environment where the speed, accuracy and quality of research results are of paramount importance. At the same time, mainly due to the great pressure for the utilization of scientific findings in clinical practice, an important factor of scientific progress has been underestimated: the **reproducibility** of results, the “openness” and the “*transparency*” of the methods used to produce them. In this problematic and fragmented research environment, it is necessary **to design and develop computational infrastructures and services that enable post-genomics and bioinformatics scientists to have unhindered access to distributed bio-data, software and distributed computational data processing resources through integrated environments synthesis, execution and sharing bioinformatics workflows**. In addition, the need for *collaboration* and *synergy* between the different communities of biomedical and post-genomic research requires the development of **collaborative** services not only in experimental processes but also in diffusion, dispersion and interpretation of relevant scientific findings. The above compose the key objectives of the research and development activities of the OpenBio-C project. ➡ The present deliverable presents the basic requirements of an integrated web environment for the synthesis and execution of bioinformatic workflows based on **extensive bibliographic research** and extensive reporting and analysis of the **state-of-the-art technologies** involved. The key findings from the OpenBio-C questionnaire, and the requirements that were recorded from the responses of post-genomics biomedicine researchers and bioinformaticians are also presented. Finally, the **features and functionality of the initial implementation of core software components of the OpenBio-C infrastructure** are presented.

## ΠΕΡΙΕΧΟΜΕΝΑ

|   |           |
|---|-----------|
| <b>1 ΕΙΣΑΓΩΓΗ.....</b>  | <b>1</b>  |
| 1.1 ΤΟ ΣΗΜΕΡΙΝΟ ΚΑΤΑΚΕΡΜΑΤΙΣΜΕΝΟ ΕΡΕΥΝΗΤΙΚΟ ΠΕΡΙΒΑΛΛΟΝ .....  | 2         |
| 1.2 Η ΚΡΙΣΗ ΑΝΑΠΑΡΑΓΩΓΙΜΟΤΗΤΑΣ ΕΠΙΣΤΗΜΟΝΙΚΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΚΑΙ ΕΥΡΗΜΑΤΩΝ .....  | 2         |
| <b>2 ΞΕΠΕΡΝΩΝΤΑΣ ΤΗ ΚΡΙΣΗ: ΑΝΑΓΚΕΣ ΚΑΙ ΟΙ ΠΡΟΫΠΟΘΕΣΕΙΣ .....</b>  | <b>2</b>  |
| 2.1 Η ΑΝΑΓΚΗ ΓΙΑ ΜΕΘΟΔΟΛΟΓΙΕΣ ‘ΑΝΟΙΧΤΗΣ ΕΠΙΣΤΗΜΗΣ’ .....  | 2         |
| 2.2 Η ΑΝΑΓΚΗ ΔΙΑΛΕΙΤΟΥΡΓΙΚΟΤΗΤΑΣ .....  | 3         |
| 2.3 ΥΠΟΣΤΗΡΙΞΗ ΣΥΝΕΡΓΑΣΙΑΣ & ΣΥΝΕΡΓΕΙΩΝ: Η ΑΝΑΓΚΗ ΣΥΝΕΡΓΑΤΙΚΩΝ ΠΕΡΙΒΑΛΛΟΝΤΩΝ .....  | 4         |
| 2.4 ΠΡΟΣ ΕΝΑ ΑΝΟΙΧΤΟ, ΔΙΑΛΕΙΤΟΥΡΓΙΚΟ ΚΑΙ ΣΥΝΕΡΓΑΤΙΚΟ ΠΕΡΙΒΑΛΛΟΝ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ.....  | 4         |
| <b>3 ΤΕΧΝΟΛΟΓΙΕΣ ΑΙΧΜΗΣ ΤΗΣ ΣΥΓΧΡΟΝΗΣ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ.....</b>  | <b>5</b>  |
| 3.1 ΠΕΡΙΒΑΛΛΟΝΤΑ ΒΙΟΠΛΗΡΟΦΟΡΙΚΩΝ ΕΠΙΣΤΗΜΟΝΙΚΩΝ ΡΟΩΝ ΕΡΓΑΣΙΑΣ .....  | 5         |
| 3.2 ΣΧΕΔΙΑΣΤΙΚΟΙ ΚΑΙ ΛΕΙΤΟΥΡΓΙΚΟΙ ΠΕΡΙΟΡΙΣΜΟΙ ΣΤΑ ΥΠΑΡΧΟΝΤΑ ΠΕΡΙΒΑΛΛΟΝΤΑ ΒΕΡΕ .....   | 7         |
| 3.3 ΠΕΡΙΒΑΛΛΟΝΤΑ ΚΑΙ ΕΡΓΑΛΕΙΑ ΥΠΟΣΤΗΡΙΞΗΣ ΤΗΣ ΣΥΝΕΡΓΑΣΙΑΣ .....   | 8         |
| 3.3.1 Εργαλεία νοητικής χαρτογράφησης .....   | 8         |
| 3.3.2 Εργαλεία υποστήριξης συνεργασίας με επιχειρήματα .....  | 12        |
| <b>4 ΣΧΕΔΙΑΖΟΝΤΑΣ ΕΥΕΛΙΚΤΑ ΚΑΙ ΑΝΟΙΧΤΑ ΠΕΡΙΒΑΛΛΟΝΤΑ ΕΠΙΣΤΗΜΟΝΙΚΩΝ ΡΟΩΝ ΕΡΓΑΣΙΑΣ: ΑΠΑΡΑΙΤΗΤΑ ΣΥΣΤΑΤΙΚΑ &amp; ΛΕΙΤΟΥΡΓΙΚΟΤΗΤΑ .....</b> | <b>16</b> |
| 4.1 ΥΠΟΣΤΗΡΙΞΗ ΕΙΚΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΠΕΡΙΒΑΛΛΟΝΤΩΝ (VIRTUALIZATION) .....   | 17        |
| 4.2 ΥΠΟΣΤΗΡΙΞΗ ΠΡΟΣΒΑΣΗΣ ΣΕ ΔΙΑΔΕΔΟΜΕΝΕΣ ΠΗΓΕΣ ΜΕΤΑΓΟΝΙΔΙΩΜΑΤΙΚΗΣ ... ΟΙ ΜΠΑΤΑΡΙΕΣ ΣΤΗ ΣΥΣΚΕΥΑΣΙΑ! .....                              | 17        |
| <b>5 ΑΠΑΙΤΗΣΕΙΣ ΤΗΣ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ ΚΟΙΝΟΤΗΤΑΣ .....</b>  | <b>18</b> |
| 5.1 ΟΙ ΑΝΑΓΚΕΣ ΤΩΝ ΚΟΙΝΟΤΗΤΩΝ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ: ΤΟ ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ ΤΟΥ OPENBIO-C.....   | 18        |
| 5.1.1 Ανάλυση απαντήσεων ερωτηματολογίου .....  | 21        |
| 5.1.2 Ανάλυση Απαιτήσεων: Γενικά συμπεράσματα από την ανάλυση του ερωτηματολογίου .....   | 27        |
| <b>6 ΠΡΩΤΟΤΥΠΗ ΑΡΧΙΚΗ ΥΛΟΠΟΙΗΣΗ ΤΟΥ OPENBIO-C.....</b>  | <b>27</b> |
| 6.1 ΤΟ ΒΑΣΙΚΟ ΠΛΑΙΣΙΟ ΚΑΙ ΥΠΟΒΑΘΡΟ ΤΗΣ ΥΛΟΠΟΙΗΣΗΣ .....   | 27        |
| 6.2 ΒΑΣΙΚΑ ΣΥΣΤΑΤΙΚΑ ΤΟΥ OPENBIO-C .....  | 28        |
| 6.3 ΟΙ ΒΑΣΙΚΕΣ ΜΟΝΑΔΕΣ ΥΛΟΠΟΙΗΣΗΣ ΤΟΥ OPENBIO-C.....  | 28        |
| 6.4 ΔΙΕΠΑΦΗ ΧΡΗΣΤΗ - OPENBIO-C .....  | 30        |
| 6.4.1 Εγγραφή & Προφίλ χρήστη.....  | 32        |
| 6.5 ΣΗΜΑΣΙΟΛΟΓΙΚΟΣ ΕΥΡΕΤΗΡΙΑΣΜΟΣ ΚΑΙ ΜΗΧΑΝΙΣΜΟΣ ΑΝΑΖΗΤΗΣΗΣ / ΕΝΤΟΠΙΣΜΟΥ ΕΡΓΑΛΕΙΩΝ.....  | 34        |
| 6.5.1 Λειτουργικότητα μηχανισμού αναζήτησης.....  | 35        |
| 6.5.2 Η “νοημοσύνη” του μηχανισμού αναζήτησης ... φέρε μου τα πιο σχετικά!.....   | 35        |
| 6.6 ΒΑΣΙΚΕΣ ΣΧΕΔΙΑΣΤΙΚΕΣ ΑΡΧΕΣ ΚΑΙ Η ΑΡΧΙΚΗ ΥΛΟΠΟΙΗΣΗ ΤΟΥ ΠΕΡΙΒΑΛΛΟΝΤΟΣ ΣΥΝΘΕΣΗΣ ΡΟΩΝ ΕΡΓΑΣΙΑΣ ΤΟΥ OPENBIO-C .....                    | 36        |
| 6.6.1 Εξαρτήσεις εργαλείων.....   | 37        |
| 6.6.2 Η δύναμη του BASH.....  | 37        |
| 6.6.3 Το σύστημα εκτέλεσης ροών εργασιών (Execution Environment) .....  | 38        |
| 6.6.4 Εισαγωγή εργαλείων στο OpenBio-C.....   | 39        |
| 6.7 ΤΟ ΣΥΝΕΡΓΑΤΙΚΟ ΠΕΡΙΒΑΛΛΟΝ ΤΟΥ OPENBIO-C.....  | 42        |
| 6.7.1 Σχεδιαστικές αρχές & λειτουργικότητα του συνεργατικού της αρχικής υλοποίησης του συνεργατικού περιβάλλοντος του OpenBio-C.....  | 42        |
| <b>7 ΥΠΟΛΟΓΙΣΤΙΚΗ ΥΠΟΔΟΜΗ ΤΟΥ OPENBIO-C: ΑΡΧΙΚΕΣ ΛΥΣΕΙΣ &amp; ΥΠΟΔΟΜΗ .....</b>   | <b>45</b> |
| <b>8 ΣΥΜΠΕΡΑΣΜΑΤΑ &amp; ΜΕΛΛΟΝΤΙΚΕΣ ΔΡΑΣΕΙΣ .....</b>   | <b>45</b> |
| <b>ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ .....</b>  | <b>48</b> |

## 1 ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια παρατηρούμε μια εντυπωσιακή έκρηξη στη παραγωγή επιστημονικών δεδομένων, όχι μόνο όσον αφορά το μεγάλο όγκο τους, αλλά και όσον αφορά την ετερογένεια και τη μορφή τους, καθώς και τις υπολογιστικές μεθοδολογίες και τα εργαλεία που χρησιμοποιούνται για την ανάλυση τους. Το φαινόμενο είναι ιδιαίτερα έντονο στον τομέα της σύγχρονης **βιοϊατρικής έρευνας**, κυρίως λόγω της ραγδαίας εξέλιξης της **γονιδιωματικής** έρευνας και των πραγματικά εντυπωσιακών επιτευγμάτων της βιοτεχνολογίας κυρίως όσον αφορά στις γονιδιωματικές πλατφόρμες *υψηλής απόδοσης* (highthroughput platforms) και τη δυνατότητα εκ 'παραλλήλου μέτρησης δεκάδων χιλιάδων έως και εκατομμυρίων μοριακών αποτυπωμάτων σε βιολογικούς ιστούς (γονιδίων, γενετικών μεταλλάξεων και παραλλαγών κλπ). Οστόσο, ο συντριπτικός ρυθμός αύξησης των γενετικών δεδομένων και η συνακόλουθη αύξηση των παραγόμενων νέων γνώσεων δημιουργούν ένα προβληματικό πεδίο βιοϊατρικής έρευνας στο οποίο, οι συμμετέχοντες ερευνητές συχνά δεν γνωρίζουν (πώς) και αδυνατούν να αποδώσουν νόημα (τι) στα παραγόμενα επιστημονικά ευρήματα και ανακαλύψεις, με αποτέλεσμα την πολύ περιορισμένη εκμετάλλευση των σχετικών επιστημονικών αποτελεσμάτων. Ενδεικτικά, έχει καταγραφεί ότι από τις 150.000 επιστημονικές δημοσιεύσεις με αναφορά στην ανακάλυψη νέων βιοδεικτών, μόνον 100 από αυτούς έχουν τύχη κάποιας κλινικής εφαρμογής (Poste, 2011). Σε αυτό το πλαίσιο, οι ερευνητές στο χώρο της **μεταγονιδιωματικής βιοϊατρικής** (post-genomics biomedicine) δυσκολεύονται να ακολουθήσουν τις πιο πρόσφορες κατευθυντήριες γραμμές και να λάβουν τεκμηριωμένες αποφάσεις για τις ερευνητικές τους προσπάθειες. Αυτό οφείλεται αφενός, στην **αδυναμία να αναλυθεί ο πραγματικά μεγάλος όγκος των σχετικών ετερογενών δεδομένων (γονιδιωματικών, γενετικών, κλινικών) και αφετέρου, στην περιορισμένη ικανότητα πιστοποίησης των παραγόμενων διεπιστημονικών ευρημάτων και γνώσεων.**

**Δεδομένα, δεδομένα παντού!** Η δυνατότητα παραγωγής **μεγάλου όγκου δεδομένων**, των αποκαλούμενων **‘big data’** (He et al., 2017) έχει καταστήσει πολλές παραδοσιακές αναλυτικές προσεγγίσεις αναποτελεσματικές. Ιδιαίτερα στο τομέα της μεταγονιδιωματικής και της σύγχρονης Βιοπληροφορικής παρουσιάζεται το πρωτοφανές φαινόμενο, ο ρυθμός παραγωγής δεδομένων να είναι υψηλότερος από τον ρυθμό ανάλυσης τους αλλά και αποτελεσματικής ερμηνείας τους ώστε να παραχθούν αξιόπιστα και ερμηνεύσιμα επιστημονικά αποτελέσματα (Marx, 2013). Αυτοί οι όγκοι επιστημονικών δεδομένων, από δεκάδες terabytes έως πολλά petabytes, είναι σχεδόν αδύνατο να ευρετηριαστούν, να αποθηκευτούν και να διαχειριστούν σε τοπικά περιβάλλοντα και βάσεις δεδομένων. Ταυτόχρονα, η έρευνα στο πεδίο των τεχνολογιών πληροφορικής έχει κάνει τεράστια πρόοδο όσον αφορά στην παροχή πραγματικά προχωρημένων υπολογιστικών περιβαλλόντων και λύσεων για την ανάλυση μεγάλου όγκου δεδομένων. Οι καταμεμημένες υπολογιστικές εφαρμογές (π.χ., MapReduce, Spark), οι βάσεις δεδομένων NoSQL (π.χ. DynamoDέχει B, BigQuery, MongoDB) και η μαζική αποθήκευση δεδομένων (π.χ., Amazon S3, Microsoft Azure, Google cloud services) αποτελούν μερικές από τις πολλά υποσχόμενες σχετικές τεχνολογίες. Ωστόσο, η εφαρμογή αυτών των τεχνολογιών σε επιστημονικά δεδομένα είναι πολύ περιορισμένη εξαιτίας της αδυναμίας προσβασιμότητας σε κάποιες από αυτές τις λύσεις, της πολυπλοκότητας της προσαρμογής τους, των απαιτούμενων δεξιοτήτων πληροφορικής και, ίσως το σημαντικότερο, του γεγονότος ότι οι λύσεις αυτές στοχεύουν και απευθύνονται περισσότερο στον ιδιωτικό επιχειρηματικό τομέα και λιγότερο τον ακαδημαϊκό χώρο.

Πιστεύουμε, κάτι που αποτελεί και ένα από τα βασικά κίνητρα και στόχους του έργου OpenBio-C, ότι η συνεχώς αυξανόμενες γνώσεις και επιτεύγματα που παράγονται από την διεπιστημονική μεταγονιδιωματική βιοϊατρική έρευνα μπορούν, αν οργανωθούν και διανεμηθούν κατάλληλα, να χρησιμοποιηθούν για την υλοποίηση περισσότερο εστιασμένων και καλά σχεδιασμένων κλινικο-γονιδιωματικών μελετών. Ο στόχος είναι να υποστηριχθούν διαδικασίες λήψης κλινικών αποφάσεων, οι οποίες να βασίζονται σε τεκμηριωμένα και επικυρωμένα διαγνωστικά, προγνωστικά και θεραπευτικά μοντέλα, συμβάλλοντας έτσι στις ερευνητικές προσπάθειες της μεταφραστικής βιοϊατρικής έρευνας<sup>1</sup>.

<sup>1</sup> <https://www.nature.com/subjects/translational-research>

## 1.1 Το σημερινό κατακερματισμένο ερευνητικό περιβάλλον

Το αποτέλεσμα είναι ο **κατακερματισμός της έρευνας**, κυρίως σε πεδία επιστημονικής έρευνας όπως η μεταγονιδιωμική βιοϊατρική, όπου οι προς διερεύνηση ερευνητικές υποθέσεις καθώς και ο προσδιορισμός της συνάφειας και πιθανής ολοκλήρωσης επιμέρους αποτελεσμάτων δεν διαθέτουν ανιχνεύσιμα και επαρκώς επικυρωμένα στοιχεία, καθιστώντας δύσκολη, και τις περισσότερες φορές ανεκπλήρωτη τη **μετάφραση** τους σε αξιόπιστες κλινικές συστάσεις και αποφάσεις. Το συγκεκριμένο περιβάλλον κατακερματισμένης έρευνας εκτείνεται σε δύο ορθογώνιες διαστάσεις: (i) στο **κατακερματισμό εντός του επιστημονικού πεδίου** όπου, η διερεύνηση της ίδιας επιστημονικής υπόθεσης γίνεται με διαφορετικές πειραματικές ρυθμίσεις και πρωτόκολλα (π.χ., χρησιμοποιώντας διαφορετικές πλατφόρμες ή μεθόδους ανάλυσης δεδομένων), και (ii) στο **κατακερματισμό μεταξύ επιστημονικών πεδίων** όπου, η ίδια ερευνητική υπόθεση διερευνάται σε διαφορετικά επίπεδα ή από διαφορετικές οπτικές γωνίες, π.χ., η συνέργεια πιθανών νευροεκφυλιστικών μηχανισμών μεταξύ μοριακών/γενετικών αποτυπωμάτων, από τη μια μεριά, και κλινικών / απεικονιστικών αποτυπωμάτων από την άλλη. Και στις δύο περιπτώσεις, τα αποτελέσματα της έρευνας και τα σχετικά ευρήματα, ακόμη και για την ίδια ερευνητική υπόθεση, εξακολουθούν να είναι δύσκολα ανιχνεύσιμα και ασύνδετα. Σε αυτό το κατακερματισμένο πλαίσιο, οι εμπλεκόμενες ερευνητικές βρίσκονται **“απομονωμένοι”** και σε αδυναμία να υποστηρίξουν όχι μόνο τα ευρήματά τους αλλά και το ίδιο το πρωτογενές ερευνητικό ερώτημα, δηλ., πώς να διαμορφώσουν μια “πειστική” προς διερεύνηση επιστημονική υπόθεση. Θεωρούμε ότι το φαινόμενο του κατακερματισμού της σύγχρονης έρευνας οφείλεται σε δύο κύριους λόγους: (α) στην απώλεια της εποπτείας και της δυνατότητας αφομοίωσης της παραγόμενης διεπιστημονικής γνώσης (μέσα από τις χιλιάδες των σχετικών δημοσιεύσεων και αναφορών), καθώς και (β) στην απουσία τεκμηρίωσης των διαδικασιών οι οποίες σχεδιάστηκαν και ακολουθήθηκαν για την παραγωγή αυτών των γνώσεων, κυρίως λόγω της περιορισμένης διαλειτουργικότητας των χρησιμοποιούμενων υπολογιστικών εργαλείων, υπηρεσιών και σχετικών επιστημονικών ροών εργασίας. Χωρίς υπερβολή θα λέγαμε ότι οι ερευνητές διεπιστημονικών πεδίων, όπως αυτός της μεταγονιδιωμικής βιοϊατρικής, μοιάζει να είναι **“χαμένοι στη μετάφραση”**! (Levin and Danesh-Meyer, 2010).

## 1.2 Η κρίση αναπαραγωγιμότητας επιστημονικών αποτελεσμάτων και ευρημάτων

Το σημερινό τοπίο της επιστημονικής έρευνας στη μεταγονιδιωμική είναι ένα άκρως ανταγωνιστικό περιβάλλον όπου η ταχύτητα, η ακρίβεια και η ποιότητα των ερευνητικών αποτελεσμάτων είναι υψίστης σημασίας. Απαιτεί από τους ερευνητές να επιτύχουν επιδόσεις υψηλής απήχησης και να παρέχουν αποτελεσματική και γρήγορη εφαρμογή των ευρημάτων τους σε αξιόπιστα κλινικά πρωτόκολλα. Αποτέλεσμα, να υπερεκτιμώνται οι επιπτώσεις (πιθανά) πρωτοποριακών αποτελεσμάτων και να υποεκτιμώνται ευρήματα μικρότερης εμβέλειας και σημασίας (Anderson *et al.*, 2007). Επιπλέον, η μεγάλη πίεση για δημοσίευση και μετάφραση αποτελεσμάτων έχει υποτιμήσει έναν σημαντικό παράγοντα επιστημονικής προόδου: την **αναπαραγωγιμότητα** των αποτελεσμάτων και τη διαφάνεια των μεθόδων που χρησιμοποιούνται για τη παραγωγή τους. Η αποκαλούμενη **“κρίση αναπαραγωγιμότητας”**<sup>2,3</sup> έχει χαρακτηριστεί ως “πανούκλα” για την επιστήμη που απειλεί τη χρηματοδότηση και στιγματίζει την ερευνητική κοινότητα<sup>4</sup> (Chen and Yang, 2009). Πολλές προσπάθειες για τη μέτρηση της αναπαραγωγιμότητας των δημοσιευμένων επιστημών αποκάλυψαν ότι το πρόβλημα είναι ευρέως διαδεδομένο και βαθιά ριζωμένο (Ioannidis, 2005; Munafò *et al.*, 2017).

## 2 ΞΕΠΕΡΝΩΝΤΑΣ ΤΗ ΚΡΙΣΗ: ΑΝΑΓΚΕΣ ΚΑΙ ΟΙ ΠΡΟΫΠΟΘΕΣΕΙΣ

### 2.1 Η ανάγκη για μεθοδολογίες ‘ανοιχτής επιστήμης’

Η δημοσίευση του πηγαίου κώδικα, των δεδομένων και άλλων στοιχείων υλοποίησης ενός ερευνητικού έργου εξυπηρετεί δύο βασικούς σκοπούς. Το πρώτο είναι να επιτρέψει στην κοινότητα να ελέγξει, να επικυρώσει και να επιβεβαιώσει την ορθότητα και την αξιοπιστία μιας

<sup>2</sup> Challenges in irreproducible research, <https://www.nature.com/collections/prbfkwmwvz/content/all-articles>

<sup>3</sup> Ένα ενδεικτικό παράδειγμα: [www.nature.com/news/error-found-in-study-of-first-ancient-african-genome-1.19258](http://www.nature.com/news/error-found-in-study-of-first-ancient-african-genome-1.19258)

<sup>4</sup> Retractions stigmatize scientific fields, study finds, <http://blogs.nature.com/news/2012/11/retractions-stigmatize-scientific-fields-study-finds.html>



ερευνητικής μεθοδολογίας. Το δεύτερο είναι να επιτρέψει στην κοινότητα να χρησιμοποιήσει σωστά τη μεθοδολογία σε νέα δεδομένα ή να την προσαρμόσει για να δοκιμάσει νέες ερευνητικές υποθέσεις. Πρόκειται για μια φυσική διαδικασία που αφορά σχεδόν κάθε νέα εφεύρεση ή εργαλείο η οποία μπορεί να πολύ εύκολα να εξελιχθεί μέσω της απλουστευμένης ακολουθίας: *δημιουργία εργαλείου => δοκιμή του εργαλείου => χρησιμοποίηση του εργαλείου*. Ωστόσο, είναι εκπληκτικό το γεγονός ότι στη βιοπληροφορική αυτή η φυσική ακολουθία δεν ήταν συνήθης πρακτική μέχρι τη δεκαετία του 1990, όταν ξεκίνησαν συγκεκριμένες ομάδες και σχετικές πρωτοβουλίες όπως το BOSC (Bioinformatics Open Source Conference) (Harris *et al.*, 2016). Πολλοί ερευνητές δηλώνουν ότι ιδανικά το τμήμα “Υλικά και Μέθοδοι” (materials and methods) μιας επιστημονικής δημοσίευσης πρέπει να είναι ένα ενιαίο αντικείμενο το οποίο ενσωματώνει την πλήρη μεθοδολογία, είναι άμεσα διαθέσιμο και εκτελέσιμο και συνοδεύει (αντί να συμπληρώνει) οποιαδήποτε βιοϊατρική έρευνα (Hettne *et al.*, 2012). Αυτό υπογραμμίζει την **ανάγκη για ένα σύνολο εργαλείων που αυτοματοποιούν την κοινή χρήση, την αναπαραγωγή και την επικύρωση αποτελεσμάτων και συμπερασμάτων από δημοσιευμένες μελέτες**.

Ήδη έχει διατυπωθεί η ανάγκη να συμπεριληφθούν εύκολα εργαλεία και δεδομένα (Goodman *et al.*, 2014) και να είναι δυνατή η δημιουργία πιο σύνθετων ή ολοκληρωμένων διαδικασιών έρευνας. Αυτή η νέα οικογένεια εργαλείων αναφέρεται ως **συστήματα διαχείρισης επιστημονικών ροών εργασίας / ΣΔΕΡΕ** (workflow management systems) (Karim *et al.*, 2017). Η εκτίμηση του αριθμού των εργαλείων που αποτελούν μέρος των αναπαραγωγικών ροών ανάλυσης δεν είναι μια τετριμμένη εργασία. Η πλατφόρμα Βιοπληροφορικής Galaxy<sup>5</sup> (Giardine *et al.*, 2005) διαθέτει μια διαθέσιμη στο κοινό αποθήκη διαθέσιμων εργαλείων όπου παρατίθενται 3356 διαφορετικά εργαλεία. Το myExperiment<sup>6</sup> (Goble *et al.*, 2010) είναι μια κοινωνική ιστοσελίδα όπου οι ερευνητές μπορούν να μοιράζονται τα ερευνητικά αντικείμενα όπως οι επιστημονικές ροές εργασίας. περιέχει περίπου 2700 ροές εργασίας. Για τη σύγκριση των βιολογικών εργαλείων (Ison *et al.*, 2016), μια κοινοτική προσπάθεια να καταγραφούν και να τεκμηριωθούν οι πόροι βιοπληροφορικής, παρατίθενται 6842 στοιχεία.<sup>2</sup> Η Bioinformatics.ca (Brazas *et al.*, 2010) ενσωματώνει μια λίστα με 1548 εργαλεία και 641 βάσεις δεδομένων. Η βάση OMICtools (Bandrowski *et al.*, 2014) περιέχουν περίπου 5000 εργαλεία και βάσεις δεδομένων και το περιοδικό Nucleic Acids Research επιμελείται μια λίστα με 1700 βάσεις δεδομένων (Galperin *et al.*, 2017). Τέλος, εκτιμάται ότι υπάρχουν περίπου 100 διαφορετικά ΣΔΕΡΕ με πολύ διαφορετικές αρχές σχεδιασμού.

## 2.2 Η ανάγκη διαλειτουργικότητας

Παρά το πλήθος των διαθέσιμων αποθετηρίων, εργαλείων, βάσεων δεδομένων και ΣΔΕΡΕ, ο συνδυασμός πολλαπλών εργαλείων σε μία ενιαία **επιστημονική ροή εργασίας** (ΕΡΕ) εξακολουθεί να θεωρείται μια πολύπλοκη διαδικασία που απαιτεί πάνω από τις μέσες δεξιότητες πληροφορικής (Leirzig, 2016). Το συγκεκριμένο έργο καθίσταται ακόμα πιο δύσκολο όταν ο στόχος είναι να συνδυαστούν πολλαπλές ΕΡΕ, να χρησιμοποιηθούν περισσότερα από ένα ΣΔΕΡΕ και τελικά, μία ολοκληρωμένη ΕΡΕ να μπορεί να εκτελεστεί σε ένα περιβάλλον υψηλής υπολογιστικής απόδοσης (high performance computing, HPC). Δεδομένου ότι η πρόοδος και η καινοτομία στη βιοπληροφορική βρίσκονται σε μεγάλο βαθμό στον σωστό συνδυασμό των υφισταμένων λύσεων, θα πρέπει να αναμένουμε σημαντική πρόοδο στην αυτοματοποίηση της κατασκευής ροών στο μέλλον (Parker, 2017). Εν τω μεταξύ, οι σημερινοί προγραμματιστές των εργαλείων και υπηρεσιών βιοπληροφορικής μπορούν να ακολουθήσουν ορισμένες κατευθυντήριες γραμμές ανάπτυξης λογισμικού που θα βοηθήσουν τους μελλοντικούς ερευνητές να κατασκευάσουν και να συνθέσουν πολύπλοκες ΕΡΕ με αυτά τα εργαλεία. Επιπλέον, οι πάροχοι δεδομένων μπορούν να ακολουθήσουν σαφείς οδηγίες προκειμένου να βελτιώσουν τη διαθεσιμότητα, τη δυνατότητα επαναχρησιμοποίησης και τη **σημαιολογική περιγραφή** αυτών των πόρων. Τέλος, οι μηχανικοί του ΣΔΕΡΕ πρέπει να ακολουθήσουν ορισμένες οδηγίες που μπορούν να αυξήσουν την συμπεριφορά, την εκφραστικότητα και την φιλικότητα προς το χρήστη αυτών των περιβαλλόντων.

<sup>5</sup> [galaxyproject.org](http://galaxyproject.org)

<sup>6</sup> [www.myexperiment.org/home](http://www.myexperiment.org/home)

## 2.3 Υποστήριξη συνεργασίας & συνεργειών: Η ανάγκη συνεργατικών περιβαλλόντων

Σήμερα, υπάρχουν πολλά πλούσια αποθετήρια εργαλείων ανοιχτού κώδικα. Επιπλέον, οι ενεργές διαδικτυακές επιστημονικές κοινότητες μπορούν να βοηθήσουν τους ερευνητές. Ωστόσο, τα υφιστάμενα διαδικτυακά επιστημονικά περιβάλλοντα δεν μπορούν να αντιμετωπίσουν με ικανοποιητικό τρόπο κάποια κοινά ερωτήματα τα οποία επανειλημμένα εμπλέκονται στις διαδικασίες ανάλυσης βιοϊατρικών μεταγονιδιωματικών δεδομένων, όπως για παράδειγμα:

*“έχω δεδομένα που περιέχουν έναν αριθμό γονιδίων ή γενετικών παραλλαγών, ποια εργαλεία υπάρχουν για την ανάλυση τους;”*

*“ποιες είναι οι υπολογιστικές απαιτήσεις; ποια από αυτά χρησιμοποιούνται πιο συχνά; που μπορώ να βρω παραδείγματα εισόδων και εξόδων αυτών των εργαλείων; ποιες δημοσιεύσεις έχουν χρησιμοποιήσει αυτά τα εργαλεία;”*

*“εάν εντοπίσω ένα λάθος στην τεκμηρίωση ή το λογισμικό ή αν βρήκα μια άτυπη χρήση ενός εργαλείου, πώς μπορώ να ενημερώσω την κοινότητα;”*

*“πώς μπορώ να αξιολογήσω ένα εργαλείο το οποίο θεωρώ εξειδικευμένο;”*

*“πώς μπορώ να προσεγγίσω εμπειρογνώμονες σχετικά με μια συγκεκριμένη ανάλυση; για ποιους σκοπούς και με ποιο τρόπο μια κοινότητα χρησιμοποιεί τις περισσότερες φορές ένα εργαλείο;”*

*“από πού μπορώ να λάβω πληροφορίες σχετικά με την ερευνητική μου υπόθεση και την προγραμματισμένη ανάλυση μου;”*

*“πώς μπορώ να συλλέξω πρόσθετα δεδομένα που θα ενισχύσουν τα στοιχεία των ανακαλύψεών μου;”*

*“από πού μπορώ να κάνω μια δεύτερη γνώμη σχετικά με ένα συγκεκριμένο αποτέλεσμα;”*

Πρόκειται για ένα μικρό σύνολο συχνών ερωτημάτων και ζητημάτων τι οποίες ένας ερευνητής στη περιοχή της ανάλυσης βιο-δεδομένων προσπαθεί να αντιμετωπίσει, συνήθως με μηχανές αναζήτησης και σχετικά επιστημονικά φόρουμ. Είναι αλήθεια ότι η περιορισμένη πληροφόρηση, η προσωπική προκατάληψη των απόψεων και η έλλειψη εμπειρογνομώνων δημιουργούν έναν κατακερματισμένο χώρο έρευνας. Έτσι, παρόλο που υπάρχει πληθώρα διαδικτυακών κοινοτήτων βιοπληροφορικής και χώρων αποθήκευσης επαρκώς περιεγραμμένων και τεκμηριωμένων αναλυτικών εργαλείων, η έλλειψη μηχανισμών και υπηρεσιών εύκολης ενσωμάτωσης τους σε ενιαία περιβάλλοντα σύνθεσης και εκτέλεσης τεκμηριωμένων και διάφανων επιστημονικών ροών εργασίας, αλλά και η ανυπαρξία των αναγκαίων περιβαλλόντων υποστήριξης και ενίσχυσης της **συνεργασίας** μεταξύ των εμπλεκόμενων ομάδων και χρηστών, συνθέτουν το παζλ του προβληματικού ερευνητικού περιβάλλοντος. Με άλλα λόγια, φαίνεται ότι λείπουν και δεν υποστηρίζονται μεθοδολογίες **ανοιχτής επιστήμης** (*open science*<sup>7</sup>). (Karacapilidis and Potamias, 2016)

Υπογραμμίζεται έτσι η **ανάγκη για συνεργασία και συνέργεια** μεταξύ των διαφορετικών ερευνητικών κοινοτήτων της βιοϊατρικής και μεταγονιδιωματικής έρευνας η οποία απαιτεί την ανάπτυξη **υποστήριξης συνεργατικών περιβαλλόντων**, όχι μόνο κατά τις πειραματικές διαδικασίες αλλά και κατά τη διάχυση, τη διασπορά και την ερμηνεία των σχετικών επιστημονικών ευρημάτων (Karacapilidis, 2014). Τα συνεργατικά ερευνητικά περιβάλλοντα απαιτούν φιλικές προς το χρήστη λύσεις οι οποίες από τη μία πλευρά, επιτρέπουν στους εμπλεκόμενους ερευνητές να χειρίζονται εύκολα μεγάλους όγκους δεδομένων και από την άλλη, τους παρέχουν κατευθύνσεις και συστάσεις στις οποίες μπορούν να στηρίξουν τις επιστημονικές αποφάσεις τους (Carusi and Reimer, 2010; Karacapilidis *et al.*, 2012).

## 2.4 Προς ένα ανοιχτό, διαλειτουργικό και συνεργατικό περιβάλλον βιοπληροφορικής

Η κύρια πρόκληση του σύγχρονου ερευνητικού οικοσυστήματος είναι να επιτρέψει τη βέλτιστη χρήση ερευνητικών δεδομένων και αναλυτικών υπολογιστικών μεθόδων. Το κίνητρο μας για τη **σχεδίαση, ανάπτυξη και εκμετάλλευση ενός ενιαίου, ανοιχτού και συνεργατικού περιβάλλοντος υποστήριξης βιοπληροφορικών ροών εργασίας** είναι η γεφύρωση των παρακάτω **χασμάτων** τα οποία χαρακτηρίζουν και διέπουν το οικοσύστημα της σύγχρονης έρευνας στο πεδίο της μεταγονιδιωματικής βιοϊατρικής και ανάλυσης σχετικών βιο-δεδομένων:

<sup>7</sup> <https://osf.io/>



- ❖ το **χάσμα δεδομένων**, το οποίο οφείλεται στην ύπαρξη πολλών μη συνδεδεμένων και διαφορετικών πηγών δεδομένων -- για να διορθωθεί αυτό, η προσέγγισή μας θα πρέπει να παρέχει: λεπτομερή περιγραφή δεδομένων συμπεριλαμβανομένης της προέλευσης τους (provenance) δεδομένων, διασύνδεση δεδομένων με εργαλεία και πλατφόρμες επεξεργασίας, έτσι ώστε να καθίσταται ανιχνεύσιμα τα κατάλληλα υπολογιστικά περιβάλλοντα και τα πιο πρόσφορα εργαλεία ανάλυσης;
- ❖ το **σημασιολογικό χάσμα**, το οποίο κυρίως οφείλεται στην ύπαρξη πολλαπλών λεξιλογίων (vocabularies) και οντολογιών για τη περιγραφή δεδομένων και μεθοδολογιών ανάλυσης τους -- για να διορθωθεί αυτό οι χρήστες πρέπει να μπορούν να συμπεριλάβουν υπάρχουσες και καθιερωμένες οντολογίες για την επισήμειωση (annotation) δεδομένων και εργαλείων;
- ❖ το **γνωσιακό χάσμα**, το οποίο οφείλεται στην ύπαρξη ετερογενών, ταχέως αναπτυσσόμενων και κατακερματισμένων πόρων γνώσης -- για να αντιμετωπιστεί αυτό, τα δεδομένα, τα εργαλεία και οι επιστημονικές ροές εργασίας πρέπει συνδέονται με σχετικές δημοσιεύσεις και αναφορές;
- ❖ το **συνεργατικό χάσμα**, το οποίο οφείλεται στην ύπαρξη διαφορών στα ζητούμενα των χρηστών, τους στόχους και τις αναλυτικές μεθοδολογίες που ακολουθούν -- για να διορθωθεί αυτό πρέπει: (i) να επιτρέπεται στους χρήστες να χτίσουν προσωπικά προφίλ όπου να δηλώνουν την εμπειρογνομosύνη τους, τις δημοσιεύσεις και τις προτιμήσεις στα προς χρήση εργαλεία ανάλυσης, οικοδομώντας έτσι το προφίλ αλλά και τη “φήμη” τους σύμφωνα με τις συνεισφορές τους σε σχετικές υπηρεσίες στην ερευνητική κοινότητα και (ii) να μπορούν να εκφράζουν τις απόψεις, αξιολογήσεις και επιχειρηματολογία τους για κάθε ερευνητικό αντικείμενο το οποίο εμπλέκεται σε μια επιστημονική ροή εργασίας (δεδομένα, μεθοδολογίες και εργαλεία ανάλυσης, δημοσιευμένα αποτελέσματα καθώς και εμπλεκόμενες ερευνητικές ομάδες και ανθρώπους)

Σύμφωνα με τα παραπάνω, το έργο OpenBio-C αναπτύσσει ένα **περιβάλλον και αντίστοιχη διαδικτυακή πλατφόρμαση** οποία δίνει τη δυνατότητα στις ερευνητικές κοινότητες βιοϊατρικής και μετα-γονιδιωματικής να **συνεργαστούν αποτελεσματικά** χάρη στην **αξιόπιστη και φιλική προς τον χρήστη πρόσβαση σε ολοκληρωμένες και διαλειτουργικές πηγές διαφορετικών τύπων** (δεδομένων, εργαλείων, υπολογιστικών πόρων και ανθρώπων/ομάδων). Το περιβάλλον προβλέπεται να ενσωματώνει και διαχειρίζεται ετερογενείς πηγές δεδομένων και γνώσης, καθώς και αντίστοιχες μεθοδολογίες και εργαλεία επεξεργασίας, μέσω ενός περιβάλλοντος επεξεργασίας και εκτέλεσης επιστημονικών εργασιών ροών εργασίας. Επιπλέον, το OpenBio-C πρέπει να παρέχει στους εμπλεκόμενους ερευνητές/χρήστες την απαιτούμενη ικανότητα να συνθέτουν και να αναπτύσσουν εύκολα το δικό τους περιβάλλον εργασίας χρησιμοποιώντας τα πιο πρόσφορα εργαλεία και μεθοδολογίες ανάλυσης βιο-δεδομένων.

### 3 ΤΕΧΝΟΛΟΓΙΕΣ ΑΙΧΜΗΣ ΤΗΣ ΣΥΓΧΡΟΝΗΣ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ

#### 3.1 Περιβάλλοντα Βιοπληροφορικών Επιστημονικών Ροών Εργασίας

Υπάρχει πλήθος εκτεταμένων και σχολαστικών ανασκοπήσεων υπαρχόντων περιβαλλόντων **Βιοπληροφορικών Επιστημονικών Ροών Εργασίας (BEPE)** οι οποίες μπορούν να αναζητηθούν σε πολλές σχετικές μελέτες και δημοσιεύσεις (Leipzig, 2016; Karim *et al.*, 2017; Spruth *et al.*, 2016; Kulkarni *et al.*, 2018). Παρόλα αυτά, αξίζει να αναφερθούμε σε μερικά από τα πιο γνωστά και πιο χρησιμοποιούμενα περιβάλλοντα και προσεγγίσεις, κάτι το οποίο θα μας βοηθήσει να εντοπίσουμε πιθανά κενά και προβλήματα και έτσι να οδηγηθούμε στο προσδιορισμό των απαιτήσεων ενός σύγχρονου, ολοκληρωμένου και ευέλικτου περιβάλλοντος σύνθεσης και εκτέλεσης BEPE.

- **Galaxy.** Το πιο σημαντικό, και ίσως πιο επιτυχημένο έως σήμερα περιβάλλον BEPE, με περισσότερο από μια δεκαετία συνεχούς ανάπτυξης και χρήσης, είναι το Galaxy (Giardine *et al.*, 2005). Το Galaxy κατάφερε να δημιουργήσει μια πολύ ενεργή κοινότητα βιολόγων και ειδικών πληροφορικής με έναν αρκετά μεγάλο αριθμό κατανεμημένων εγκαταστάσεων σχετικών εξυπηρετητών (servers) σε πολλά ερευνητικά ιδρύματα παγκοσμίως. Τα βασικά χαρακτηριστικά τα οποία συμβάλουν στην επιτυχία του Galaxy είναι: (α) η δυνατότητα δημιουργίας ανεξάρτητων αποθετηρίων ροών εργασίας και εργαλείων και η δυνατότητα συμπερίληψής τους σε διαφορετικούς εξυπηρετητές (servers); (β) η ύπαρξη ενός πολύ βασικού και απλού μηχανισμού ενσωμάτωσης διαφορετικών υπολογιστικών εργαλείων (Krieger *et al.*, 2017); και (γ) η προσφορά ενός απλού διαδραστικού διαδικτυακού περιβάλλοντος σύνθεσης ροών

εργασίας βασισμένο σε γραφήματα. Παρά την επιτυχία του, το Galaxy εξακολουθεί να στερείται αρκετών ποιοτικών κριτηρίων, με τα πιο σημαντικά να είναι: η δυνατότητα να μπορούν οι χρήστες να συνεργάζονται κατά τη διάρκεια σύνθεσης ροών, να τις ανταλλάσσουν και να τις διαμοιράζονται (*sharing*), καθώς και τη δυνατότητα να αξιολογούν και να βαθμολογούν τα αποτελέσματα (*rating*), ανταλλάσσοντας ιδέες, προτάσεις και εναλλακτικές προσεγγίσεις.

- **Taverna.** Το δεύτερο πιο γνωστό περιβάλλον BEPE είναι το Taverna (Wolstencroft *et al.*, 2013). Το Taverna έχει, τουλάχιστον όσον αφορά στο Galaxy, περιορισμένη χρήση στον τομέα της μεταγονιδιωματικής έρευνας. Ο λόγος ότι στηρίζεται σε ένα μη-διαδικτυακό και σχετικά πιο περίπλοκο πλαίσιο σύνθεσης ροών εργασίας, υποχρεώνοντας τους χρήστες να τηρούν οι ίδιοι βάσεις με συγκεκριμένα πρότυπα ροών εργασίας, όπως για παράδειγμα το BioMoby<sup>8</sup>. Το Taverna υπολείπεται στη χρήση του από τις κοινότητες βιοπληροφορικής σε σχέση με το Galaxy, αν και όλα τα σχετικά BEPE τα οποία έχουν αναπτυχθεί με το Taverna περιγράφονται και διατίθενται στο γνωστό αποθετήριο myExperiment, βλέπε τις σχετικές τάσεις στο [imgur.com/a/7Zzlu](https://imgur.com/a/7Zzlu)<sup>9</sup>. Αυτό οφείλεται κυρίως στο γεγονός ότι ο χρόνος από τον εντοπισμό μιας χρήσιμης BEPE έως τη πραγματική εκτέλεση της μπορεί να πάρει πολλές μέρες ανάλογα με τις ικανότητες των χρηστών όσον αφορά στη κατανόηση και ανάπτυξη του αναγκαίου υπολογιστικού κώδικα και την προθυμία του να κατανοήσει τις ιδιαιτερότητες του προσφερόμενου πολύπλοκου περιβάλλοντος. Το συγκεκριμένο πρόβλημα θα πρέπει να μελετηθεί πιο βαθιά ώστε να εξαχθούν πολύτιμα συμπεράσματα σε σχέση με τη σχεδίαση και ανάπτυξη σύγχρονων και ευέλικτων περιβαλλόντων BEPE.
- **Molgenis.** Το περιβάλλον σύνθεσης και εκτέλεσης BEPE Molgenis προσφέρει δυνατότητες ενσωμάτωσης εργαλείων και πιο απλών, σε σχέση με το Galaxy και Taverna, περιγραφών και λειτουργιών σύνθεσης BEPE (Swertz *et al.*, 2010). Το Molgenis έχει ενσωματωμένα πολλά γνωστά και καθιερωμένα εργαλεία γονιδιωματικών/γενετικών δεδομένων με κύρια αναφορά σε υπολογιστικές τεχνικές συσχέτισης φαινοτύπου-γονοτύπου (phenotype-to-genotype association analysis) και ανάλυσης δεδομένων αλληλούχισης νέας γενιάς (NGS, next generation sequencing) (Byelas *et al.*, 2012). Το Molgenis προσφέρει επίσης λειτουργίες εξαγωγής και εκτέλεσης των συντεθειμένων BEPE σε καταμετρημένα υπολογιστικά περιβάλλοντα, π.χ., clusters/Grids (Byelas *et al.*, 2011, 2012).
- **Προγραμματικές Δέσμες.** Εκτός από ολοκληρωμένα περιβάλλοντα σύνθεσης και εκτέλεσης BEPE θα πρέπει να αναφερθούμε και σε λύσεις οι οποίες στηρίζονται σε γλώσσες προγραμματισμού και παρέχουν τη δυνατότητα σύνθεσης ροών μέσω σχετικών **προγραμματικών δεσμών εντολών** (programmatic scripts). Τέτοιες λύσεις παρέχονται από σχετικά πακέτα Python όπως το Luigi και το bcbio-nextgen (Guimera, 2012), και πακέτα Java όπως το bripe (Sadedin *et al.*, 2012). Άλλες και αρκετά ευέλικτες λύσεις είναι τα: Snakemake<sup>10</sup>, Nextflow<sup>11</sup> και BigDataScript<sup>12</sup> (Cingolani *et al.*, 2015), τα οποία παρέχουν υπολογιστικά περιβάλλοντα (στη πραγματικότητα νέες γλώσσες προγραμματισμού) αφιερωμένα και εστιασμένα στη σύνθεση BEPE οι οποίες υλοποιούν αποκλειστικά pipelines βιοπληροφορικής<sup>13</sup>. Η περιγραφή μιας ροής εργασίας σε αυτά τα πακέτα δίνει στον ερευνητή τη δυνατότητα να εκτελούν τις αναλύσεις τους εύκολα σε πληθώρα περιβαλλόντων, αλλά δυστυχώς οι παρεχόμενες υπηρεσίες στοχεύουν ειδικευμένους χρήστες με εκτεταμένες δυνατότητες προγραμματισμού. Τέλος, υπολογιστικές γλώσσες για τη περιγραφή και την ανταλλαγή BEPE είναι οι YAWL (van der Aalst and ter Hofstede, 2005), η CWL – common workflow language<sup>14</sup> και η WDL – open workflow description language<sup>15</sup>. Η υποστήριξη μιας ή περισσότερων γλωσσών για τη περιγραφή μια BEPE αποτελούν ένα από τα σημαντικότερα ποιοτικά κριτήρια ενός σύγχρονου και ευέλικτου περιβάλλοντος σύνθεσης, διαμοιρασμού και εκτέλεσης BEPE.
- Άλλα ολοκληρωμένα περιβάλλοντα BEPE με πολλά υποσχόμενα χαρακτηριστικά περιλαμβάνουν το Omics Pipe (Fisch *et al.*, 2015), το EDGE (Li *et al.*, 2017) για δεδομένα NGS και το Chipster (Kallio *et al.*, 2011) για ανάλυση δεδομένων μικροσυστοιχιών (microarrays). Όλα αυτά τα περιβάλλοντα υποστηρίζουν ότι είναι αποκεντρωμένα και βασίζονται στις συνεισφορές των εμπλεκόμενων ερευνητικών κοινοτήτων, παρόλο που η ευρεία υιοθέτησή τους από αυτές είναι ακόμη περιορισμένη.

<sup>8</sup> [biomoby.open-bio.org](https://biomoby.open-bio.org)

<sup>9</sup> Έχει παραχθεί μέσω του [gist.github.com/kantale/01e5b42289c37786652ceadce57c07dd](https://gist.github.com/kantale/01e5b42289c37786652ceadce57c07dd)

<sup>10</sup> [bitbucket.org/snakemake/snakemake/wiki/Home](https://bitbucket.org/snakemake/snakemake/wiki/Home)

<sup>11</sup> [www.nextflow.io](https://www.nextflow.io)

<sup>12</sup> [pcingola.github.io/BigDataScript](https://pcingola.github.io/BigDataScript)

<sup>13</sup> [www.biostars.org/p/91301](https://www.biostars.org/p/91301)

<sup>14</sup> [www.commonwl.org](https://www.commonwl.org)

<sup>15</sup> [www.openwdl.org](https://www.openwdl.org)

### 3.2 Σχεδιαστικοί και λειτουργικοί περιορισμοί στα υπάρχοντα περιβάλλοντα BEPE

Σήμερα υπάρχει μια πληθώρα από περιβάλλοντα σύνθεσης και εκτέλεσης BEPE. Μια σημαντική και αρκετά ενδεικτική παρατήρηση σχετικά με τη χρήση τους συνοψίζεται στο εξής: "όσο πιο δύσκολο είναι για έναν ερευνητή να χρησιμοποιήσει ένα σύστημα σε σύγκριση με ένα *ad-hoc script* ή ίσως με την εκμετάλλευση μη βέλτιστων, αυτόνομων και μη διαλειτουργικά συνδεδεμένων εργαλείων, τόσο χαμηλότερη είναι η ευρεία αποδοχή της ροής εργασίας από τις κοινότητες βιοπληροφορικής και υπολογιστικής βιολογίας" (Sprjuth *et al.*, 2015). Επιπλέον, οι επιστημονικές συνεργασίες γίνονται ολοένα και πιο παγκόσμιες, πολυ-πολικές (multi-site) και διαδικτυωμένες, κάτι που απαιτεί **καινοτόμα, δυναμικά και ανά-πάσα στιγμή διαθέσιμα ερευνητικά περιβάλλοντα** όπου οι διασκορπισμένοι επιστήμονες μπορούν να έχουν απρόσκοπτη πρόσβαση και στα εμπλεκόμενα δεδομένα και στο λογισμικό καθώς και στους αναγκαίους καταναμημένους υπολογιστικούς πόρους επεξεργασίας, μέσω απλών διαδικτυακών εφαρμογών (Smith and Sotola, 2011).

Επιδιώκοντας τη διαμόρφωση μιας πιο εμπειριστατωμένης άποψης σύγχρονων BEPE, μπορούμε να εντοπίσουμε τρία σημαντικά μειονεκτήματα στα διαθέσιμα περιβάλλοντα:

- **Κλειστή φύση.** Ίσως το πιο σημαντικό χαρακτηριστικό για τη διάδοση σύγχρονων περιβαλλόντων BEPE, είναι η παροχή ανοικτού κώδικα (φυσικά στα πλαίσια συγκεκριμένων πλαισίων άδειας για τη χρήση του), κάτι που αποτελεί ένα από τα σημαντικότερα επιτεύγματα της σύγχρονης επιστημονικής κοινότητας εν γένει. Οι μεθοδολογίες της **ανοιχτής επιστήμης**<sup>16</sup> (open science) εκτός από τα ανοικτά εργαλεία και τις ροές εργασίας, υποστηρίζουν ανοικτές υπηρεσίες, ανοικτά αποθετήρια, ανοιχτές πολιτικές δεδομένων και ανοιχτές οντολογίες περιγραφής τους, ανοιχτές μετρήσεις καθώς και ανοιχτά περιβάλλοντα σύνθεσης και εκτέλεσης επιστημονικών ροών εργασίας. Αλλά "ανοιχτό" δεν ισοδυναμεί με το ότι κάτι "δεν είναι κλειστό"! αλλά κάτι ευρύτερο. Σημαίνει ότι ένας χρήστης μπορεί εύκολα και χωρίς περιορισμούς να χρησιμοποιεί, να δημιουργεί, να ανακαλύπτει, να επεξεργάζεται, να βαθμολογεί, να σχολιάζει, να μοιράζεται και να αναφέρεται σε όλα τα επιμέρους συστατικά μιας ροής εργασίας (Badia *et al.*, 2017).
- **Σύνδρομο 'κλειδώματος'.** Τα σύγχρονα περιβάλλοντα BEPE επιτρέπουν την κατασκευή ροών ανάλυσης σύμφωνα με τις αντίστοιχες εσωτερικές τους ιδιαιτερότητες. Για παράδειγμα στο Galaxy, ο χρήστης πρέπει να μετατρέψει τα διαθέσιμα, εξωτερικά ως προς το Galaxy, εργαλεία σε εργαλεία Galaxy, να μετατρέψει τα βήματα εκτέλεσης σε βήματα Galaxy, να ορίσει τις προγραμματιστικές παραμέτρους του Galaxy, και στη συνέχεια να συνθέσει ροές εργασίας που μπορούν να εκτελεστούν αποκλειστικά σε περιβάλλον εκτέλεσης Galaxy! Αλλά πάντα θα υπάρχει το ερώτημα: *αξίζει τον κόπο*; Η ανάγνωση των σελίδων τεχνικής τεκμηρίωσης του Galaxy, που πολλές φορές υπερβαίνουν την τεχνογνωσία του χρήστη είναι μια σημαντική επένδυση χρόνου. Ίσως πιο σημαντικό, η συνεργασία μεταξύ διαφορετικών περιβαλλόντων είναι πρακτικά αδύνατη. Οι προσπάθειες συνδυασμού πολλαπλών ροών εργασίας από διαφορετικά περιβάλλοντα και η ενσωμάτωση ενός περιβάλλοντος σε ένα άλλο ή η ενσωμάτωση του σε εξωτερικά εργαλεία είναι τεχνικά πολύ δύσκολο να διεκπεραιωθεί. Είναι ενδεικτικό ότι η συνέργεια μεταξύ Galaxy και Taverna (Wolstencroft *et al.*, 2013) είναι τόσο δύσκολη ώστε έπρεπε να δημιουργηθεί ένα νέο υβριδικό περιβάλλον, το TavernaX! (Abouelhoda *et al.*, 2012). Ακόμη και στις λύσεις οι οποίες προσφέρουν πλούσιες περιγραφές των συστατικών μιας επιστημονικής ροής εργασίας, των αποκαλούμενων **ερευνητικών αντικειμένων** (research objects) (Hettne *et al.*, 2014), λείπει ένα ζωτικό στοιχείο: η "χρήση τώρα". Εν ολίγοις, τα σύγχρονα περιβάλλοντα BEPE απαιτούν εκτεταμένες επενδύσεις χρόνου, απαιτούν την αναμόρφωση των αρχικών πλάνων του χρήστη σύμφωνα με τις λεπτομέρειες του περιβάλλοντος, και θα λέγαμε ότι είναι γενικά "εχθρικές" ως προς την ενσωμάτωση εξωτερικών (στο συγκεκριμένο προς χρήση περιβάλλον) υλοποιήσεων.
- Ο αγνοούμενος **κοινωνικός παράγοντας**. Καθώς η έρευνα γίνεται πιο διεπιστημονική, αναμένουμε ότι ο αριθμός των συγγραφέων σε δημοσιευμένα έγγραφα θα αυξηθεί. Το 2016, δημοσιεύθηκαν επιστημονικές εργασίες ακόμη και με περισσότερους από 1000 συντάκτες! (Leung *et al.*, 2015). Προφανώς, όλοι αυτοί οι συγγραφείς συμμετείχαν σε μια πολύ δημιουργική συζήτηση σχετικά με διάφορες πτυχές της έρευνας, οι οποίες έλαβαν χώρα πριν από τη δημοσίευση. Η πλήρης **επιχειρηματολογία** από πολλούς ανθρώπους με διαφορετική εμπειρία όσον αφορά τα εργαλεία, τις

<sup>16</sup> [www.fosteropenscience.eu/content/what-open-science-introduction](http://www.fosteropenscience.eu/content/what-open-science-introduction)

μεθόδους ανάλυσης και τις συγκεκριμένες τεχνικές που οδήγησαν σε μια δημοσίευση υψηλού προφίλ είναι ένα “επιστημονικό χρυσωρυχείο” το οποίο δε συνοδεύει τη δημοσίευση. Συνήθως οι ΒΕΡΕ ανάλυσης βιο-δεδομένων σπάνια συνοδεύουν ένα δημοσιευμένο έγγραφο. Τα σύγχρονα περιβάλλοντα ΒΕΡΕ αγνοούν εντελώς αυτό το κρίσιμο ζήτημα. Αντιμετωπίζοντας ένα τέτοιο περιοριστικό πλαίσιο, οι ερευνητές πρέπει να επιλέξουν μεταξύ δύο μη ικανοποιητικών επιλογών: αφενός, να περιορίσουν την επιστήμη τους σύμφωνα με τις ιδιαιτερότητες ενός περιβάλλοντος ΒΕΡΕ, προκειμένου να δημοσιεύσουν και να εξελίσουν τη καριέρα τους (*publish or perish!*) και αφετέρου, να θυσιάσουν την **αναπαραγωγικότητα** της εργασίας και των αποτελεσμάτων τους, επειδή στη πραγματικότητα δε στηρίζονται και δε χρησιμοποιούν ένα δομημένο περιβάλλον για τις διαδικασίες επεξεργασίας και ανάλυσης που ακολούθησαν. Καταλήγουμε έτσι στην εξής σημαντική διαπίστωση: **οι σύγχρονες τεχνολογίες ΒΕΡΕ πρέπει να επικεντρώνονται σε στρατηγικές που μειώνουν την “αυτοματοποίηση των πάντων” και στη βελτίωση της ποιότητας των παρεχόμενων λύσεων, εμπλέκοντας τους κατάλληλους ανθρώπους την κατάλληλη στιγμή.** Με άλλα λόγια θα πρέπει να συστηματοποιούν, να υποστηρίζουν και να ενισχύουν προσεγγίσεις “συνεισφοράς πλήθους” (crowdsourcing) στη σχεδίαση, στη σύνθεση, στη συντήρηση και διάδοση επιστημονικών ροών εργασίας (Shah et al., 2016).

### 3.3 Περιβάλλοντα και εργαλεία υποστήριξης της συνεργασίας

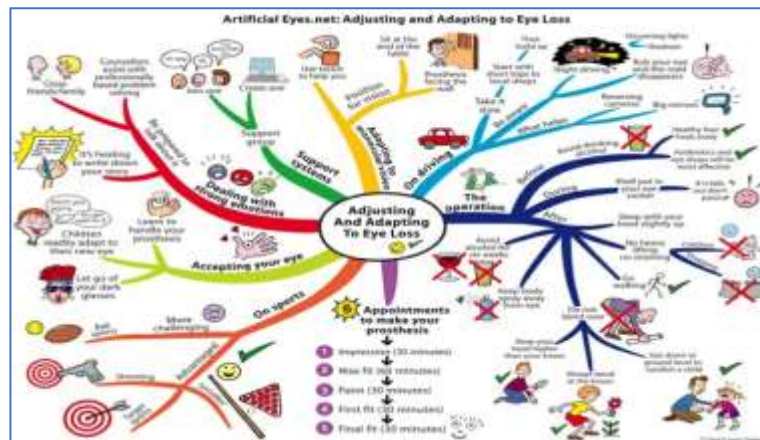
Ο όρος **λογισμικό υποστήριξης της συνεργασίας** (collaboration support software) αναφέρεται σε λογισμικό που έχει σχεδιαστεί για να υποστηρίξει μια ομάδα ανθρώπων προκειμένου να επιτύχουν τους στόχους τους στο πλαίσιο της εμπλοκής τους σε μια κοινή εργασία<sup>17</sup>. Η έλευση της εποχής του Web 2.0 εισήγαγε μια πληθώρα εργαλείων υποστήριξης της συνεργασίας που εξασφαλίζουν την αλληλεπίδραση των χρηστών σε μαζική κλίμακα. Τα εργαλεία αυτά καλύπτουν ένα ευρύ φάσμα αναγκών που περιλαμβάνουν την απλή ανταλλαγή γνώσεων, την ανταλλαγή και την τοποθέτηση ετικετών (tags), την κοινωνική δικτύωση, την ομαδική συγγραφή, τη νοητική χαρτογράφηση (mind mapping) και τη συζήτηση. Στις επόμενες ενότητες, εξετάζονται διεξοδικά δυο βασικές κατηγορίες εργαλείων Web 2.0 (Χριστοδούλου, 2015). Έχοντας ως βασικό στόχο να σχεδιαστεί ένα εργαλείο που θα υποστηρίξει αποδοτικά τη συνεργασία και τη λήψη αποφάσεων σε περιβάλλοντα που χαρακτηρίζονται από πληθώρα χρηστών και αυξημένο (όπως και πολύπλοκο) όγκο πληροφοριών, παρουσιάζονται μερικά αντιπροσωπευτικά εργαλεία κάθε κατηγορίας σε μια προσπάθεια να εντοπιστούν λειτουργίες αιχμής και να σχεδιαστούν οι προτεινόμενες λύσεις στα πλαίσια του OpenBio-C. Οι δυο αυτές κατηγορίες είναι: (α) **Εργαλεία νοητικής χαρτογράφησης** (mind mapping tools): επιτρέπουν την αναπαράσταση ιδεών και εννοιών που μπορούν να διασυνδεθούν προκειμένου να σχηματίσουν χρήσιμα διαγράμματα. Δίνουν ιδιαίτερη έμφαση στην οπτικοποίηση και στη δομή των ιδεών με σκοπό τη μελέτη και την οργάνωση πληροφοριών ώστε, τελικά, να επιτευχθεί η επίλυση προβλημάτων και η λήψη αποφάσεων, και (β) **Εργαλεία διαλεκτικής συνεργασίας** (argumentative collaboration tools): επιτρέπουν τη δημιουργία και διαχείριση διαλεκτικών συζητήσεων, όπου μια ομάδα ανθρώπων ανταλλάσσει θέσεις και επιχειρήματα, προκειμένου να επιτευχθεί συναίνεση σχετικά με ένα προς συζήτηση θέμα.

#### 3.3.1 Εργαλεία νοητικής χαρτογράφησης

Τα εργαλεία νοητικής χαρτογράφησης επιτρέπουν τη δημιουργία και επεξεργασία των λεγόμενων “νοητικών χαρτών” (mind maps). Ένας νοητικός χάρτης είναι ένα διάγραμμα που μπορεί να αναπαραστήσει λέξεις, ιδέες, εργασίες ή άλλα στοιχεία που συνδέονται και είναι διατεταγμένα γύρω από μια κεντρική λέξη ή ιδέα (Σχήμα 1). Οι νοητικοί χάρτες χρησιμοποιούνται κυρίως για την παραγωγή, οπτικοποίηση, δόμηση και ταξινόμηση ιδεών, και ενεργούν ως βοήθημα για την μελέτη και την οργάνωση των πληροφοριών, την επίλυση προβλημάτων και τη διευκόλυνση της διαδικασίας λήψης αποφάσεων. Ένας νοητικός χάρτης θα μπορούσε να θεωρηθεί ως μια απεικόνιση/επισκόπηση ενός συγκεκριμένου κομματιού γνώσης. Είναι μια συλλογή από “θέματα” τα οποία συγκεντρώνονται γύρω από μια κεντρική ιδέα (μπορεί να είναι μια μόνο λέξη ή μια ολόκληρη φράση) που αποτελεί και το κεντρικό θέμα. Γύρω από την κεντρική ιδέα, σε κυκλική συνήθως διάταξη και με γραμμές που τις συνδέουν με την κεντρική ιδέα, προστίθενται άλλες ιδέες και έννοιες. Αντιπροσωπευτικά εργαλεία αυτής της κατηγορίας είναι:

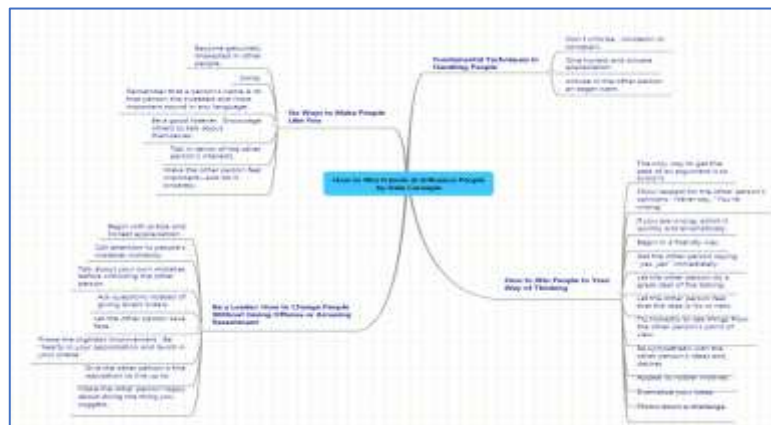
<sup>17</sup> [en.wikipedia.org/wiki/Collaborative\\_software](http://en.wikipedia.org/wiki/Collaborative_software)





Σχήμα 1. Ένα παράδειγμα ενός νοητικού χάρτη<sup>18</sup>

**MindMeister.** Το *MindMeister*<sup>19</sup> (Error! Reference source not found.) είναι ένα διαδικτυακό εργαλείο που αξιοποιεί τους νοητικούς χάρτες με στόχο, κατά κύριο λόγο, την υποστήριξη της διαχείρισης φαινομένων *καταιγισμού ιδεών*. Επιτρέπει τη δημιουργία νοητικών χαρτών, έχοντας στο κέντρο τη βασική ιδέα του υπό εξέταση θέματος. Οι χρήστες του MindMeister μπορούν να εργαστούν παράλληλα στον ίδιο νοητικό χάρτη, σε πραγματικό χρόνο, ενώ είναι εφικτή και η εξαγωγή νοητικών χαρτών σε διάφορες μορφές αρχείων. Παράλληλα, παρέχεται η δυνατότητα ασύγχρονης δημιουργίας και επεξεργασίας ενός νοητικού χάρτη (offline mode) μέσω της χρήσης της μη διαδικτυακής (desktop) έκδοσης που επιτρέπει τη δημιουργία και επεξεργασία νοητικών χαρτών και, στη συνέχεια, την ανάρτησή τους στη διαδικτυακή έκδοση του εργαλείου. Η Διεπαφή Προγραμματισμού Εφαρμογών (Application Programming Interface - API) του MindMeister επιτρέπει την πρόσβαση και επεξεργασία των διαθέσιμων νοητικών χαρτών από εφαρμογές τρίτων. Προκειμένου να ενισχυθεί η συνεργασία σε περιβάλλοντα με αυξημένο όγκο δεδομένων και με πολλαπλούς χρήστες, παρέχονται κατάλληλες ειδοποιήσεις μέσω e-mail ή SMS προκειμένου ένας χρήστης να είναι ενήμερος για τις αλλαγές που πραγματοποιούν άλλοι χρήστες σε νοητικούς χάρτες στη δημιουργία των οποίων έχει συμβάλει ο συγκεκριμένος χρήστης.



Σχήμα 2. Ένας νοητικός χάρτης στο MindMeister<sup>20</sup>

Οι χρήστες μπορούν επίσης να χρησιμοποιήσουν το εργαλείο εστίασης (zoom in/out) προκειμένου να περιηγηθούν και να εργαστούν σε νοητικούς χάρτες με μεγάλο αριθμό θεμάτων. Σε μια τέτοια περίπτωση, μπορεί να χρησιμοποιηθεί η λειτουργία κατάρρευσης/επέκτασης (collapse/expand) των επιμέρους θεμάτων του κεντρικού θέματος, η οποία βοηθά ώστε να δοθεί μια συνολική εικόνα του νοητικού χάρτη. Το εργαλείο φιλτραρίσματος (filtering) μπορεί να αξιοποιηθεί προκειμένου να απομονωθεί ένα μέρος του νοητικού χάρτη με τη χρήση συγκεκριμένων κριτηρίων που περιλαμβάνουν το όνομα του χρήστη που έχει επεξεργαστεί μέρος του νοητικού χάρτη, το κείμενο μιας ιδέας σε ένα νοητικό χάρτη ή τα προκαθορισμένα εικονίδια (που ισοδυναμούν με μια σειρά από ετικέτες) που πιθανόν έχουν προστεθεί σε μια ιδέα. Τέλος,

<sup>18</sup> Πηγή: [artificialeyes.net/files/adjusting-to-eye-loss-mind-map-770.jpg](http://artificialeyes.net/files/adjusting-to-eye-loss-mind-map-770.jpg)

<sup>19</sup> [www.mindmeister.com](http://www.mindmeister.com)

<sup>20</sup> Πηγή: [www.mindmeister.com/40950677/how-to-win-friends-influence-people](http://www.mindmeister.com/40950677/how-to-win-friends-influence-people)



είναι διαθέσιμα σε κάθε χρήστη τα πλήρη στοιχεία της ιστορικότητας κάθε νοητικού χάρτη. Με αυτόν τον τρόπο, παρέχεται δυνατότητα ανάκτησης οποιασδήποτε έκδοσης του νοητικού χάρτη, από τη φάση της δημιουργίας του μέχρι την τελευταία του μορφή (versioning).

**Mindomo.** Το *Mindomo*<sup>21</sup> (Σχήμα 3) έχει σχεδιαστεί για να υποστηρίξει τη δημιουργία νοητικών χαρτών και να ενισχύσει τη συνεργασία σε εφαρμογές που είτε είναι ατομικές είτε σχετίζονται με επιχειρήσεις και την εκπαίδευση. Οι νοητικοί χάρτες του Mindomo είναι σε δενδρική μορφή και ξεκινούν από ένα και μόνο σημείο (το οποίο μπορεί να είναι είτε μια λέξη είτε μια ολόκληρη πρόταση). Ένας νοητικός χάρτης του Mindomo έχει ένα κεντρικό θέμα, το οποίο στη συνέχεια «αποσυντίθεται» σε άλλα θέματα ή υποκατηγορίες. Ο χρήστης του Mindomo είναι σε θέση να διασυνδέει το κεντρικό θέμα με τα υπόλοιπα θέματα ή υποκατηγορίες χρησιμοποιώντας τα εργαλεία συσχέτισης ή να ομαδοποιήσει ένα υποσύνολο των αντικειμένων του νοητικού χάρτη χρησιμοποιώντας μια σειρά από εργαλεία ομαδοποίησης. Εκτός από τη διαδικτυακή έκδοση, υπάρχει και μια μη διαδικτυακή έκδοση που επιτρέπει την δημιουργία/επεξεργασία νοητικών χαρτών και, στη συνέχεια, τη δημοσιοποίησή τους στη διαδικτυακή έκδοση του Mindomo.



**Σχήμα 3.** Ένας νοητικός χάρτης στο Mindomo<sup>22</sup>

Για την ενίσχυση της συνεργασίας σε περιβάλλοντα με αυξημένο όγκο δεδομένων, το Mindomo υποστηρίζει τη λειτουργία της επέκτασης/κατάρρευσης τμημάτων του νοητικού χάρτη. Επιπλέον, υποστηρίζεται και η προβολή/απόκρυψη τμημάτων του νοητικού χάρτη που ικανοποιούν επιλεγμένα από το χρήστη κριτήρια. Στην οπτική αναπαράσταση, οι χρήστες μπορούν να προσθέσουν ετικέτες (ειδικά εικονίδια) προκειμένου να είναι δυνατή η παροχή περισσότερων πληροφοριών σχετικά με το θέμα (όπως για παράδειγμα, το πόσο σημαντικό είναι αυτό το θέμα). Το παρεχόμενο εργαλείο εστίασης (zoom in/out) καθιστά ευκολότερη την περιήγηση σε μεγάλους χάρτες. Η ιστορικότητα ενός νοητικού χάρτη του Mindomo ουσιαστικά είναι μια λίστα με όλες τις ενέργειες που έχουν πραγματοποιηθεί στο συγκεκριμένο νοητικό χάρτη. Το εργαλείο φιλτραρίσματος μπορεί να χρησιμοποιηθεί για την απομόνωση τμημάτων του χάρτη που ικανοποιούν μια σειρά διαφορετικών κριτηρίων. Τέλος, το εργαλείο αναζήτησης επιτρέπει την αναζήτηση ενός θέματος και υποθεμάτων των οποίων το περιεχόμενο ταιριάζει με το συγκεκριμένο κείμενο.



**Σχήμα 4.** Ένας νοητικός χάρτης στο Bubbl.us<sup>23</sup>

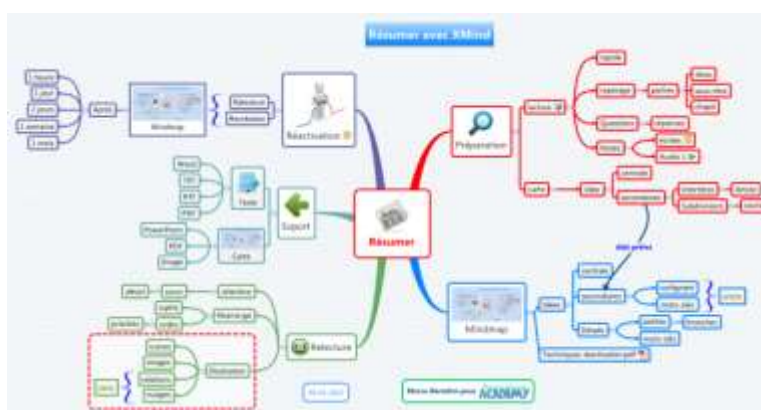
21 [www.mindomo.com](http://www.mindomo.com)

<sup>22</sup> [www.mindomo.com/it/mindmap/mindomo-tutorial-learn-by-doing-57885d9859e13b331edfdfac56e1b859](http://www.mindomo.com/it/mindmap/mindomo-tutorial-learn-by-doing-57885d9859e13b331edfdfac56e1b859)

<sup>23</sup> Πηγή: [bubbl.us/examples](https://bubbl.us/examples)

**Bubbl.us.** Το *Bubbl.us*<sup>24</sup> (Error! Reference source not found.) είναι ένα εργαλείο δημιουργίας νοητικών χαρτών, παρόμοιο με το MindMeister, που αποσκοπεί στην υποστήριξη της διαχείρισης φαινομένων καταιγισμού ιδεών. Με το Bubbl.us, ο χρήστης είναι σε θέση να δημιουργήσει διαδικτυακούς νοητικούς χάρτες και να τους διαμοιραστεί με άλλους χρήστες. Ένας νοητικός χάρτης που δημιουργήθηκε με αυτό το εργαλείο μπορεί να ενσωματωθεί σε ένα blog ή μια ιστοσελίδα, να αποθηκευτεί ως εικόνα ή να αποσταλεί μέσω ηλεκτρονικού ταχυδρομείου. Ένας τέτοιος χάρτης αποτελείται από φυσαλίδες (bubbles). Το κείμενο, το μέγεθος και το χρώμα της φυσαλίδας είναι χαρακτηριστικά που μπορεί να καθορίσει ο χρήστης. Δύο φυσαλίδες μπορεί να συνδέονται με μια κατευθυνόμενη γραμμή (σύνδεσμος). Το εργαλείο αναίρεσης (Undo) επιτρέπει στους χρήστες να ανακαλέσουν όλα τα βήματα που έχουν λάβει χώρα για τη δημιουργία ενός νοητικού χάρτη. Οι χάρτες μπορούν να εξαχθούν σε διάφορες μορφές (XML, HTML, διάφοροι τύποι εικόνας). Για να αντιμετωπιστούν καταστάσεις που εμπεριέχουν υπερφόρτωση πληροφοριών, το Bubbl.us παρέχει ένα εργαλείο εστίασης (zoom in/out) που μπορεί να χρησιμοποιηθεί αποτελεσματικά ειδικότερα στις περιπτώσεις που απαιτείται περιήγηση και κύλιση σε νοητικούς χάρτες με πολλές φυσαλίδες. Η συνεπεξεργασία νοητικών χαρτών είναι επίσης εφικτή κάτι που συνεπάγεται, ουσιαστικά, την ανταλλαγή νοητικών χαρτών μεταξύ των μελών μιας ομάδας χρηστών.

**XMind.** Ο νοητικός χάρτης του *XMind*<sup>25</sup> (Σχήμα 5) ακολουθεί την κλασική μορφή των νοητικών χαρτών που απαντάται και σε άλλα παρόμοια και δημοφιλή εργαλεία νοητικής χαρτογράφησης: η κύρια ιδέα είναι η ρίζα που βρίσκεται στο κέντρο του χάρτη και κλαδιά καταλήγουν στην κεντρική ρίζα του δένδρου. Για να υποστηριχθούν διαδικασίες που σχετίζονται με τον τομέα των επιχειρήσεων, το XMind παρέχει δυνατότητες που σχετίζονται με τη διαχείριση έργου (project management) και τα ορόσημα (milestones) του έργου. Παρέχεται ένα σύνολο από διαφορετικές απεικονίσεις μέσω διαγραμμάτων τύπου fisheye και ishikawa, δενδρικών διαγραμμάτων και οργανογραμμάτων. Σε ένα νοητικό χάρτη είναι δυνατό να προστεθούν όρια, συσχετίσεις, δείκτες, ετικέτες, σημειώσεις, ηχητικές σημειώσεις, υπερσύνδεσμοι και γραφικά. Μπορούν επίσης να προστεθούν αρχεία, είτε ως νέα θέματα είτε ως συνημμένα, σε ήδη υπάρχοντα θέματα που λειτουργούν ως δευτερεύοντα. Υποστηρίζεται η εξαγωγή νοητικών χαρτών σε μια ποικιλία διαφορετικών μορφών. Υποστηρίζεται η διαδικτυακή συνεργασία πολλαπλών χρηστών. Για την αντιμετώπιση προβλημάτων που προκύπτουν σε νοητικούς χάρτες που περιλαμβάνουν αυξημένο όγκο δεδομένων, το XMind υποστηρίζει ένα μηχανισμό φιλτραρίσματος που επιτρέπει στους χρήστες να εμφανίσουν ή να αποκρύψουν ένα συγκεκριμένο τμήμα του νοητικού χάρτη επιλέγοντας συγκεκριμένα κριτήρια (δείκτες ή ετικέτες). Τα θέματα που πληρούν τα κριτήρια φιλτραρίσματος που παρέχει ο χρήστης επισημαίνονται κατάλληλα, ενώ τα θέματα που δεν πληρούν τα κριτήρια αποκρύπτονται. Το εργαλείο επέκτασης/κατάρρευσης επιμέρους θεμάτων ενός θέματος μπορεί επίσης να είναι χρήσιμο, ιδιαίτερα σε περιπτώσεις χαρτών με μεγάλο αριθμό θεμάτων.



Σχήμα 5. Ένας νοητικός χάρτης στο XMind<sup>26</sup>

<sup>24</sup> [www.bubbl.us](http://www.bubbl.us)

<sup>25</sup> [www.xmind.net](http://www.xmind.net)

<sup>26</sup> Πηγή: [trouvetavoie.files.wordpress.com/2012/01/rc3a9sumer-avec-xmind.png?w=700](http://trouvetavoie.files.wordpress.com/2012/01/rc3a9sumer-avec-xmind.png?w=700)

### 3.3.2 Εργαλεία υποστήριξης συνεργασίας με επιχειρήματα

Η **επιχειρηματολογία** είναι μια λεκτική δραστηριότητα, η οποία συνήθως διεξάγεται σε απλή γλώσσα. Ένας άτομο που εμπλέκεται στην επιχειρηματολογία, χρησιμοποιεί λέξεις και φράσεις για να δηλώσει, να ρωτήσει ή να αρνηθεί κάτι, να ανταποκριθεί στις δηλώσεις, στις ερωτήσεις και στις αρνήσεις κάποιου άλλου και ούτω καθεξής (van Eemeren *et al.*, 1996). Η επιχειρηματολογία είναι επίσης μια κοινωνική δραστηριότητα η οποία, καταρχήν, απευθύνεται σε άλλους ανθρώπους και είναι άμεσα συνδεδεμένη με την εξαγωγή συμπερασμάτων μέσω τεκμηριωμένου συλλογισμού. Αφορά πάντα μια συγκεκριμένη άποψη για ένα συγκεκριμένο θέμα και η ανάγκη για επιχειρηματολογία προκύπτει όταν υπάρχει διάσταση απόψεων σχετικά με το θέμα αυτό. Τα συστήματα υποστήριξης επιχειρηματολογίας είναι συστήματα λογισμικού που έχουν σχεδιαστεί για να βοηθούν τους ανθρώπους να λάβουν μέρος σε διάφορους τύπους διαλόγων κατά τη διάρκεια των οποίων ανταλλάσσονται επιχειρήματα. Τέτοια συστήματα έχουν χρησιμοποιηθεί σε πεδία όπως το εμπόριο, η εκπαίδευση, το δίκαιο και ο σχεδιασμός (van Eemeren *et al.*, 1996). Οι τεχνολογίες που υποστηρίζουν τη συνεργασία με επιχειρήματα συνήθως παρέχουν τα κατάλληλα μέσα για τη δόμηση της συζήτησης, την οπτικοποίηση της, την ανταλλαγή εγγράφων και τη διαχείριση των χρηστών. Υποστηρίζουν συνεργασία με επιχειρήματα σε διάφορα επίπεδα και έχουν δοκιμαστεί με διαφορετικές ομάδες χρηστών και διαφορετικά πλαίσια. Επιπλέον, στοχεύουν στο να χρησιμοποιήσουν την επιχειρηματολογία ως μέσο για να δημιουργηθεί μια κοινή βάση μεταξύ των διαφόρων ενδιαφερομένων φορέων ώστε να γίνουν κατανοητές συγκεκριμένες θέσεις σε θέματα, να προκύψουν παραδοχές/κριτήρια και να δομηθεί συλλογική συναίνεση. Αντιπροσωπευτικά εργαλεία αυτής της κατηγορίας περιγράφονται στη συνέχεια.

**Araucaria.** Το *Araucaria*<sup>27</sup> επιτρέπει την ανάλυση επιχειρημάτων μέσω διαγραμμάτων. Τα επιχειρήματα στο *Araucaria* δημιουργούνται από την επιλογή φράσεων ενός κειμένου που επιλέγει ο χρήστης. Κάθε επιλεγμένο τμήμα κειμένου αντιστοιχεί σε έναν κόμβο στο διάγραμμα επιχειρημάτων και οι γραμμές (συσχετίσεις) μπορούν να μετακινηθούν από ένα κόμβο (υπόθεση) σε έναν άλλο κόμβο (συμπέρασμα). Ένας χρήστης μπορεί να προσθέσει υποθέσεις στο διάγραμμα, να δημιουργήσει/ τροποποιήσει/ ανακτήσει/ αποθηκεύσει ένα σύστημα επιχειρηματολογίας ή να διαγράψει τα συστατικά στοιχεία (κόμβοι/συσχετίσεις) του διαγράμματος. Τα επιχειρήματα αποθηκεύονται στο δίσκο σε μια, κατάλληλη για μεταφορά σε άλλα συστήματα, μορφή που βασίζεται στη γλώσσα AML (Argument Markup Language) η οποία προσομοιάζει με την XML.



Σχήμα 6. Ένας χάρτης ιδεών στο DebateGraph<sup>28</sup>

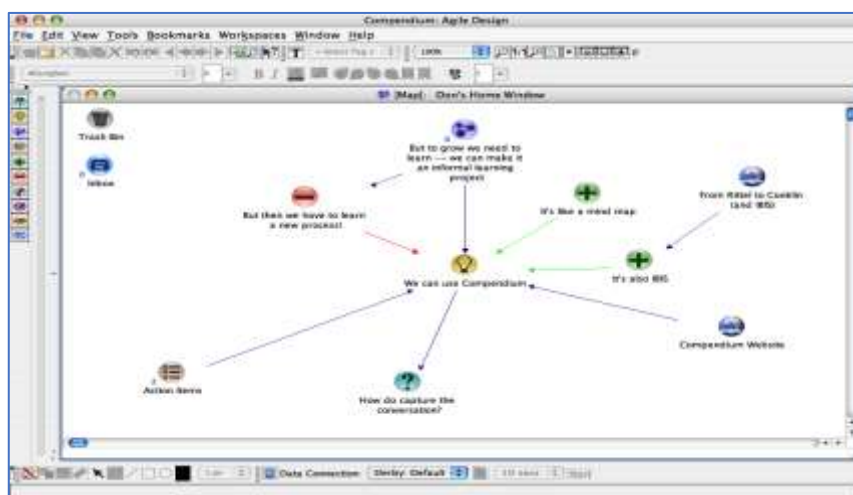
**DebateGraph.** Το *DebateGraph*<sup>29</sup> (Error! Reference source not found.) είναι μια άλλη ενδιαφέρουσα προσέγγιση για την υποστήριξη συνεργασίας με επιχειρήματα. Ένας χάρτης στο DebateGraph περιλαμβάνει δύο κύριες συνιστώσες, τις «ιδέες» και τις «συσχετίσεις» μεταξύ δύο ιδεών. Διάφοροι μηχανισμοί έχουν υλοποιηθεί για την υποστήριξη μεγάλης κλίμακας επιχειρηματολογίας και τη συνεργασία σε περιβάλλοντα με αυξημένο όγκο δεδομένων. Για παράδειγμα, ο μηχανισμός ιστορικότητας επιτρέπει σε

<sup>27</sup> [araucaria.computing.dundee.ac.uk/doku.php](http://araucaria.computing.dundee.ac.uk/doku.php)

<sup>28</sup> Πηγή: [debategraph.org/Stream.aspx?nid=610&vt=bubble&dc=focus](http://debategraph.org/Stream.aspx?nid=610&vt=bubble&dc=focus)

29 [debategraph.org](http://debategraph.org)

ένα χρήστη να περιηγηθεί στις ενέργειες των άλλων χρηστών. Επιπλέον, υποστηρίζεται μια προοδευτική απεικόνιση του χάρτη επιχειρηματολογίας με την έννοια ότι μόνο ένα μέρος του χάρτη εμφανίζεται στην οθόνη του χρήστη (μια «ιδέα» και οι «ιδέες» που είναι άμεσα συνδεδεμένες με αυτή). Ένας χρήστης μπορεί να πλοηγηθεί σε ένα χάρτη μετακινούμενος κατά ένα επίπεδο κάθε φορά. Το DebateGraph παρέχει επίσης μια σειρά από μηχανισμούς ενημέρωσης. Ένας χρήστης μπορεί να ενημερώνεται για τις αλλαγές σε ένα χάρτη, είτε μέσω ηλεκτρονικού ταχυδρομείου είτε μέσω RSS. Παρέχεται ακόμα ένας μηχανισμός αναζήτησης προκειμένου να είναι εφικτή η αναζήτηση ενός κειμένου τόσο σε ιδιωτικούς όσο και σε δημόσιους χάρτες. Τέλος, ο χρήστης έχει στη διάθεσή του διαφορετικές προβολές ενός χάρτη μεταξύ των οποίων η προβολή εστίασης-μεγέθυνσης, η δενδρική προβολή (ιεραρχική προβολή), η προβολή εξερεύνησης (συμπεριλαμβανομένου ενός εργαλείου εστίασης) και οι προβολές "macro" και "hub" (που περιλαμβάνει ένα χώρο συζήτησης για κάθε χάρτη).



Σχήμα 7. Ένας χάρτης στο Compendium<sup>30</sup>

**Compendium.** Το *Compendium*<sup>31</sup> (Error! Reference source not found.) είναι ένα ακόμη εργαλείο που έχει σχεδιαστεί για τη χαρτογράφηση πληροφοριών, ιδεών και επιχειρημάτων. Οι ιδέες σε ένα χάρτη του Compendium εκφράζονται με διαφορετικούς τύπους κόμβων που συνδέονται μεταξύ τους με διαφορετικούς τύπους συσχετίσεων. Για τη διαχείριση χαρτών με αυξημένο όγκο δεδομένων, το Compendium περιλαμβάνει μια σειρά από χαρακτηριστικά. Πιο συγκεκριμένα, υποστηρίζονται χάρτες πολλαπλών επιπέδων (ένας χάρτης μπορεί να περιλαμβάνει έναν άλλο χάρτη). Υπάρχει επίσης ένα εργαλείο εστίασης που διευκολύνει την εστίαση και την περιήγηση σε χάρτες με μεγάλο αριθμό κόμβων. Ο μηχανισμός αναζήτησης που έχει υλοποιηθεί λαμβάνει υπόψη μια σειρά από κριτήρια, όπως την ετικέτα ενός κόμβου, το όνομα του δημιουργού του, την ημερομηνία και το κείμενο του κόμβου. Προκειμένου να αλλάξει τη θέση των κόμβων σε ένα χάρτη, ο χρήστης μπορεί να επιλέξει και να μεταφέρει πολλαπλούς κόμβους (περισσότεροι από ένας κόμβοι επιλέγονται δημιουργώντας ένα ορθογώνιο που τους περιέχει και, στη συνέχεια, σύρονται μαζί). Τέλος, ένας χρήστης είναι σε θέση να αποθηκεύσει «σελιδοδείκτες» των κόμβων σε ένα χάρτη ώστε, σε μετέπειτα στάδιο, να είναι εύκολος ο εντοπισμός τους, ιδιαίτερα στις περιπτώσεις μεγάλων χαρτών.

**CoPe\_it!** Το *CoPe\_it!*<sup>32</sup> (Σχήμα 8) (Karakapilidis et al, 2009) είναι ένα Web 2.0 εργαλείο που έχει σχεδιαστεί για να ενισχύσει τη συνεργασία μέσω της ανταλλαγής απόψεων και πόρων σε κοινότητες πρακτικής. Ένας χρήστης του CoPe\_it! μπορεί να δημιουργήσει ένα προσωπικό ή ένα συνεργατικό (κοινό) χώρο εργασίας, να συμμετέχει και να συμβάλλει σε έναν υπάρχοντα χώρο εργασίας και να προσθέσει ή να διαμοιραστεί πόρους μέσω ενός χώρου εργασίας. Το CoPe\_it! επιτρέπει τη δημιουργία πολλών και διαφορετικών αντικειμένων που χρησιμοποιούνται για την ανάπτυξη της επιχειρηματολογίας, τα οποία μπορεί να συνδέονται μεταξύ τους με μια σειρά από συσχετίσεις. Οι χρήστες μπορούν να συνεργαστούν είτε με ασύγχρονο είτε με σύγχρονο τρόπο. Η προσέγγιση που έχει ακολουθηθεί στο CoPe\_it! είναι επεκτάσιμη (scalable) προκειμένου να είναι εφικτή η υποστήριξη των ποικιλόμορφων αιτημάτων των χρηστών: τα χαρακτηριστικά και οι

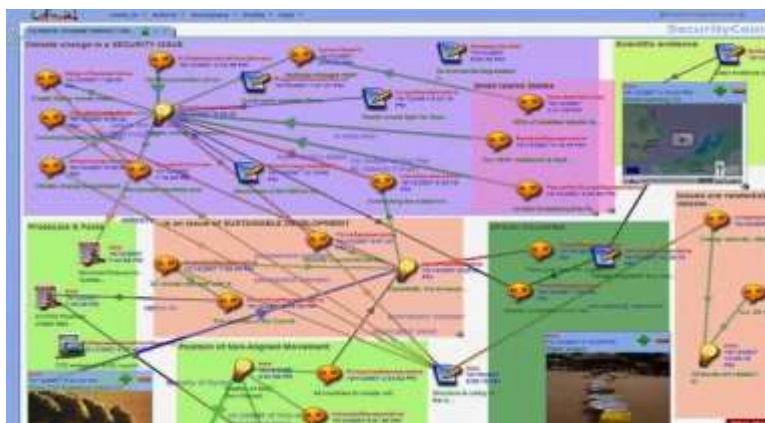
<sup>30</sup> Πηγή: [bddl.blogspot.gr/2009/11/planning-in-agile-learning-design.html](http://bddl.blogspot.gr/2009/11/planning-in-agile-learning-design.html)

<sup>31</sup> [compendium.open.ac.uk](http://compendium.open.ac.uk)

<sup>32</sup> [copeit.cti.gr](http://copeit.cti.gr)



λειτουργίες του εργαλείου διαφοροποιούνται προκειμένου να είναι κατανοητά από τον άνθρωπο (human understandable) αλλά, παράλληλα, να είναι και κατανοητά από τη μηχανή (machine understandable). Για την επίτευξη αυτού του στόχου, έχουν υλοποιηθεί στο πλαίσιο του CoPe\_it! πολλαπλές προβολές του χώρου εργασίας. Το CoPe\_it! υποστηρίζει μια σειρά από χαρακτηριστικά (Karacapilidis *et al.*, 2009) για να ενισχύσει τη συνεργασία σε περιπτώσεις με αυξημένο όγκο δεδομένων.



Σχήμα 8. Η αναπαράσταση ενός χώρου εργασίας στο CoPe\_it!<sup>33</sup>

Για παράδειγμα, παρέχεται μια μικρογραφία του χώρου εργασίας (minimap) προκειμένου να είναι διαθέσιμη μια επισκόπηση των αντίστοιχων περιεχομένων του. Επίσης, διατίθεται και ένας μηχανισμός ιστορικότητας, μέσω του οποίου μπορεί κανείς να παρακολουθήσει την εξέλιξη ενός χώρου εργασίας στη διάρκεια του χρόνου. Πολλά στοιχεία ενός χώρου εργασίας μπορούν να ομαδοποιηθούν ώστε να είναι ευκολότερη η διαχείρισή τους. Τέλος, ένας μηχανισμός φιλτραρίσματος επιτρέπει σε ένα χρήστη να εμφανίσει τα αντικείμενα της επιχειρηματολογίας που πληρούν συγκεκριμένα κριτήρια (ή να αποκρύψει αντικείμενα που δεν πληρούν κάποια κριτήρια). Παραδείγματα τέτοιων κριτηρίων είναι ο τίτλος, η ημερομηνία, ο συγγραφέας και ο τύπος του αντικειμένου επιχειρηματολογίας.

**Cohere.** Το Cohere<sup>34</sup> (Σχήμα 9) είναι ένα διαδικτυακό εργαλείο που χρησιμοποιείται προκειμένου μια ομάδα χρηστών να δημιουργήσει και να διαμοιράσει τις ιδέες της, οι οποίες μπορούν να αποθηκευτούν σε ιδιωτικούς ή δημόσιους χώρους εργασίας καθώς και σε χώρους εργασίας που επιτρέπουν την πρόσβαση μόνο στα μέλη μιας επιλεγμένης ομάδας. Το Cohere μιμείται τους πιο δημοφιλείς διαδικτυακούς κοινωνικούς τόπους καθώς υιοθετεί την ιδέα των μελών, των σελίδων των μελών και των ομάδων. Οι ιδέες που δημιουργούνται από έναν χρήστη συγκεντρώνονται στην προσωπική ιστοσελίδα του χρήστη (μια μορφή προσωπικής ιστοσελίδας-σημειωματάριου των ιδεών του συγκεκριμένου χρήστη). Το σημειωματάριο του Cohere περιλαμβάνει τις ιστοσελίδες που έχουν δημιουργηθεί από τους χρήστες, τα σχολιασμένα κομμάτια τους, τις συνδέσεις που έχουν δημιουργηθεί μεταξύ των ιδεών και τον κατάλογο των ατόμων/ομάδων με τους οποίους ο χρήστης διαμοιράζεται κάποιο περιεχόμενο (Liddo and Shum, 2010).



Σχήμα 9. Ένας χώρος εργασίας στο Cohere<sup>35</sup>

<sup>33</sup> Πηγή: [copeit.cti.gr](http://copeit.cti.gr)

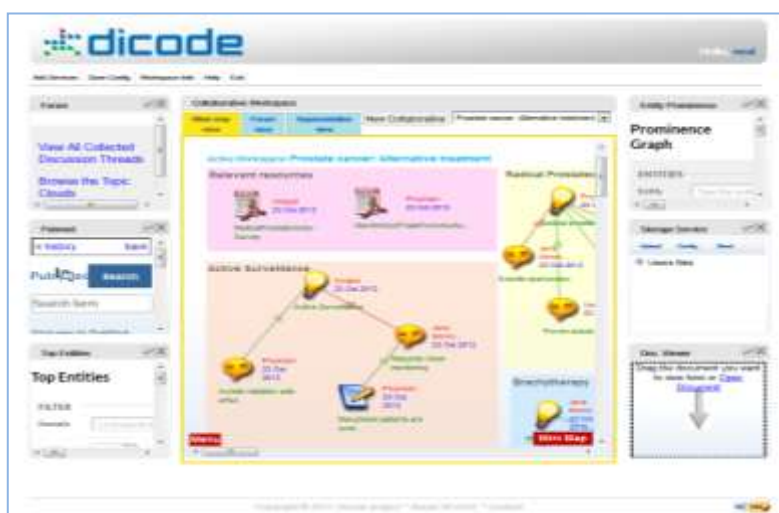
<sup>34</sup> [cohere.open.ac.uk](http://cohere.open.ac.uk)

<sup>35</sup> Πηγή: [learningemergence.net/wp-content/uploads/2010/10/FirefoxScreenSnapz469.jpg](http://learningemergence.net/wp-content/uploads/2010/10/FirefoxScreenSnapz469.jpg)



Για την αντιμετώπιση καταστάσεων που περιλαμβάνουν αυξημένο όγκο δεδομένων, το Cohere προσφέρει διάφορους μηχανισμούς φιλτραρίσματος (π.χ. οι ιδέες και οι συνδέσεις τους μπορεί να φιλτράρονται ανάλογα με τον τύπο τους). Είναι επίσης εφικτός και ο σχολιασμός-προσθήκη ετικετών (tagging) των ιδεών, ενώ ο μηχανισμός αναζήτησης μπορεί να αξιοποιήσει το κείμενο των ιδεών, τις σχετικές ετικέτες, το όνομα του χρήστη (ή της ομάδας) και το κείμενο που εμφανίζεται σε μια σύνδεση ώστε να παρέχει αξιόπιστα αποτελέσματα αναζήτησης όταν ο χρήστης το επιθυμεί.

**Dicode.** Ο στόχος του ευρωπαϊκού ερευνητικού έργου Dicode<sup>36</sup>, το οποίο υλοποιήθηκε στα πλαίσια του Ευρωπαϊκού Προγράμματος FP7, ήταν να διευκολύνει και να βελτιώσει τη συνεργασία και την ομαδική λήψη αποφάσεων σε περιβάλλοντα που χαρακτηρίζονται από προβλήματα υπερφόρτωσης πληροφοριών και γνωστικής πολυπλοκότητας. Το λογισμικό που αναπτύχθηκε στα πλαίσια του έργου μπορεί να θεωρηθεί ως ένας καινοτόμος «πάγκος εργασίας» (workbench) που ενσωματώνει και εννοχρηστρώνει ένα σύνολο διαλειτουργικών υπηρεσιών που έχουν ως στόχο να μειώσουν δραστικά τα παραπάνω προβλήματα, οδηγώντας τα σε διαχειρίσιμα επίπεδα (βλ. Σχήμα 10). Με αυτό τον τρόπο, οι εμπλεκόμενοι φορείς-χρήστες γίνονται περισσότερο παραγωγικοί στο έργο τους και μπορούν να επικεντρωθούν σε δημιουργικές και καινοτόμες δραστηριότητες. Κεντρικό ρόλο μεταξύ των υπηρεσιών του πάγκου εργασίας του Dicode κατέχουν οι υπηρεσίες συνεργασίας και λήψης αποφάσεων οι οποίες, μεταξύ άλλων, διευκολύνουν τη σύγχρονη και ασύγχρονη συνεργασία των χρηστών μέσω προσαρμοζόμενων χώρων εργασίας, διαχειρίζονται αποτελεσματικά την αναπαράσταση και την οπτικοποίηση των αποτελεσμάτων των υπηρεσιών εξόρυξης δεδομένων (μέσω εναλλακτικών και εστιασμένων σχημάτων οπτικοποίησης δεδομένων), και επιτρέπουν την εννοχρήστρωση μιας σειράς δράσεων για τον κατάλληλο χειρισμό των δεδομένων. Όσον αφορά στη λήψη αποφάσεων, οι υπηρεσίες αυτές: (α) αποσκοπούν στην ενίσχυση (σε ατομικό και ομαδικό επίπεδο) της διαδικασίας εξαγωγής νοήματος (sense making), (β) υποστηρίζουν αποτελεσματικά τους χρήστες όσον αφορά τη διαδικασία του εντοπισμού, της ανάκτησης και της ανάπτυξης επιχειρημάτων για σχετικές πληροφορίες και γνώσεις, και (γ) παρέχουν κατάλληλα ενημερωτικά μηνύματα και συστάσεις (που λαμβάνουν υπόψη παραμέτρους όπως οι προτιμήσεις, οι ικανότητες, η εμπειρία των χρηστών κλπ). Οι υπηρεσίες αυτής της κατηγορίας βασίζονται σε μια βαθμιαία “αυστηροποίηση της αναπαράστασης” (formalization) της συνεργασίας και αξιοποιούν μια σειρά από μηχανισμούς συλλογισμού για την υποστήριξη των χρηστών σε καθημερινές διαδικασίες λήψης αποφάσεων (Tzagarakis *et al.*, 2014).



Σχήμα 10. Ο πάγκος εργασίας του Dicode<sup>37</sup>

Η προτεινόμενη προσέγγιση προσφέρει τις ακόλουθες εναλλακτικές απεικονίσεις (Views) ενός χώρου συνεργασίας: • *Forum View*. Ένας χώρος συνεργασίας εμφανίζεται ως μια παραδοσιακή διαδικτυακή συζήτηση όπου οι θέσεις των μελών της συζήτησης εμφανίζονται με αύξουσα χρονολογική σειρά. Οι χρήστες είναι σε θέση να δημοσιεύσουν νέα μηνύματα στο χώρο συνεργασίας τα οποία προβάλλονται στο τέλος της

<sup>36</sup> [dicode-project.cti.gr](http://dicode-project.cti.gr)

<sup>37</sup> Πηγή: Dicode Deliverable D5.4.2, [dicode-project.cti.gr](http://dicode-project.cti.gr)

λίστας των μηνυμάτων. Ο στόχος αυτής της απεικόνισης είναι να επιτραπεί ο συλλογισμός και η ανταλλαγή απόψεων, χωρίς να περιορίζεται η εκφραστικότητα των συμμετεχόντων. • *Mind-map View*. Ένας χώρος συνεργασίας εμφανίζεται ως νοητικός χάρτης που επιτρέπει μια μη αυστηρή (informal) αναπαράσταση και συσχέτιση μεταξύ των αντικειμένων συνεργασίας, ενώ παράλληλα υιοθετεί και μια συγκεκριμένη σημασιολογία όσον αφορά στους τύπους των αντικειμένων συνεργασίας και τις μεταξύ τους συσχετίσεις. • *Neighborhood View*. Η συγκεκριμένη απεικόνιση επιτρέπει στους χρήστες να επιλέξουν ένα συγκεκριμένο αντικείμενο συνεργασίας από το νοητικό χάρτη και να εμφανίσουν μόνο εκείνα τα αντικείμενα συνεργασίας που σχετίζονται άμεσα με το υπό εξέταση αντικείμενο. • *Formal View*. Αυτή η απεικόνιση υιοθετεί ένα συγκεκριμένο μοντέλο επιχειρηματολογίας (συγκεκριμένα, το IBIS (Kunz and Rittel, 1970)) και περιλαμβάνει μια σειρά από αυστηρούς μηχανισμούς αξιολόγησης και συλλογισμού προκειμένου να βοηθήσει τους χρήστες να αντιληφθούν το αποτέλεσμα μιας συνεργασίας και να οδηγηθούν στη λήψη της καλύτερης (καλύτερα τεκμηριωμένης) απόφασης. • *Multi-Criteria Decision-Making View*. Παρέχει μια σειρά από αλγορίθμους που βασίζονται σε πολλαπλά κριτήρια λήψης αποφάσεων προκειμένου να υπολογίζεται κάθε φορά η λίστα των πιο ισχυρών εναλλακτικών λύσεων.

Κατά τη διάρκεια μιας συνεργασίας, κάθε χρήστης μπορεί να επιλέξει την απεικόνιση την οποία θεωρεί καταλληλότερη για το συγκεκριμένο στάδιο της συνεργασίας. Πρέπει να σημειωθεί ότι όλες οι απεικονίσεις συνεργασίας είναι ισοδύναμες, με την έννοια ότι αποτελούν απεικόνιση του ίδιου χώρου συνεργασίας, ενώ διαφέρουν στο ότι παρέχουν διαφορετικούς μηχανισμούς προκειμένου να υποστηρίξουν τη συνεργασία και, σε τελική φάση, τη λήψη αποφάσεων.

#### 4 ΣΧΕΔΙΑΖΟΝΤΑΣ ΕΥΕΛΙΚΤΑ ΚΑΙ ΑΝΟΙΧΤΑ ΠΕΡΙΒΑΛΛΟΝΤΑ ΕΠΙΣΤΗΜΟΝΙΚΩΝ ΡΟΩΝ ΕΡΓΑΣΙΑΣ: ΑΠΑΡΑΙΤΗΤΑ ΣΥΣΤΑΤΙΚΑ & ΛΕΙΤΟΥΡΓΙΚΟΤΗΤΑ

Παρόλο που η έννοια των ΒΕΡΕ είναι μια απλή ιδέα με ιστορική παρουσία στην κοινότητα βιοπληροφορικής, υπάρχουν πολλοί παράγοντες που επηρεάζουν τη συνολική τους ποιότητα και εξακολουθούν να αγνοούνται. Σε αυτή την ενότητα παρουσιάζουμε ορισμένους από αυτούς τους παράγοντες.

**Ενσωματώσου και ενσωμάτωσε.** Οποιαδήποτε ΒΕΡΕ θα πρέπει να έχει τη δυνατότητα να ενσωματώνει άλλες ΒΕΡΕ. Όπως προαναφέραμε, τα σύγχρονα περιβάλλοντα ΒΕΡΕ τείνουν να υποφέρουν από το “*σύνδρομο του κλειδώματος*”. Δηλαδή, απαιτούν από τους χρήστες να επενδύσουν σημαντικό χρόνο και πόρους για να ανασυστήσουν ένα εξωτερικό στοιχείο με τα απαιτούμενα δεδομένα μεταδεδομένων, βιβλιοθήκες και αρχεία διαμόρφωσης σε μια ροή εργασιών. Οι ροές εργασίας πρέπει να είναι *αγνωστικιστικές* (agnostic) σχετικά με τα πιθανά συστατικά μέρη τους και πρέπει να παρέχουν αποτελεσματικούς μηχανισμούς για την ενσωμάτωσή τους με ελάχιστη προσπάθεια. Ομοίως, ένα περιβάλλον ΒΕΡΕ δεν πρέπει να υποθέτει ότι θα είναι το κύριο εργαλείο με το οποίο ο ερευνητής κάνει την πλήρη ανάλυση. Αυτή η συμπεριφορά είναι “*εγωιστική*” και αποκαλύπτει την επιθυμία “*κυριαρχίας*” και όχι να συμβάλει σε μια ερευνητική κοινότητα. Με βάση αυτή τη διαπίστωση, τα περιβάλλοντα ΒΕΡΕ θα πρέπει να προσφέρουν στους ερευνητές τη δυνατότητα να εξαγουν την πλήρη ανάλυση τους σε μορφές που μπορούν εύκολα να αφομοιωθούν από άλλα συστήματα. Παραδείγματα περιλαμβάνουν απλά σενάρια BASH με *μεταδεδομένα* που περιγράφονται σε γνωστές και εύκολα διαχειρίσιμες μορφές όπως π.χ., XML, JSON ή YAML. Μια άλλη επιλογή είναι οι προγραμματικές δέσμες ενεργειών (scripts) σε σειριοποιημένα (serialized) αντικείμενα όπως η PICKLE και η CAMEL<sup>38</sup> που μπορούν εύκολα να ενσωματωθούν στο σύνολό τους από άλλα εργαλεία.

**Σημασιολογική και οντολογική υποστήριξη.** Τα συστήματα ροής εργασίας τείνουν να επικεντρώνονται περισσότερο στο αναλυτικό τμήμα του εργαλείου και να παραμελούν το σημασιολογικό μέρος. Ο **σημασιολογικός εμπλουτισμός** (semantic enrichment) μιας ροής εργασίας μπορεί να επιτευχθεί με τη συμμόρφωση και τη χρήση καθιερωμένων οντολογιών. Επομένως, η σημασιολογική ολοκλήρωση μεθόδων και δεδομένων μέσω οντολογιών μπορεί να αποτελέσει κοινό σημείο για άμεση σύγκριση των επιστημονικών ευρημάτων. Οι οντολογίες, για βιοτρόπεζες (Pang *et al.*, 2015), για κατηγορίες λειτουργικότητας γονιδίων (Gene Ontology Consortium, 2015), για τη γενετική ποικιλομορφία (Byrne *et al.*, 2012) και για τη περιγραφή και αναλυτική τεκμηρίωση φαινοτύπων (Robinson and Mundlos, 2010) είναι εξαιρετικά χρήσιμες. Φυσικά, οι οντολογίες δεν είναι η πανάκεια για την επίλυση των προαναφερθέντων

<sup>38</sup> [eev.ee/blog/2015/10/15/dont-use-pickle-use-camel](http://eev.ee/blog/2015/10/15/dont-use-pickle-use-camel)

προβλημάτων καθότι έχουν τα δικά τους ζητήματα που πρέπει να ληφθούν υπόψη, κυρίως η λειτουργική, αποδοτική και ομογενοποιημένη χρήση τους (Malone *et al.*, 2016).

#### 4.1 Υποστήριξη εικονικών υπολογιστικών περιβαλλόντων (virtualization)

Οι υπολογιστικές απαιτήσεις μιας BEPE είναι συχνά άγνωστες στον συγγραφέα ή χρήστη της. Επιπλέον, το υποκείμενο υπολογιστικό περιβάλλον έχει τις δικές του απαιτήσεις (π.χ. το λειτουργικό σύστημα, τις εγκατεστημένες βιβλιοθήκες και τα προεγκατεστημένα πακέτα). Επειδή ένα περιβάλλον BEPE έχει τις δικές του απαιτήσεις και εξαρτήσεις, είναι δυσκίνητο, ακόμη και για ειδικευμένους και καλά εκπαιδευμένους στην πληροφορική βιοπληροφορικούς. Συνεπώς, δαπανάται ένας σημαντικός όγκος ερευνητικού χρόνου για τη διαμόρφωση και τη δημιουργία ενός περιβάλλοντος BEPE. Επιπλέον, η έλλειψη τεκμηρίωσης, η απειρία σε δεξιότητες πληροφορικής και η πίεση του χρόνου οδηγούν σε λανθασμένα διαμορφωμένα περιβάλλοντα που με τη σειρά τους οδηγούν σε σπατάλη πόρων, ακόμη και σε εσφαλμένα αποτελέσματα. Μια λύση για αυτό μπορεί να είναι η **εικονικοποίηση** (virtualization) (Juve and Deelman, 2011). Η εικονικοποίηση είναι η "συσκευασία σε ένα πακέτο" όλων των απαιτούμενων συστατικών λογισμικών, λειτουργικού συστήματος, βιβλιοθηκών και πακέτων σε ένα μοναδικό αντικείμενο (συνήθως ένα αρχείο) που καλείται "εικόνα" (image) ή "container". Αυτή η εικόνα μπορεί να εκτελεστεί σχεδόν σε όλα τα γνωστά λειτουργικά συστήματα με τη βοήθεια συγκεκριμένου λογισμικού, καθιστώντας τη συγκεκριμένη τεχνική μια πολύ πρόσφορη προσέγγιση ενσωμάτωσης διαφορετικών και ετερογενών συστατικών από διαφορετικές ροές εργασίας. Οποιοδήποτε περιβάλλον σύνθεσης και εκτέλεσης ροών εργασίας που μπορεί να εικονικοποιηθεί είναι αυτόματα εύκολο να ενσωματωθεί σε οποιοδήποτε άλλο σύστημα. Ορισμένα ενδεικτικά παραδείγματα περιλαμβάνουν την κοινοπραξία *I2B2* (Informatics for Integrating Biology and the Bedside)<sup>39</sup>, η οποία προσφέρει την ολοκληρωμένη ροή εργασίας σε μια εικόνα VMare (Uzuner *et al.*, 2011). Ένα άλλο είναι το λογισμικό transMART που προσφέρεται σε ένα container VMware ή VirtualBox (Athey *et al.*, 2013). Το *Docker* ίσως το πιο διαδεδομένο περιβάλλον και υποδομή εικονοποίησης προσφέρει ένα ανοικτό αποθετήριο όπου οι χρήστες μπορούν να περιηγηθούν, να κατεβάσουν και να εκτελέσουν μια τεράστια συλλογή από containers. Η αξία του *Docker* στις σύγχρονες ερευνητικές διαδικασίες ανάλυσης δεδομένων έχει ήδη δειχθεί (Boettiger, 2015; Di Tommaso *et al.*, 2015). Για παράδειγμα, το BioSha-Dock είναι ένα αποθετήριο Docker εργαλείων που μπορούν να εκτελεστούν χωρίς προβλήματα στο Galaxy (Moreews *et al.*, 2015).

#### 4.2 Υποστήριξη Πρόσβασης σε Διαδεδομένες Πηγές Μεταγονιδιωματικής ... οι μπαταρίες στη συσκευασία!

**Υποστήριξη πρόσβασης και ανάκλησης δεδομένων.** Τα τελευταία χρόνια, ένας αυξανόμενος αριθμός μεγάλων ερευνητικών προγραμμάτων δημιούργησε μεγάλους όγκους ποιοτικών δεδομένων που αποτελούν μέρος σχεδόν όλων των αναπτυσσόμενων και διαθέσιμων BEPE. Μέχρι στιγμής, παρόλο που ο εντοπισμός και η συλλογή δεδομένων είναι σχετικά απλές διαδικασίες, ο όγκος τους και οι διασκορπισμένες περιγραφές τους καθιστούν δύσκολη την ενσωμάτωσή τους. Τα σύγχρονα περιβάλλοντα BEPE θα πρέπει να προσφέρουν αυτόματες μεθόδους συλλογής δεδομένων από γνωστές, αξιόπιστες και ευρέως διαδεδομένες πηγές βιο-δεδομένων, όπως για παράδειγμα: από την *ExAC*<sup>40</sup> (Lek *et al.*, 2016) η οποία διαθέτει ~60.000 πλήρως χαρακτηρισμένα εξώματα (exomes); από το *1000 Genomes Project*<sup>41</sup> (The 1000 Genomes Project Consortium, 2012) το οποίο διαθέτει ~2.500 πλήρη γονιδιώματα, το έργο *ENCODE*<sup>42</sup> (Consortium Project ENCODE, 2004); και το *TCGA*<sup>43</sup> (The Cancer Genome Atlas Program) (Tomczak *et al.*, 2015). Ένα βασικό εμπόδιο είναι ότι τα δεδομένα τα οποία διατίθενται από διεθνείς κοινοπραξίες στο πεδίο της μεταγονιδιωματικής έρευνας δημοσιεύονται με διαφορετικές μορφές συγκατάθεσης και αντίστοιχες πολιτικές πρόσβασης και ανάκλησης δεδομένων, όπως για παράδειγμα το Ευρωπαϊκό Αρχείο Γονοτύπου-

<sup>39</sup> [www.i2b2.org](http://www.i2b2.org)

<sup>40</sup> [exac.broadinstitute.org](http://exac.broadinstitute.org)

<sup>41</sup> [www.internationalgenome.org](http://www.internationalgenome.org)

<sup>42</sup> [www.encodeproject.org](http://www.encodeproject.org)

<sup>43</sup> [www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga](http://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga)

Φαινόματος<sup>44</sup> (EGA – European Genome-Phenome Archive). Ένα άλλο σημαντικό χαρακτηριστικό το οποίο πρέπει να χαρακτηρίζει ένα σύγχρονο και ολοκληρωμένο περιβάλλον σύνθεσης και εκτέλεσης BEPE είναι η αυτόματη παραγωγή τεχνητών δεδομένων, όποτε αυτό απαιτείται για σκοπούς δοκιμής και προσομοίωσης (Mu *et al.*, 2015).

**Υποστήριξη πρόσβασης και ενσωμάτωσης εργαλείων ανάλυσης και ροών εργασίας.** Η συμπερίληψη διαθέσιμων, επικυρωμένων και ευρέως διαδεδομένων εργαλείων ανάλυσης βιο-δεδομένων σε σύγχρονα περιβάλλοντα BEPE προσφέρει ένα πολύτιμο πλαίσιο ανάπτυξης ροών εργασίας το οποίο μπορεί εύκολα και γρήγορα να προσελκύσει χρήστες να συμμετάσχουν και να συμβάλλουν με τις υλοποιήσεις τους. Στο πεδίο της μεταγονιδιωματικής βιοϊατρικής έρευνας και βιοπληροφορικής τα σχετικά εργαλεία και ροές εργασίας μπορούν να χωριστούν στις ακόλουθες κατηγορίες:

- Εργαλεία τα οποία είναι απαραίτητα και μπορούν να χρησιμοποιηθούν για ανάλυση συσχετίσεων γονοτύπου-φαινοτύπου, όπως το plink<sup>45</sup> (Purcell *et al.*, 2007) και το GATK<sup>46</sup> (DePristo *et al.*, 2011).
- Εργαλεία επισημείωσης (annotation) γονιδιωμάτων και λειτουργικότητας, όπως για παράδειγμα τα: Avia<sup>47</sup> (Vuong *et al.*, 2015), ANNOVAR<sup>48</sup> (Wang *et al.*, 2010), GEMINI<sup>49</sup> (Paila *et al.*, 2013), SIFT<sup>50</sup> (Schneider *et al.*, 2012) και Polyphen-2 (Adzhubei *et al.*, 2013).
- Ροές εργασίας για ανάλυση δεδομένων αλληλούχισης νέας γενιάς (next generation sequencing / NGS), π.χ., iSPV (Mimori *et al.*, 2013); RNA-sequencing, π.χ., PRADA (Torres-Garcia *et al.*, 2014), ο προσδιορισμός νέων μεταλλάξεων (Samocha *et al.*, 2014); και imputation ελλειπόντων γονοτύπων (missing genotypes) (Kanterakis *et al.*, 2015).

**Υποστήριξη Οπτικοποίησης Δεδομένων και Αποτελεσμάτων.** Ένα χαρακτηριστικό που υποστηρίζεται εν μέρει από υπάρχοντα περιβάλλοντα BEPE είναι η εγγενής υποστήριξη **οπτικοποίησης** (visualization) δεδομένων. Στη μεταγονιδιωματική έρευνα, ορισμένες μέθοδοι απεικόνισης έχουν καταστεί πρότυπο και είναι εύκολα ερμηνευμένες και ευρέως κατανοητές από τις εμπλεκόμενες κοινότητες. Για παράδειγμα, σε μελέτες ανάλυσης σε επίπεδο γονιδιώματος (GWAS – genome wide association analysis), τα **διαγράμματα Manhattan** για την ανίχνευση της σημασίας συγκεκριμένων αποτελεσμάτων<sup>51</sup>, τα **διαγράμματα QQ** για ανίχνευση πληθωρισμού στα δεδομένα<sup>52</sup> καθώς και τα διαγράμματα **κύριας ανάλυσης συστατικών** (principal components) για την ανίχνευση της στρωματοποίησης του πληθυσμού αποτελούν βασικά συστατικά BEPE.

## 5 ΑΠΑΙΤΗΣΕΙΣ ΤΗΣ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ ΚΟΙΝΟΤΗΤΑΣ

### 5.1 Οι ανάγκες των κοινοτήτων Βιοπληροφορικής: Το ερωτηματολόγιο του OpenBio-C

Στα πλαίσια της ανάλυσης των απαιτήσεων των υποψήφιων χρηστών συντάχτηκε ένα ερωτηματολόγιο (Σχήμα 11). Σκοπός του ερωτηματολογίου ήταν να αξιολογηθούν οι ανάγκες της πλατφόρμας και να καταγραφούν τα κύρια εμπόδια που αντιμετωπίζουν οι ερευνητές στην αναπαραγωγή της έρευνας.

Οι ερωτήσεις του ερωτηματολογίου χωρίστηκαν στις παρακάτω κατηγορίες με βάση το περιεχόμενό τους:

- Προφίλ χρήστη (Profile)
- Υπολογιστικά περιβάλλοντα και εργαλεία Βιοπληροφορικής (Computers and Bioinformatics tools)
- Συστήματα διαχείρισης BEPE (Workflow Management Systems)
- Διεξαγωγή έρευνας στο πεδίο της βιοπληροφορικής (Carrying out bioinformatics research)

<sup>44</sup> [ega-archive.org](http://ega-archive.org)

<sup>45</sup> [zzz.bwh.harvard.edu/plink](http://zzz.bwh.harvard.edu/plink)

<sup>46</sup> [software.broadinstitute.org/gatk](http://software.broadinstitute.org/gatk)

<sup>47</sup> [avia-abcc.ncicrf.gov/apps/site/sub\\_analysis/?id=3](http://avia-abcc.ncicrf.gov/apps/site/sub_analysis/?id=3)

<sup>48</sup> [annovar.openbioinformatics.org/en/latest](http://annovar.openbioinformatics.org/en/latest)

<sup>49</sup> [gemini.readthedocs.io](http://gemini.readthedocs.io)

<sup>50</sup> [sift.bii.a-star.edu.sg](http://sift.bii.a-star.edu.sg)

<sup>51</sup> [en.wikipedia.org/wiki/Manhattan\\_plot](http://en.wikipedia.org/wiki/Manhattan_plot)

<sup>52</sup> [en.wikipedia.org/wiki/Q-Q\\_plot](http://en.wikipedia.org/wiki/Q-Q_plot)



Σχήμα 11. Η πρώτη σελίδα του ερωτηματολογίου του OpenBio-C

Το ερωτηματολόγιο διαδόθηκε διεθνώς και στάλθηκε σε ερευνητές στο πεδίο της μεταγονιδιωμικής έρευνας και βιοπληροφορικής καθώς και σε μεταπτυχιακούς φοιτητές στον τομέα της βιοπληροφορικής, καθώς και σε forum σχετικού ερευνητικού ενδιαφέροντος, π.χ.: Reddit/ bioinformatics, Reddit/ SampleSize, Biostars.org; Twitter/@openbioe.

- **Προφίλ (Profile, Σχήμα 12).** Οι πρώτες πέντε ερωτήσεις έχουν σαν στόχο την συγκέντρωση προσωπικών πληροφοριών του χρήστη που αφορούν την ηλικία, το εκπαιδευτικό επίπεδο, το περιβάλλον εργασίας και το λειτουργικό σύστημα με το οποίο είναι εξοικειωμένοι. Οι ερωτήσεις σε αυτήν την κατηγορία δεν είναι υποχρεωτικές καθώς αφορούν προσωπικά δεδομένα, αν και σχεδόν το 100% των συμμετεχόντων απάντησαν σε αυτή την ενότητα.

Σχήμα 12. Ερώτηση του ερωτηματολογίου στην κατηγορία «Προφίλ».



- **Υπολογιστές και εργαλεία βιοπληροφορικής** (Computers and Bioinformatics tools, Σχήμα 13). Οι επόμενες πέντε ερωτήσεις αφορούν τη χρήση και την εξοικείωση με εργαλεία βιοπληροφορικής. Ξεκινώντας με δύο πιο γενικές ερωτήσεις που αφορούν την εμπειρία στον προγραμματισμό και τη χρήση του λειτουργικού συστήματος Linux, ακολουθούν ερωτήσεις για το επίπεδο γνώσης και εξειδίκευσης στην βιοπληροφορική αλλά και της χρήσης ειδικών τεχνικών βιοπληροφορικής.

**Computers and Bioinformatics tools**

What is your experience with computer programming? \*

☐ No experience

☐ Novice / entry level

☐ Advanced

☐ Professional

Σχήμα 13. Ερώτηση του ερωτηματολογίου στην κατηγορία «Υπολογιστές και εργαλεία βιοπληροφορικής».

- **Συστήματα διαχείρισης ροής εργασίας** (Workflow Management Systems, Σχήμα 14). Οι ερωτήσεις στην ενότητα αυτή έχουν στόχο τη συλλογή πληροφοριών για τη χρήση υπάρχουσας τεχνολογίας και τις εργασίες για τις οποίες χρησιμοποιούνται συστήματα διαχείρισης ροής εργασίας. Επίσης, περιλαμβάνονται ερωτήσεις για την εμπειρία χρήσης ανάλογων συστημάτων.

How often do you use the following environments for your bioinformatics analysis? \*

|                                    | Very often            | Often                 | Occasionally          | Rarely                | Never                 |
|------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Personal computer                  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| High-End workstation               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Campus/Institution cluster         | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Nation/Region wide cluster or grid | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Cloud                              | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

BACK NEXT

Page 3 of 4

Σχήμα 14. Ερώτηση του ερωτηματολογίου στην κατηγορία «Συστήματα διαχείρισης ροής εργασίας».

- **Διεξαγωγή έρευνας στο πεδίο της βιοπληροφορικής** (Carrying out bioinformatics research, Σχήμα 15). Οι ερωτήσεις στην τελευταία ενότητα αναφέρονται στη διεξαγωγή/αναπαραγωγή έρευνας και στη χρήση εργαλείων για την έρευνα. Η τελευταία ερώτηση έχει ως στόχο τη αξιολόγηση των πιο σημαντικών χαρακτηριστικών για την δημιουργία μιας συνεργατικής πλατφόρμας.

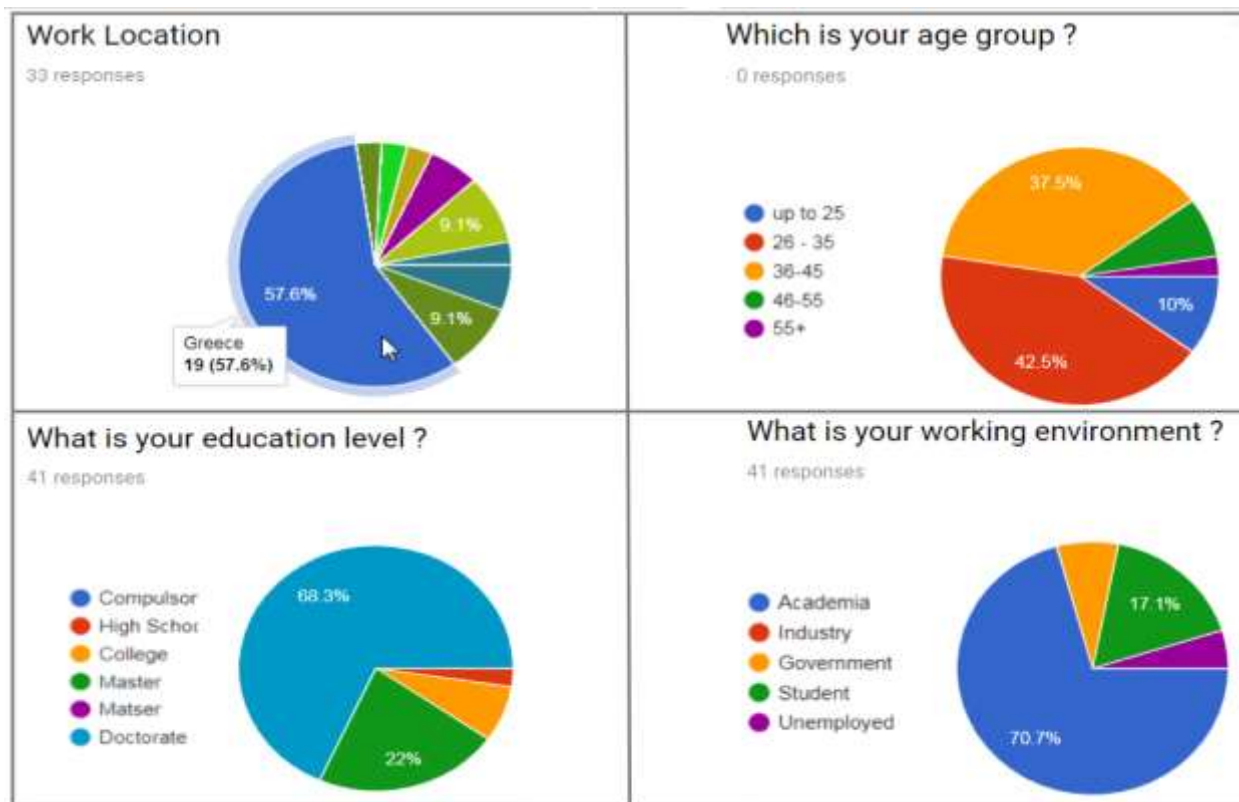
In average how straightforward is to reproduce the analysis of a research paper?

|           | 1                     | 2                     | 3                     | 4                     | 5                     |                |
|-----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|----------------|
| Very easy | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Very difficult |

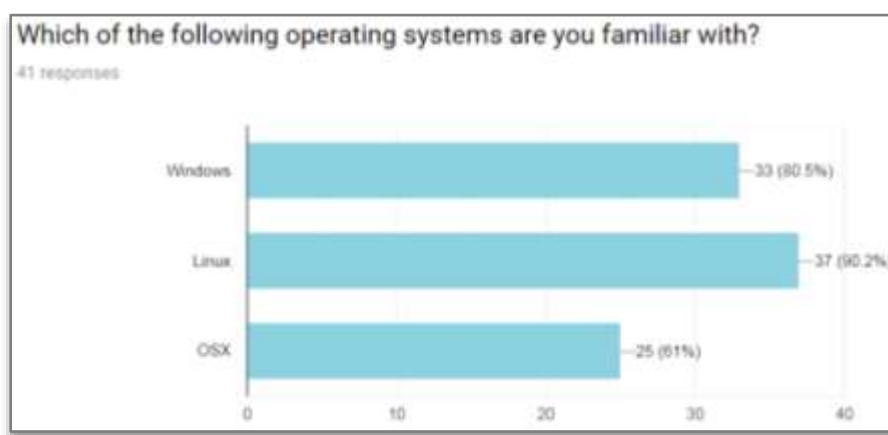
Σχήμα 15. Ερώτηση του ερωτηματολογίου στην κατηγορία «Διεξαγωγή έρευνας βιοπληροφορικής».

### 5.1.1 Ανάλυση απαντήσεων ερωτηματολογίου

**Προφίλ (Profile).** Οι απαντήσεις στην κατηγορία αυτή δείχνουν ότι η πλειοψηφία των συμμετεχόντων δουλεύουν στην Ελλάδα (57.6%), ανήκουν στο ηλικιακό γκρουπ 26-35 (42,5%) και ένα σημαντικό ποσοστό ανήκει στο ηλικιακό γκρουπ 36-45 (37,5%). Επίσης οι ερωτηθέντες, κατά κύριο λόγο, εργάζονται στον ακαδημαϊκό χώρο (70,7%) και είναι κάτοχοι διδακτορικού (68,3%) (Εικόνα 16). Τα λειτουργικά συστήματα με τα οποία είναι εξοικειωμένοι είναι: Linux (90,2%), Windows (80,5%) και OSX (61%), Σχήμα 17.

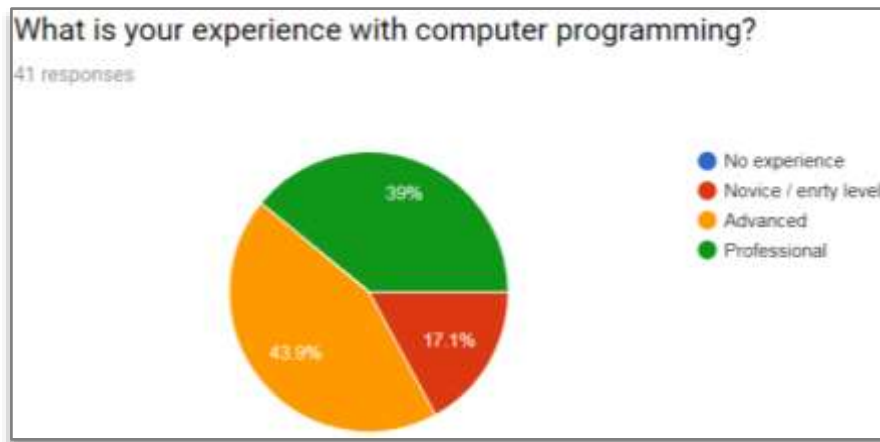


Εικόνα 16. Κατανομή απαντήσεων ανά τόπο/χώρα εργασίας, ηλικία, επίπεδο εκπαίδευσης και εργασιακό περιβάλλον



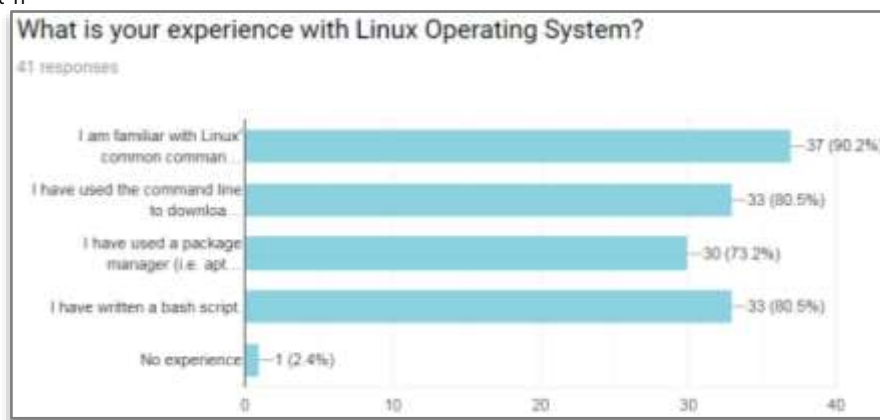
Σχήμα 17. Ερώτηση πολλαπλής επιλογής για την εξοικείωση με τα λειτουργικά συστήματα Windows, Linux, OSX ή και προσθήκη κάποιο άλλου.

**Υπολογιστές και εργαλεία βιοπληροφορικής (Computers and Bioinformatics tools, Σχήμα 18).** Με βάση τις απαντήσεις για τη χρήση υπολογιστών και εργαλείων πληροφορικής οι ερωτηθέντες είναι πολύ εξοικειωμένοι με τον προγραμματισμό (43,9 % προχωρημένοι και 39% επαγγελματίες).



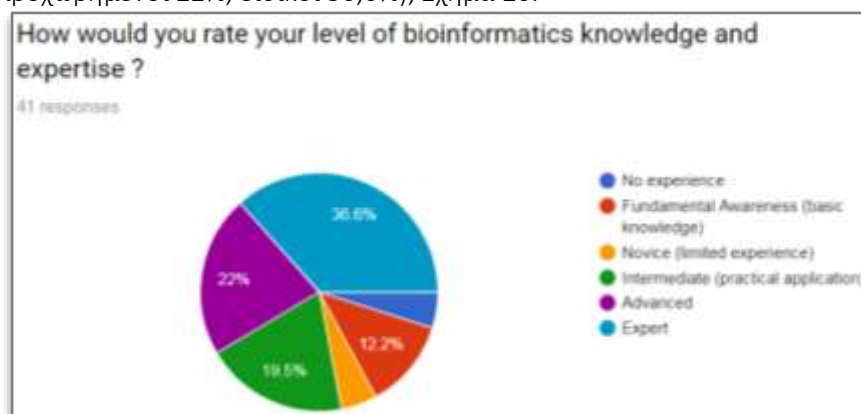
**Σχήμα 18.** Ερώτηση επιλογής για την εμπειρία (καμία εμπειρία, αρχάριος, προχωρημένος και επαγγελματίας) στον τομέα του προγραμματισμού.

Ένα πολύ μικρό ποσοστό δεν έχει εμπειρία με την χρήση Linux (2,4 %) ενώ το μεγαλύτερο ποσοστό έχει εμπειρία με βασικές εντολές Linux (90,2%) αλλά και με πιο εξειδικευμένες διεργασίες (80,5% έχει γράψει bash scripts), Σχήμα 19.



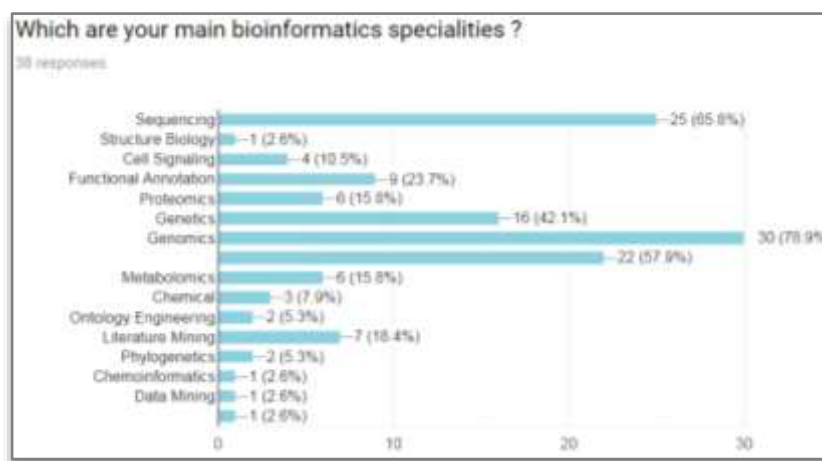
**Σχήμα 19.** Ερώτηση πολλαπλής επιλογής για την εμπειρία χρήσης του λειτουργικού συστήματος Linux.

Στην ερώτηση για το επίπεδο εμπειρίας στον τομέα της βιοπληροφορικής οι περισσότεροι χρήστες είναι προχωρημένοι (προχωρημένοι 22%, ειδικοί 36,6%), Σχήμα 20.



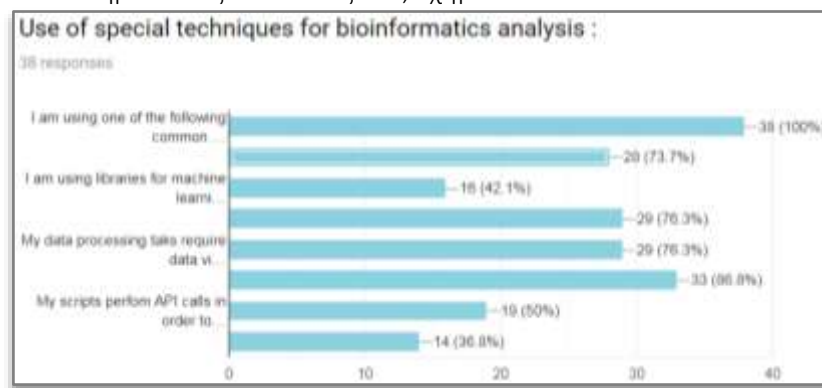
**Σχήμα 20.** Ερώτηση επιλογής για επίπεδο εμπειρίας (καμία εμπειρία, βασικές γνώσεις, αρχάριος, προχωρημένος, πρακτική εφαρμογή, ειδικός) στο τομέα της βιοπληροφορικής.

Οι ειδικότητες που συναντήσαμε πιο συχνά στον τομέα της βιοπληροφορικής είναι: Genomics (78,9%), Sequencing (65,8%) και Gene Expression (57,9%), Σχήμα 21.



**Σχήμα 21.** Προαιρετική ερώτηση πολλαπλής επιλογής για τις ειδικότητες (π.χ. sequencing, data mining) στον τομέα της βιοπληροφορικής. Ο χρήστης μπορεί να επιλέξει πάνω από μία ή και καμία ειδικότητα.

Για τη χρήση τεχνικών βιοπληροφορικής τα αποτελέσματα του ερωτηματολογίου δείχνουν ότι όλοι όσοι απάντησαν (100%) χρησιμοποιούν Python, R και Perl και ένα μεγάλο ποσοστό(86,8%) χρειάζεται πάνω από ένα εργαλεία για να ολοκληρώσει τις αναλύσεις του, Σχήμα 22.



**Σχήμα 22.** Προαιρετική ερώτηση πολλαπλής επιλογής η οποία αφορά την χρήση συγκεκριμένων τεχνικών βιοπληροφορικής.

**Συστήματα διαχείρισης ροής εργασίας (Workflow Management Systems).** Στην κατηγορία των ερωτήσεων για τα συστήματα διαχείρισης ροής εργασίας οι απαντήσεις των χρηστών έδειξαν ότι ένα μεγάλο ποσοστό δεν τα χρησιμοποιεί (48,8%). Ανάμεσα σε αυτούς που χρησιμοποιούν τέτοια συστήματα οι προτιμήσεις κυμαίνονται στο Galaxy (17,1%), Nextflow(17,1%) και Snakemake(14,6%), Σχήμα 23.



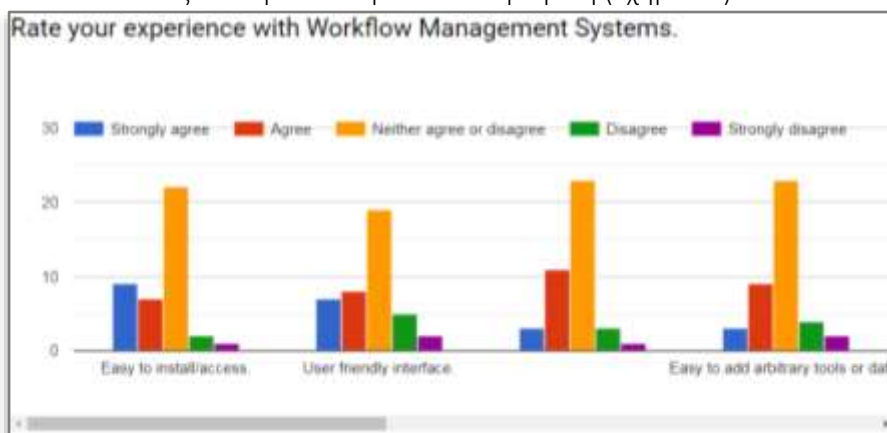
**Σχήμα 23.** Υποχρεωτική ερώτηση πολλαπλής επιλογής όπου ο χρήστης καλείται αν επιλέξει (ή και να συμπληρώσει) ποιο από τα υπάρχοντα συστήματα διαχείρισης ροής εργασιών (π.χ. “Galaxy”, “Taverna”) χρησιμοποιεί.

Οι χρήστες οι οποίοι χρησιμοποιούν συστήματα διαχείρισης ροής δεδομένων, τα θεωρούν χρήσιμα κυρίως για την δημιουργία αγωγών ανάλυσης (analysis pipeline), Σχήμα 24.



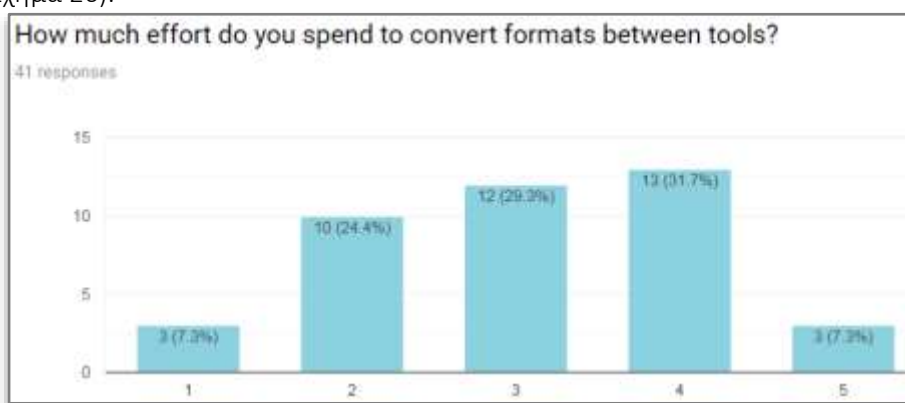
**Σχήμα 24.** Υποχρεωτική ερώτηση που έχει στόχο να καταγράψει τη συχνότητα χρήσης συστημάτων διαχείρισης για συγκεκριμένες εργασίες (π.χ. για την εγκατάσταση και χρήση κάποιο εργαλείου βιοπληροφορικής).

Η εμπειρία των χρηστών των συστημάτων διαχείρισης ροής εργασιών φαίνεται από τις παρακάτω απαντήσεις ότι δεν είναι ούτε ξεκάθαρα θετική ούτε και αρνητική (Σχήμα 25).



**Σχήμα 25.** Υποχρεωτική ερώτηση που έχει στόχο να καταγράψει την εμπειρία χρήσης που αφορά συγκεκριμένες λειτουργίες των συστημάτων διαχείρισης ροής εργασιών.

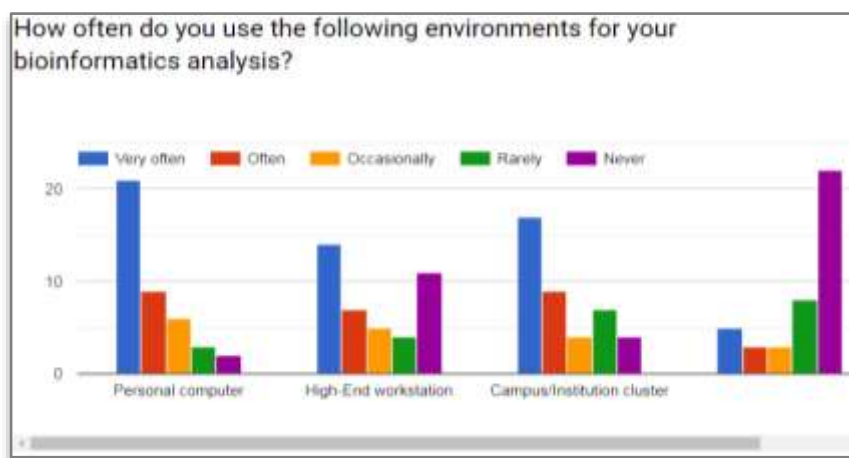
Από τις απαντήσεις στην ερώτηση για την προσπάθεια που χρειάζεται να καταβάλουν οι χρήστες για την μετατροπή τύπων μεταξύ εργαλείων διαπιστώσαμε ότι η πλειοψηφία θεωρεί ότι χρειάζεται αρκετή προσπάθεια (Σχήμα 26).



**Σχήμα 26.** Υποχρεωτική ερώτηση με βαθμολόγηση (1: very low - 5: very high) για την προσπάθεια που χρειάζεται για την μετατροπή τύπων μεταξύ εργαλείων.

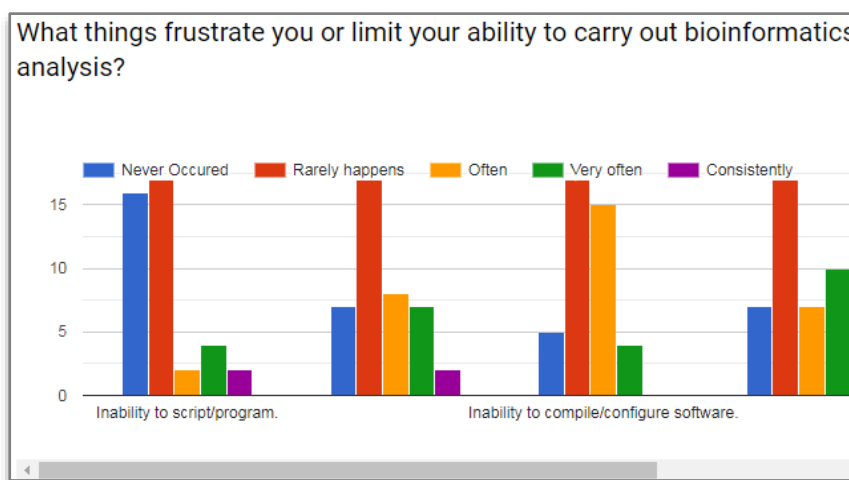
Για τις αναλύσεις (βιοπληροφορικής) χρησιμοποιείται κυρίως ο προσωπικός υπολογιστής με βάση τις απαντήσεις στην παρακάτω ερώτηση (Σχήμα 27):





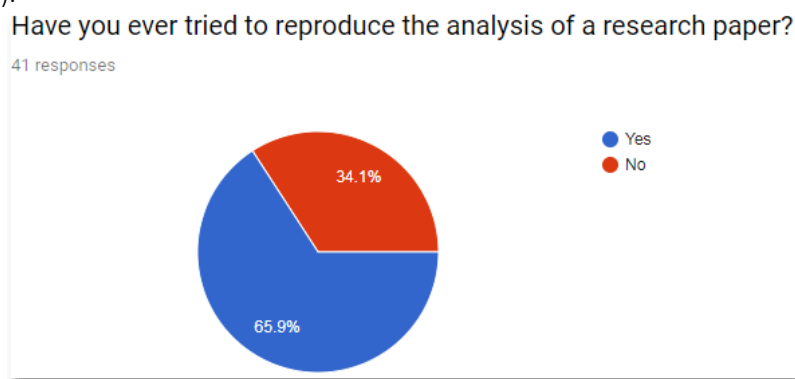
Σχήμα 27. Υποχρεωτική ερώτηση που έχει στόχο να καταγράψει τη συχνότητα χρήσης διαφορετικών περιβαλλόντων (π.χ. προσωπικός υπολογιστής, cloud) για αναλύσεις βιοπληροφορικής.

**Διεξαγωγή έρευνας βιοπληροφορικής** (Carrying out bioinformatics research). Σύμφωνα με τις απαντήσεις που λάβαμε το πιο συχνό πρόβλημα κατά την διεξαγωγή βιοπληροφορικών αναλύσεων είναι η έλλειψη τεκμηρίωσης. Ένα άλλο σημαντικό πρόβλημα φαίνεται ότι είναι τα προσαρμοσμένα δεδομένα (custom data), Σχήμα 28.



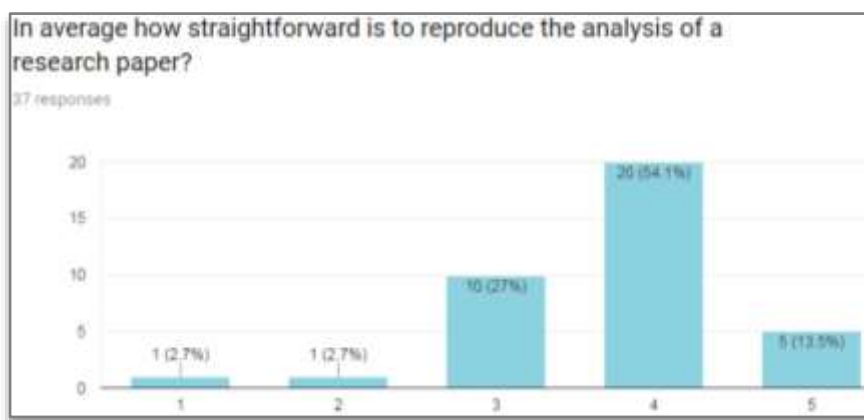
Σχήμα 28. Μη υποχρεωτική ερώτηση που έχει στόχο την καταμέτρηση της συχνότητας εμφάνισης προβληματικών κατά την διεξαγωγή βιοπληροφορικών αναλύσεων

Οι περισσότεροι από τους συμμετέχοντες (65,9%) έχουν προσπαθήσει να αναπαράγουν ερευνητικές εργασίες (Σχήμα 29).



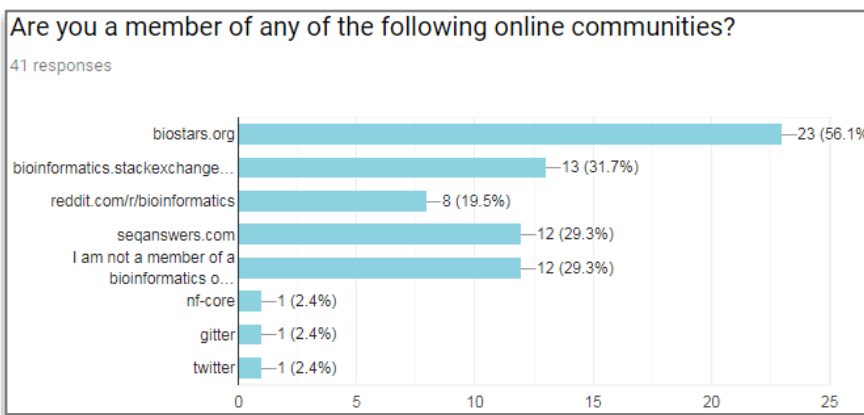
Σχήμα 29. Ερώτηση για την καταμέτρηση των συμμετεχόντων που έχουν δοκιμάσει να αναπαράγουν μια ερευνητική εργασία.

Από τους συμμετέχοντες που έχουν δοκιμάσει να αναπαράγουν κάποια ερευνητική εργασία οι περισσότεροι (54,1%) πιστεύουν ότι είναι μια δύσκολη διαδικασία (Σχήμα 30).



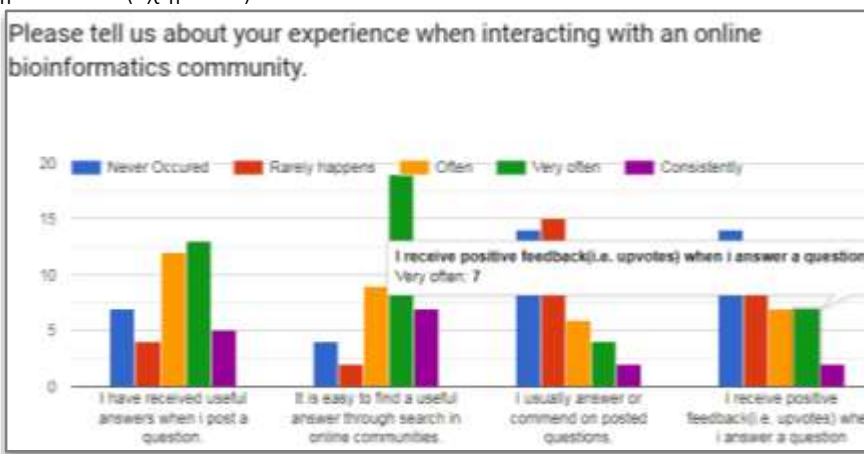
**Σχήμα 30.** Ερώτηση με στόχο να βαθμολογηθεί η ευκολία/δυσκολία της αναπαραγωγής μιας ερευνητικής εργασίας.

Το 56, 1% απάντησαν ότι είναι μέλη της διαδικτυακής κοινότητας biostars.org (Σχήμα 31).



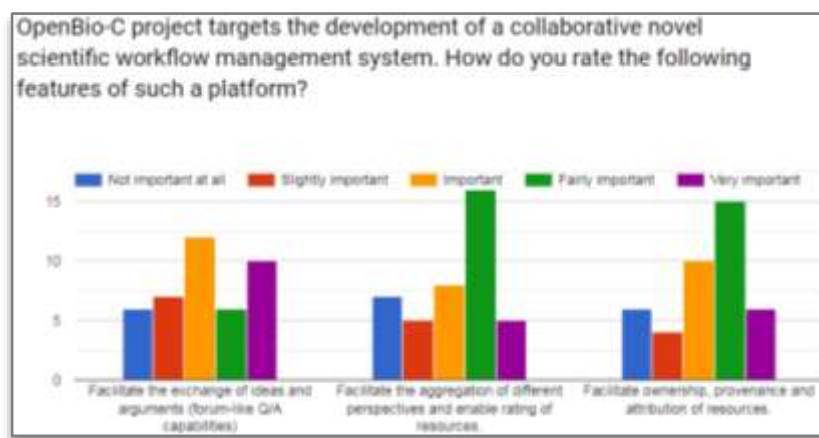
**Σχήμα 31.** Ερώτηση για την καταγραφή των συμμετεχόντων σε διαδικτυακές κοινότητες.

Μέσα από τη συμμετοχή σε διαδικτυακές κοινότητες ένα μεγάλο ποσοστό των χρηστών δείχνει ότι έχει βρει συχνά κάποια χρήσιμη απάντηση σε υπάρχουσες συζητήσεις ή και έχει πάρει κάποια χρήσιμη απάντηση σε ερώτηση που δημοσίευσε (Σχήμα 32).



**Σχήμα 32.** Μη υποχρεωτική ερώτηση που έχει στόχο την καταμέτρηση της συχνότητας εμφάνισης συγκεκριμένων κατά την συμμετοχή σε διαδικτυακές κοινότητες βιοπληροφορικής.

Το πιο χρήσιμο στοιχείο που προτείνεται μέσα από το ερωτηματολόγιο για την πλατφόρμα OpenBio-C είναι η διευκόλυνση της ανταλλαγής ιδεών και επιχειρημάτων (φόρουμ τύπου Q / A ικανότητες). Πολύ σημαντικά φαίνεται ότι είναι, με ελάχιστη διαφορά, η συνάθροιση διαφορετικών προοπτικών και η αξιολόγηση πόρων καθώς και η διευκόλυνση της κυριότητας της προέλευσης και της κατανομής πόρων (Σχήμα 33).



Σχήμα 33. Ερώτηση για την βαθμολόγηση σπουδαιότητας προτεινόμενων εργασιών για την πλατφόρμα του OpenBio-C.

### 5.1.2 Ανάλυση Απαιτήσεων: Γενικά συμπεράσματα από την ανάλυση του ερωτηματολογίου

Από τις απαντήσεις που ελήφθησαν μπορούμε να συμπεράνουμε τις ακόλουθες τάσεις:

- Το γενικότερο μορφωτικό επίπεδο των χρηστών είναι πολύ ψηλό. Αυτό σημαίνει ότι πρέπει να αναμένουμε ότι οι χρήστες έχουν εξειδικευμένες γνώσεις βιολογίας και γενετικής. Άρα θα πρέπει να υποστηρίξουμε αρκετά πολύπλοκες δομές BEPE και μεγάλη ετερογένεια στα εργαλεία.
- Οι περισσότεροι χρήστες είναι εξοικειωμένοι με δημοφιλή εργαλεία προγραμματισμού και στατιστικής επεξεργασίας όπως είναι η Python η R και η Perl.
- Οι περισσότεροι συμμετέχοντες είναι εξοικειωμένοι με το λειτουργικό σύστημα Linux.
- Η πλειοψηφία των χρηστών, παρά την εξοικείωση με τα εργαλεία που αναφέρθηκαν και τις διαδικασίες που ακολουθούν, δε χρησιμοποιεί περιβάλλοντα για BEPE.
- Παρατηρείται εξοικείωση με τη χρήση και συμμετοχή σε διαδικτυακές κοινότητες(π.χ. forum) και γενικότερα τάση προς δικτύωση, διάδοση πληροφοριών και εξωστρέφεια.
- Οι περισσότεροι χρήστες αναπαράγουν ερευνητικές εργασίες όπου και απάντησαν ότι είναι μία δύσκολη διαδικασία.

## 6 ΠΡΩΤΟΤΥΠΗ ΑΡΧΙΚΗ ΥΛΟΠΟΙΗΣΗ ΤΟΥ OPENBIO-C

### 6.1 Το βασικό πλαίσιο και υπόβαθρο της υλοποίησης

Το OpenBio-C είναι μια διαδικτυακή πλατφόρμα μέσω της οποίας οι χρήστες *εισάγουν, επεξεργάζονται, επικυρώνουν, αξιολογούν και σχολιάζουν* ψηφιακά **ερευνητικά αντικείμενα**<sup>53</sup> (Research Objects). Τα ερευνητικά αντικείμενα αποτελούν μία σύνθετη δομή δεδομένων μέσω της οποίας επιτυγχάνεται η πλούσια **σημασιολογική περιγραφή και επισημείωση** (semantic description and annotation) όλων των συστατικών ρών εργασιών (εργαλείων, δεδομένων καθώς και άλλων ρών); των ανθρώπων/ερευνητών και ομάδων που εμπλέκονται στη σύνθεση και εκτέλεση ρών εργασίας και εν-γένει εκτελούν μεταγονιδιωματικές μελέτες και πειράματα με χρήση μεθοδολογιών βιοπληροφορικής; αλλά και στις συνεργατικές διαδικασίες οι οποίες λαμβάνουν χώρα για τη σχεδίαση και εκτέλεση τέτοιων μελετών/πειραμάτων και για τη παραγωγή και διερμήνευση των αντίστοιχων αποτελεσμάτων (Hettne *et al.*, 2014).

Εφαρμόζοντας το μοντέλο των ερευνητικών αντικειμένων για τη μοντελοποίηση, αναπαράσταση και διαχείριση ολοκληρωμένων βιοπληροφορικών διαδικασιών μπορούμε να ανακτήσουμε και να διαχειριστούμε τα συμπεράσματα του πειράματος στο πλαίσιο της αρχικής ερευνητικής υπόθεσης, των

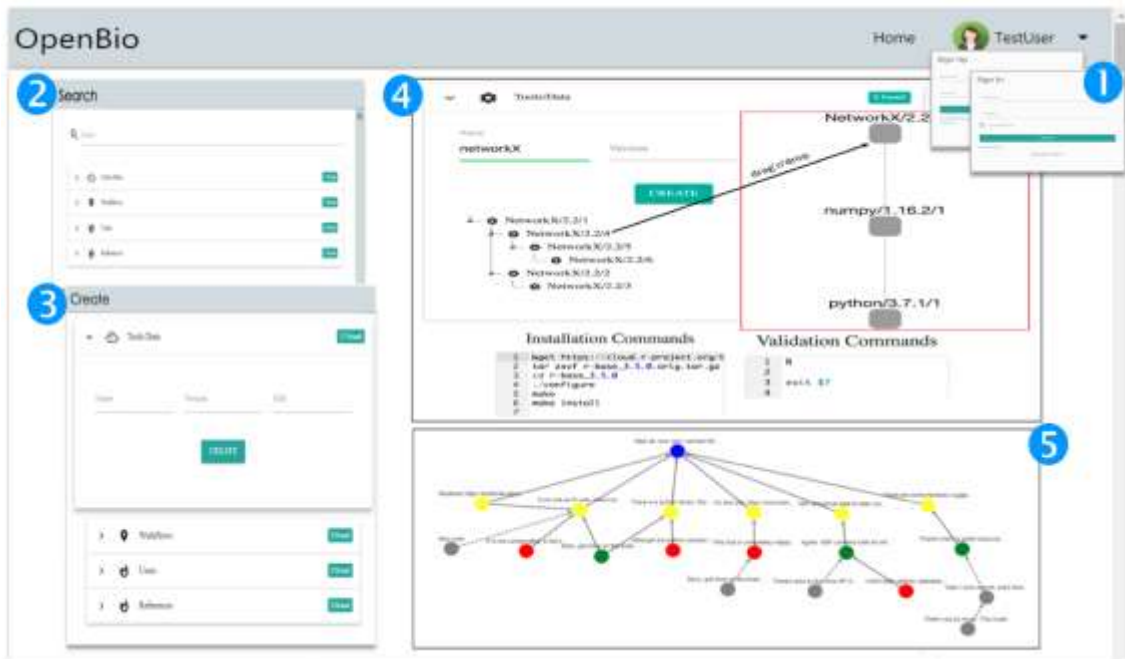
<sup>53</sup> <http://www.researchobject.org/>

εκτελούμενων ρών εργασίας καθώς και των εμπλεκόμενων πηγών δεδομένων και γνώσεων ... από την υπόθεση στο πείραμα και στη διερμηνύση αποτελεσμάτων! Ταυτόχρονα, το όλο περιβάλλον καθίσταται συμβατό με τις σύγχρονες αρχές FAIR (Findable, Accessible, Interoperable and Reusable) δυνάμενο να προσφέρει υπηρεσίες 'ανοιχτής' επιστήμης (Wilkinson *et al.*, 2016).

## 6.2 Βασικά συστατικά του OpenBio-C

Η αρχική σχεδίαση και υλοποίηση του OpenBio-C περιλαμβάνει 5 συστατικά υποσυστήματα (βλ. Σχήμα 1):

1 **Εγγραφής / (δημιουργίας) Προφίλ** [Sign Up] και **Σύνδεσης** [Sign In] του χρήστη 2 **Αναζήτησης/Εντοπισμού** [Search] ερευνητικών αντικειμένων (εργαλείων, ρών εργασίας, δεδομένων, ανθρώπων/ομάδων, αναφορών) 3 **Δημιουργίας** [Create] και εισαγωγής ερευνητικών αντικειμένων 4 **Σύνθεσης** ρών εργασίας με την **Εισαγωγή** (drag-and-drop) δημιουργημένων και εισηγμένων ερευνητικών αντικειμένων και την υποστήριξη Γραφικού περιβάλλοντος 5 **Περιβάλλον Συνεργασίας** και **Επιχειρηματολόγησης** με την υποστήριξη γραφικού περιβάλλοντος. Οι επί μέρους μονάδες (modules) οι οποίες υλοποιούν τα υποσυστήματα αυτά περιγράφονται με λεπτομέρεια στα επόμενα υπο-κεφάλαια.



Εικόνα 34. Τα βασικά λειτουργικά υποσυστήματα του OpenBio-C

## 6.3 Οι βασικές μονάδες υλοποίησης του OpenBio-C

Όλη η υλοποίηση του OpenBio-C βασίζεται σε ανοιχτό κώδικα. Οι βασικές μονάδες (modules) της διαδικτυακής πλατφόρμας περιγράφονται παρακάτω (βλ. Σχήμα 35):

1. **Η μονάδα υποστήριξης (backend).** Η μονάδα υποστήριξης, είναι το λειτουργικό κομμάτι το οποίο 'τρέχει' σε αποκλειστικά αφιερωμένους εξυπηρετητές και υποστηρίζει τις ακόλουθες λειτουργίες / υπηρεσίες: (i) την εγγραφή, διαπίστευση και καταγραφή (εσωτερικά στη πλατφόρμα) και των χρηστών, (ii) την επικοινωνία με τη βάση δεδομένων, (iii) την επικοινωνία με το γραφικό περιβάλλον σύνθεσης, αξιολόγησης, επικύρωσης και εκτέλεσης ρών εργασίας και (iv) τον γενικότερο έλεγχο όλων των πράξεων των χρηστών ώστε να διατηρηθεί η ακεραιότητα και η ορθότητα των δεδομένων. Για αυτό το υποσύστημα έχουμε επιλέξει το περιβάλλον ανάπτυξης ιστοτόπων Django<sup>54</sup>.
2. **Βάση Δεδομένων/Πληροφοριών.** Η βάση δεδομένων περιέχει τα στοιχεία των χρηστών και των ερευνητικών δεδομένων. Περιέχει επίσης στατιστικά στοιχεία χρήσης για κάθε ερευνητικό αντικείμενο. Για τα εργαλεία περιέχει: (α) Σε πόσες ροές εργασίες συμμετέχουν, (β) πόσοι χρήστες τα έχουν χρησιμοποιήσει, (γ) ο μέσος όρος χρόνου/μήνης/αποθηκευτικού χώρου που χρειάζονται για να

<sup>54</sup> <https://www.djangoproject.com/>



εγκατασταθούν και να τρέξουν και (δ) τη βαθμολογία και τα σχόλια των χρηστών. Για τους χρήστες περιέχει: Πόσα και ποια ερευνητικά αντικείμενα έχουν δημιουργήσει, τα σχόλια και τη βαθμολογία που έχουν δώσει σε άλλα ερευνητικά αντικείμενα καθώς και εκτενής πληροφορία σχετικά με το ακαδημαϊκό τους προφίλ. Η βάση δεδομένων που έχει επιλεγεί είναι η **PostgreSQL**.

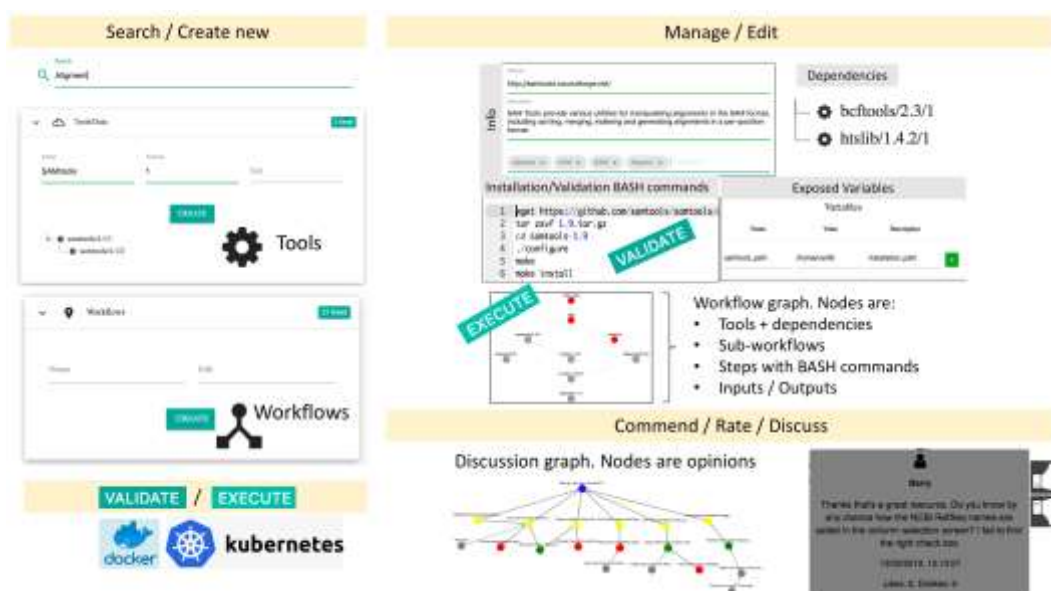
3. **Μονάδες διεπαφών (frontend).** Η μονάδες διεπαφών υλοποιούν όλη τη “λογική” της διαδικτυακής εφαρμογής OpenBio-C και του αντίστοιχου ιστοτόπου, επικυρώνει τις πράξεις των χρηστών και όταν επιτρέπεται, επικοινωνεί με το σύστημα υποστήριξης. Οι μονάδες διεπαφών ενσωματώνουν και αναφέρονται στις παρακάτω επί μέρους υπο-μονάδες:

■ **Κεντρική διεπαφή.** Η κεντρική διεπαφή περιέχει τον σχεδιασμό όλων των στοιχείων. Είναι υλοποιημένη σε Materialize το οποίο είναι ένα σύγχρονο σύστημα διαχείρισης των οπτικών στοιχείων της διεπαφής (χρώματα, στυλ, περιθώρια κτλ). Επίσης περιέχει οπτικά βοηθήματα ώστε ο χρήστης να οδηγείται στο τι πρέπει να κάνει, ποια πεδία να συμπληρώσει και με τι τύπου τιμές. Το περιβάλλον επίσης παρέχει εναλλακτικές διεπαφές για συσκευές σύνδεσης στο Ίντερνετ (π.χ., κινητά τηλέφωνα και tablets).

■ **Γραφικό περιβάλλον διαχείρισης ροών εργασιών.** Δεδομένου ότι οι ροές εργασίες είναι στην ουσία γράφοι, ο φυσικός τρόπος αναπαράστασής τους είναι σε ένα γραφικό περιβάλλον όπου οι χρήστες μπορούν να αλληλοεπιδρούν με τους κόμβους και τις ακμές του. Για το σκοπό αυτό έχουμε επιλέξει το **Cytoscape** (Franz *et al.*, 2016) το οποίο είναι μία javascript βιβλιοθήκη δημιουργίας, αναπαράστασης και επεξεργασίας γράφων. Αν και αρχικά το Cytoscape δημιουργήθηκε κυρίως για βιολογικούς γράφους, σήμερα χρησιμοποιείται σε όλους τους υπολογιστικούς τομείς όπου απαιτείται απεικόνιση γράφων. Στο OpenBio-C ο χρήστης μπορεί να εισάγει ένα εργαλείο (μαζί με τις εξαρτήσεις του) με μία απλή κίνηση μεταφοράς και απόθεσης (drag-and-drop). Ομοίως μπορεί να επεξεργαστεί όλα τα σημασιολογικά στοιχεία του γράφου όπως τα επιμέρους υπολογιστικά βήματα, τις παραμέτρους και τα τελικά αποτελέσματα.

4. **Μονάδα συνεργασίας.** Το συνεργατικό περιβάλλον που υλοποιείται στα πλαίσια του OpenBio-C προσφέρει λειτουργικότητα **Επιχειρηματολόγησης** (argumentation) υλοποιώντας σχετικό **γραφικό/δενδρικό περιβάλλον** και υποστηρίζει τα εξής βασικά χαρακτηριστικά: (α) Εισαγωγή/Επεξεργασία/Διαγραφή κόμβων στον γράφο της συζήτησης; (β) Επεξεργασία του κειμένου ενός κόμβου του γράφου συζήτησης; (γ) Παροχή σύντομων πληροφοριών μέσω ειδικού tooltip.

Όλα τα παραπάνω συντονίζονται μέσω της javascript βιβλιοθήκης **Angular**<sup>55</sup>. Μια γενική εικόνα της αρχιτεκτονικής της αρχικής υλοποίησης του OpenBio-C φαίνεται στο Σχήμα 35.



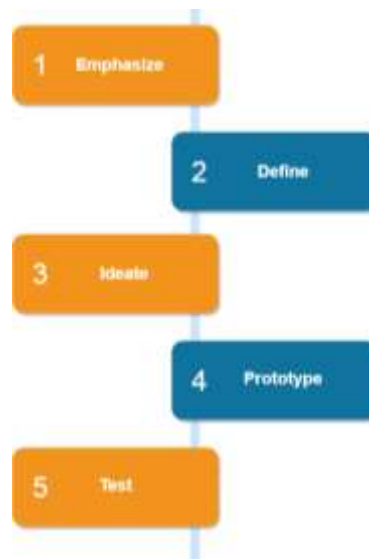
Σχήμα 35. Τα βασικά λειτουργικά συστήματα και αρχιτεκτονική του OpenBio-C.

<sup>55</sup> <https://angular.io/guide/creating-libraries>

## 6.4 Διεπαφή Χρήστη - OpenBio-C

Η πλατφόρμα OpenBio-C επιτρέπει την πρόσβαση σε ολοκληρωμένους και διαλειτουργικούς (interoperable) πόρους διαφορετικών τύπων και ενσωματώνει και διαχειρίζεται ετερογενείς πηγές δεδομένων. Οι χρήστες μπορούν να εισάγουν, επεξεργάζονται, αξιολογούν και σχολιάζουν ψηφιακά ερευνητικά αντικείμενα (research objects). Για την ομαλή υποστήριξη των πολυμορφικών και διαφορετικών λειτουργιών, και με βασικό στόχο την δημιουργία ενός περιβάλλοντος φιλικού προς το χρήστη ένα σημαντικό κομμάτι της υλοποίησης αφορά το σχεδιασμό της διεπαφής (interface) του ιστοτόπου του OpenBio-C.

Ο σχεδιασμός μιας διεπαφής χρήστη (user interface) με προσανατολισμό στις ανθρώπινες ανάγκες αποτελεί βασικό αντικείμενο έρευνας του επιστημονικού πεδίου της αλληλεπίδρασης ανθρώπου - υπολογιστή (HCI- Human Computer Interaction). Η αύξηση της πολυπλοκότητας και της πολυμορφίας των διεπαφών χρήστη τα τελευταία χρόνια έχει οδηγήσει στην επέκταση του πεδίου της αλληλεπίδρασης ανθρώπου-υπολογιστή και στην ανάπτυξη νέων μεθοδολογιών. Σε αυτό το πλαίσιο έχει αναπτυχθεί η διαδικασία βελτίωσης της ικανοποίησης των χρηστών με ένα προϊόν (διεπαφή) που συναντάται με τον όρο σχεδιασμός εμπειρίας χρήστη (UX - user experience design<sup>56</sup>). Η μεθοδολογία που προτείνεται σε αυτή την διαδικασία αποτελείται από πέντε βήματα και περιγράφεται στο παρακάτω Σχήμα 36:

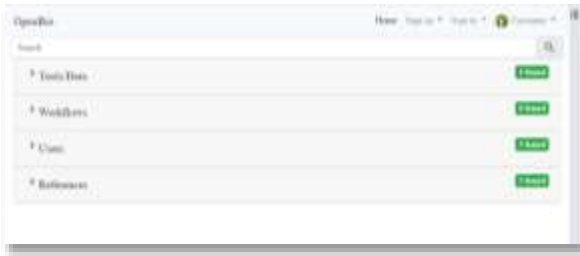


Σχήμα 36. Τα πέντε στάδια της μελέτης του σχεδιασμού (βλ. Stanford's Hasso-Plattner Institute of Design<sup>57</sup>).

Το πρώτο βήμα (**Emphasize**) έχει στόχο την κατανόηση του προβλήματος στο οποίο προσπαθούμε να δώσουμε λύση. Σε αυτό το στάδιο απαιτείται επικοινωνία με χρήστες για την καλύτερη κατανόηση του προβλήματος. Στο δεύτερο βήμα (**Define**) θα πρέπει να αναλύσουμε τα δεδομένα από το πρώτο βήμα και να καθορίσουμε τα βασικά προβλήματα τα οποία έχουμε εντοπίσει. Μετά από την ολοκλήρωση των διαδικασιών αυτών είμαστε έτοιμοι για το τρίτο βήμα (**Ideate**) όπου μπορούμε να δημιουργήσουμε ιδέες να εντοπίσουμε λύσεις με βάση τις πληροφορίες που έχουμε. Για την ολοκλήρωση αυτών των τριών βημάτων, όπως ήδη έχουμε περιγράψει, συγκεντρώσαμε στοιχεία που αφορούν τις υπάρχουσες ανάγκες, τις τεχνολογίες αιχμής στον τομέα και τα προβλήματα που παραμένουν. Υπό αυτό το πρίσμα, το τέταρτο βήμα της μεθοδολογίας (**Prototype**) και έχοντας τις προαπαιτούμενες πληροφορίες συνεχίσαμε με τον σχεδιασμό δύο διαφορετικών δειγμάτων διεπαφών, όπου χρησιμοποιήθηκαν οι εργαλειακές **Bootstrap** (Σχήμα 37) και **Materialize** (Σχήμα 38) αντίστοιχα.

<sup>56</sup> [careerfoundry.com/en/blog/ux-design/the-difference-between-ux-and-ui-design-a-laymans-guide](https://careerfoundry.com/en/blog/ux-design/the-difference-between-ux-and-ui-design-a-laymans-guide)

<sup>57</sup> [www.interaction-design.org/literature/article/5-stages-in-the-design-thinking-process](https://www.interaction-design.org/literature/article/5-stages-in-the-design-thinking-process)



**Σχήμα 37.** Υλοποίηση πρωτοτύπου με τη χρήση του Bootstrap.



**Σχήμα 38.** Υλοποίηση πρωτοτύπου με τη χρήση του Materialize.

Η επιλογή της τελικής διεπαφής (**Βήμα 5 - Test**) ώστε να προχωρήσουμε στην παραμετροποίηση του και ενσωμάτωση με την πλατφόρμα έγινε από την ομάδα του OpenBio-C μετά από δοκιμές χρήσης και των δύο προτύπων. Για την επιλογή και την περαιτέρω ανάπτυξη των διεπαφών της αρχικής διαδικτυακής πλατφόρμας OpenBio-C βασιστήκαμε σε κάποιες βασικές αρχές σχεδιασμού διεπαφής χρήστη<sup>58,59,60</sup>.

- Έλεγχος της διεπαφής από τον χρήστη – αυτό επιτυγχάνεται με τη δυνατότητα αναίρεσης ενεργειών από το χρήστη, με την εύκολη πλοήγηση στην διεπαφή, την παροχή ενημερωτικών σχολίων και ορατότητα κατάστασης συστήματος (Σχήμα 39).



**Σχήμα 39.** Ενημερωτικά μηνύματα από τον ισότοπο του OpenBio-C με στόχο την ενημέρωση και καθοδήγηση του χρήστη.



- *Εύκολη αλληλεπίδραση με την διεπαφή* – για την ευκολότερη αλληλεπίδραση του χρήστη με την διεπαφή του OpenBio-C είναι προτιμότερο σε κάθε βήμα να είναι ορατές μόνο οι απαραίτητες πληροφορίες. ➡ Χρησιμοποιώντας *μενού* – *ακορντεόν* (accordion) (Σχήμα 40), ο χρήστης μπορεί να επιλέγει τα δεδομένα τα οποία θα είναι ορατά και το πλήθος των υπόλοιπων στοιχείων να είναι πρόσκαιρα ανενεργά.

**Σχήμα 40.** Μενού-Ακορντεόν στο περιβάλλον αναζήτησης εργαλείων.

<sup>58</sup> 10 Usability Heuristics for User Interface Design. [www.nngroup.com/articles/ten-usability-heuristics](http://www.nngroup.com/articles/ten-usability-heuristics)

<sup>59</sup> *The Eight Golden Rules of Interface Design*. [www.cs.umd.edu/users/ben/goldenrules.html](http://www.cs.umd.edu/users/ben/goldenrules.html)

<sup>60</sup> *The 4 Golden Rules of UI Design.* [theblog.adobe.com/4-golden-rules-ui-design](http://theblog.adobe.com/4-golden-rules-ui-design)

- *Συνέπεια στο σχεδιασμό της διεπαφής* – η λειτουργική συνέπεια και η συνέπεια στην μορφή των στοιχείων που απαρτίζουν τη διεπαφή διευκολύνουν το χρήστη στην εκμάθηση των διαθέσιμων λειτουργιών. ➔ Η διαδικτυακή πλατφόρμα του OpenBio-C απαρτίζεται από διαφορετικά δομικά στοιχεία, που όμως ανήκουν σε συγκεκριμένες κατηγορίες π.χ. φόρμες συμπλήρωσης στοιχείων, μενού. Βασικό χαρακτηριστικό του σχεδιασμού του OpenBio-C είναι η συνέπεια μεταξύ των στοιχείων της ίδιας κατηγορίας αλλά και όλων των στοιχείων όσον αφορά τα χρώματα, τις γραμματοσειρές και άλλα κοινά χαρακτηριστικά.
- *Μείωση γνωστικού φορτίου* – αναφέρεται στην επεξεργασία που απαιτείται από το χρήστη για την χρήση της πλατφόρμας OpenBio-C. Ο σχεδιασμός θα πρέπει να προωθεί την αναγνώριση των πληροφοριών και λειτουργιών. Εικόνες όπως η δισκέτα παραπέμπουν στην λειτουργικότητα του κουμπιού και διευκολύνουν το χρήστη να αναγνωρίζει τη χρησιμότητά του εύκολα χωρίς να χρειάζεται να ανακαλεί αυτήν την πληροφορία. Επεξηγηματικά κείμενα που εμφανίζονται κάθε φορά όταν ο χρήστης τοποθετήσει τον κέρσορα στο εκάστοτε στοιχείο διευκολύνουν τη χρήση κατά τον ίδιο τρόπο (Σχήμα 41).



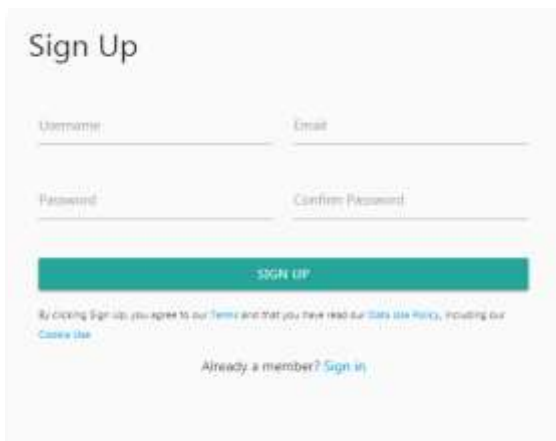
Σχήμα 41. Μείωση γνωστικού φορτίου με τη χρήση tooltips

#### 6.4.1 Εγγραφή & Προφίλ χρήστη

Για την δυνατότητα πλήρους πρόσβασης στις διαφορετικές λειτουργικότητες της πλατφόρμας οι χρήστες καλούνται να δημιουργήσουν ένα λογαριασμό χρήστη. Για τη δημιουργία του λογαριασμού απαιτείται να συμπληρωθεί η παρακάτω φόρμα (

**Σχήμα 42)** η οποία περιέχει τα πεδία όνομα, κωδικό πρόσβασης και λογαριασμό ηλεκτρονικού ταχυδρομείου (email). Η εγγραφή του χρήστη είναι απαραίτητη να γίνει μόνο μία φορά, έπειτα για την σύνδεση στο σύστημα χρειάζεται μόνο η εισαγωγή του ονόματος και του κωδικού (**Error! Reference source not found.**) που δημιούργησε ο χρήστης.





Sign Up

Username: \_\_\_\_\_ Email: \_\_\_\_\_

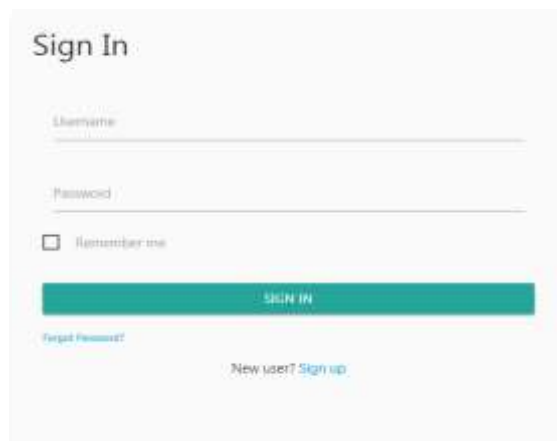
Password: \_\_\_\_\_ Confirm Password: \_\_\_\_\_

**SIGN UP**

By clicking Sign up, you agree to our [Terms](#) and that you have read our [Data Use Policy](#), including our [Cookie Use](#)

Already a member? [Sign in](#)

Σχήμα 42. Φόρμα για τη δημιουργία λογαριασμού.



Sign In

Username: \_\_\_\_\_

Password: \_\_\_\_\_

☐ Remember me

**SIGN IN**

[Forgot Password?](#)

New user? [Sign up](#)

Σχήμα 43. Φόρμα για τη σύνδεση του χρήστη.

Κατά τη διάρκεια της σύνδεσης του χρήστη στη πλατφόρμα OpenBio-C, ένα εικονίδιο με το όνομα του χρήστη εμφανίζεται στο πάνω μέρος της οθόνης για να είναι ενήμερος ότι είναι ήδη συνδεδεμένος (Σχήμα 44).



Σχήμα 44. Avatar συνδεδεμένου χρήστη.

Για την ενημέρωση και περαιτέρω παραμετροποίηση του προφίλ του, ο χρήστης πατώντας στο παραπάνω εικονίδιο μπορεί να μεταφερθεί στην οθόνη του προφίλ (Σχήμα 45):



Σχήμα 45. Οθόνη του προφίλ χρήστη.

Η εγγραφή του χρήστη θα μπορεί να γίνει στο μέλλον μέσω της σύνδεσης του OpenBio-C με το παγκόσμιο ευρετήριο ερευνητών, ORCID<sup>61</sup>.

Στο προφίλ υπάρχουν όλα τα στατιστικά στοιχεία που προκύπτουν από την αλληλεπίδραση του χρήστη με τη πλατφόρμα. Συγκεκριμένα θα υπάρχουν πληροφορίες σχετικά με:

- ❖ Ποια ερευνητικά αντικείμενα έχει δημιουργήσει, διακλαδώσει (fork) και έχει χρησιμοποιήσει;
- ❖ Πόσο δημοφιλής είναι τα ερευνητικά αντικείμενα που έχει δημιουργήσει. Για παράδειγμα πόσοι χρήστες κατά μέσο όρο τα χρησιμοποιούν πόσες διακλαδώσεις(forks) έχουν γίνει και τι βαθμολογία έχουν;
- ❖ Τα σχόλια του χρήστη προς διάφορα ερευνητικά αντικείμενα καθώς και η βαθμολογία (upvotes / downvotes) άλλων χρηστών προς αυτά τα σχόλια;
- ❖ Μία συνολική βαθμολογία σχετικά με τη θετική δράση και επίδραση του χρήστη στη κοινότητα του OpenBio-C. Αυτό θα εμφανίζεται με τη μορφή “παρασημοφορήσεων” (badges) και συνολικής βαθμολογίας. Η συγκεκριμένη λειτουργικότητα ακολουθείται από άλλα περιβάλλοντα ερωτήσεων/απαντήσεων όπως είναι το *stackoverflow.com* και *biostars.org*;

Επίσης ο χρήστης θα μπορεί να προσθέτει στο προφίλ οποιεσδήποτε πληροφορίες σχετικά με το ακαδημαϊκή του δραστηριότητες. Τέτοιες πληροφορίες είναι: ①Ακαδημαϊκά ενδιαφέροντα, τίτλος και ειδικότητα ②Το ακαδημαϊκό ίδρυμα στο οποίο εργάζεται (affiliation), ή έχει εργαστεί στο παρελθόν ③Δημοσιεύσεις.

## 6.5 Σημασιολογικός Ευρετηριασμός και Μηχανισμός Αναζήτησης / Εντοπισμού Εργαλείων

Σκοπός είναι να αξιοποιήσουμε υπάρχοντα ερευνητικά αποτελέσματα και να τα επεκτείνουμε με στόχο να παρέχουμε στους χρήστες του OpenBio-C ένα ισχυρό εργαλείο αναζήτησης πόρων, με τη βοήθεια της **εξόρυξης πληροφοριών που καθοδηγούνται από οντολογίες**.

<sup>61</sup> [orcid.org](http://orcid.org)

### 6.5.1 Λειτουργικότητα μηχανισμού αναζήτησης

Η αναζήτηση ερευνητικών αντικειμένων (εργαλείων, ροών εργασίας, δεδομένων) γίνεται με απλό και κατανοητό τρόπο και σκοπό έχει να μπορεί να χρησιμοποιηθεί από χρήστες οι οποίες δεν έχουν μεγάλη ερευνητική εξειδίκευση αλλά και από χρήστες οι οποίες δεν έχουν πολύ συγκεκριμένα κριτήρια αναζήτησης. Για παράδειγμα ένας χρήστης μπορεί να αναζητήσει με βάση κάποιον πολύ γενικό όρο (π.χ. genetics) ή κάτι πολύ συγκεκριμένο (Structural Variant Calling). Στο OpenBio-C υπάρχουν δύο τρόποι αναζήτησης, είτε με ελεύθερο κείμενο είτε με συγκεκριμένα κριτήρια. Στη παρούσα φάση έχουμε υλοποιήσει την αναζήτηση με ελεύθερο κείμενο. Συγκεκριμένα έχουμε υλοποιήσει διαδοχική αναζήτηση (incremental search), όπου τα αποτελέσματα της αναζήτησης ανανεώνονται κατά τη διάρκεια της εισαγωγής λέξεων κλειδιών. Το αποτέλεσμα της αναζήτησης οργανώνονται και παρουσιάζονται με μία δενδρική δομή η οποία περιέχει όλα τα ερευνητικά αντικείμενα τα οποία ικανοποιούν τα κριτήρια ανίχνευσης. Η δενδρική αυτή δομή περιέχει τα δένδρα διακλαδώσεων (fork tree) των ερευνητικών αντικειμένων (

Σχήμα 46). Ο χρήστης μπορεί να επιλέξει ένα από τα αντικείμενα του δένδρου (με κλικ πάνω στο δένδρο) όπου και εμφανίζονται όλες οι πληροφορίες για το συγκεκριμένο αντικείμενο.

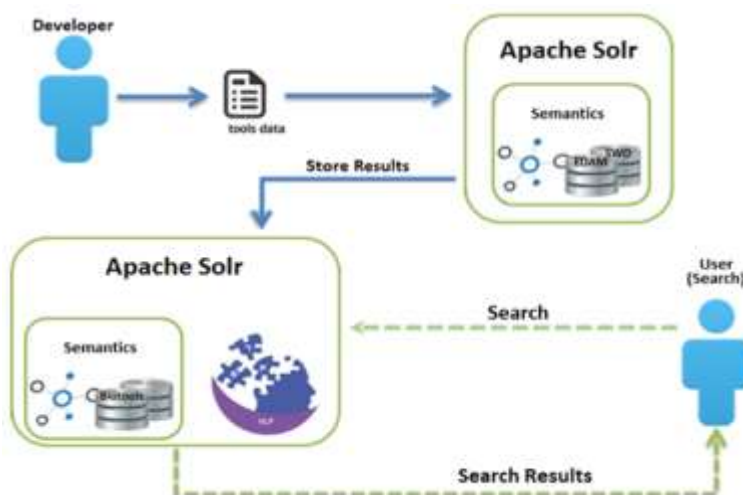


**Σχήμα 46.** Το δένδρο διακλαδώσεων (fork tree) το οποίο εμφανίζεται μετά από την εισαγωγή ενός φίλτρου αναζήτησης (σε αυτή τη περίπτωση: “NetworkX”. Για παράδειγμα το εργαλείο NetworkX/2.2/6 έχει δημιουργηθεί από τη διακλάδωση του εργαλείου NetworkX/2.2/5. Επίσης το εργαλείο NetworkX/2.2/1 έχει διακλαδοθεί δύο φορές και έχουν δημιουργηθεί τα εργαλεία: NetworkX/2.2/4 και NetworkX/2.2/2.

Τόσο τα στοιχεία που εμφανίζονται στο δένδρο, όσο και το πλήθος των φίλτρων αναζήτησης θα αυξηθεί στις επόμενες εκδόσεις.

### 6.5.2 Η “νοημοσύνη” του μηχανισμού αναζήτησης ... φέρε μου τα πιο σχετικά!

Για την υλοποίηση της μηχανής αναζήτησης χρησιμοποιούμε τεχνολογίες αιχμής στον τομέα της αναζήτησης πληροφορίας, ανοικτού πηγαίου κώδικα όπως ο Apache Solr<sup>62</sup> (Smiley *et al.*, 2015) και οι οντολογίες Embrace Data and Methods (EDAM)<sup>63</sup> (Ison *et al.*, 2013) and the Software Ontology (SWO)<sup>64</sup> (Malone *et al.*, 2014). Η αρχιτεκτονική του υπο-συστήματος αναζήτησης και εντοπισμού εργαλείων φαίνεται στο Σχήμα 47.



**Σχήμα 47:** Αρχιτεκτονική συστήματος αναζήτησης πληροφορίας

<sup>62</sup> [lucene.apache.org/solr](http://lucene.apache.org/solr)

<sup>63</sup> [edamontology.org](http://edamontology.org)

<sup>64</sup> [www.obofoundry.org/ontology/swo.html](http://www.obofoundry.org/ontology/swo.html)

Η διαδικασία αναζήτησης και εντοπισμού εξελίσσεται σε τρία βήματα:

- i. Όταν πρέπει να προστεθεί ένα νέο εργαλείο (είτε και ολόκληρη ροή εργασίας), χρησιμοποιούμε όλες τις διαθέσιμες περιγραφές τους σαν ερωτήματα ώστε να ανακτήσαμε αυτόματα τους πέντε (5) πιο ενδεικτικούς όρους με το υψηλότερο *σκορ-σημαντικότητας*, το οποίο αποδίδεται από τον solr με βάση τις αντίστοιχες και πιο κοντινές περιγραφές κάθε οντολογίας (έχουμε ήδη τροφοδοτήσει τον εξυπηρετητή του Apache solr με τις οντολογίες EDAM & SWO).
- ii. Στη συνέχεια δημιουργούμε ένα νέο αρχείο JSON, με όλη την πληροφορία/περιγραφές του εργαλείου και προσθέτουμε τους πέντε πιο ενδεικτικούς όρους από κάθε οντολογία ως ελεύθερο κείμενο. Αυτή η πληροφορία αποθηκεύεται σε ένα νέο πυρήνα Apache solr, εγκαταστημένο στη κεντρική υπολογιστική υποδομή του OpenBio-C.
- iii. Κατά τη διάρκεια της αναζήτησης, ο χρήστης μπορεί να χρησιμοποιήσει κείμενο ή λέξεις κλειδιά ως είσοδο. Ακολουθώντας την ίδια διαδικασία κατά την αναζήτηση του χρήστη, μεταφράζουμε το ερώτημα του χρήστη σε όρους οντολογίας και χρησιμοποιώντας έναν αλγόριθμο αντιστοίχισης εντοπίζουμε τα πιο όμοια (για το ερώτημα του χρήστη) εργαλεία.

**Αξιολόγηση και πιστοποίηση μηχανισμού αναζήτησης και εντοπισμού εργαλείων.** Για την επαλήθευση της μεθοδολογίας αναζήτησης και εντοπισμού εργαλείων και δεδομένων, ανακτήσαμε όλα τα δεδομένα από όλα τα εργαλεία από το μητρώο εργαλείων βιοπληροφορικής *biotools* (Ison *et al.*, 2016), το οποίο έχει υιοθετηθεί από την Ευρωπαϊκή υποδομή ELIXIR<sup>65</sup> και το οποίο περιέχει ~11.000 εργαλεία σχετικά με βιοπληροφορική. Τα εργαλεία αυτά, με τις περιγραφές τους, έχουν ήδη εισαχθεί και αποθηκευθεί σε σχετική βάση δεδομένων του OpenBio-C. Ακολούθως, αναζητήσαμε και ανασύραμε κάποιες ενδεικτικές ερωτήσεις από επιστημονικά forums βιοπληροφορικής (SeqAnswers<sup>66</sup>) και προσπαθήσαμε να ανακτήσουμε τα προτεινόμενα (από την κοινότητα του forum) εργαλεία με βάση μόνο την αρχική ερώτηση. Τα αποτελέσματα του πειράματος επαλήθευσης είναι ενθαρρυντικά δίνοντας σωστές απαντήσεις με ακρίβεια της τάξης ~80-85%, με την απόκριση του συστήματος να κινείται στη τάξη των msec, καθιστώντας τη συγκεκριμένη προσέγγιση χρησιμοποιήσιμη σε πραγματικό χρόνο, κάτι που αποτελεί αναγκαία προϋπόθεση για την αποδοχή μιας διαδικτυακής εφαρμογής. ➡ Θα πρέπει να σημειώσουμε ότι για την εισαγωγή νέων ερευνητικών αντικειμένων ο χρήστης χρησιμοποιεί το ίδιο περιβάλλον με αυτό της αναζήτησης. Όταν η αναζήτηση δεν εντοπίζει κάποιο αντικείμενο τότε εμφανίζεται η επιλογή για δημιουργία καινούργιου. Στην επόμενη έκδοση του μηχανισμού αναζήτησης και εντοπισμού εργαλείων και δεδομένων προτιθέμεθα να εκμεταλλευτούμε και να προσαρμόσουμε τεχνικές *βαθιάς μάθησης* (deep learning) οι οποίες έχουν αποδειχθεί ιδιαίτερα αποδοτικές και ακριβείς στον εντοπισμό ομοίων αναφορών ελεύθερου κειμένου / περιγραφών στο πεδίο της *διαχείρισης πληροφοριών* (Mitra and Craswell, 2017; Craswell *et al.*, 2018).

## 6.6 Βασικές σχεδιαστικές αρχές και η αρχική υλοποίηση του περιβάλλοντος σύνθεσης ροών εργασίας του OpenBio-C

Μία από τις κυριότερες προκλήσεις του έργου είναι η παροχή δυνατότητας σε όλους τους χρήστες να δημιουργούν δικές τους “εκδόσεις” από ερευνητικά αντικείμενα που εισάγουν (ή μεταβάλλουν) άλλοι χρήστες, χωρίς όμως να μπορούν να μεταβάλλουν τα ερευνητικά αντικείμενα άλλων χρηστών.

Στο OpenBio-C δίνουμε την εγγύηση ότι ένα ερευνητικό αντικείμενο το οποίο σώζεται στη βάση δεδομένων, θα παραμείνει αμετάβλητο για πάντα. Για να το υλοποιήσουμε αυτό υιοθετήσαμε τη λογική των *forks* (διακλαδωμένες υλοποιήσεις) τα οποία αποτελούν μία πολύ δημοφιλή έννοια στη διαχείριση λογισμικού. Κάθε ερευνητικό αντικείμενο συνοδεύεται από ένα όνομα και έναν μοναδικό αριθμό για αυτό το όνομα. Τον αριθμό αυτό τον ονομάζουμε “έκδοση” (edit). Κάθε αντικείμενο το οποίο έχει σωθεί στη βάση και περιέχει κάποια συγκεκριμένα “έκδοση”, δεν μπορεί να αλλάξει (θεωρείται immutable, αμετάβλητο). Ο οποιοσδήποτε χρήστης όμως, και με το πάτημα ενός κουμπιού, μπορεί να κάνει “fork” οποιοδήποτε αντικείμενο, δημιουργώντας ένα αντίγραφο του αντικειμένου αυτού, το οποίο έχει ακριβώς τα ίδια στοιχεία εκτός από τον αριθμό της “έκδοσής” του. Αυτό το νέο αντικείμενο μπορεί να το μεταβάλει ο χρήστης όπως θέλει, μέχρι να το αποθηκεύσει στη βάση οπότε πάλι θεωρείται αμετάβλητο. Έτσι, για όλα τα ερευνητικά αντικείμενα υπάρχει μία *δενδρική δομή* με όλες τις εκδόσεις που έχουν φτιάξει οι χρήστες. Κάθε αντικείμενο έχει τα δικά του στατιστικά χρήσης, τον δικό του σχολιασμό από χρήστες, και τα δικά του χαρακτηριστικά ποιότητας που αποδίδονται από το σύστημα.

<sup>65</sup> <https://www.elixir-europe.org/>

66 <http://seganswers.com/>

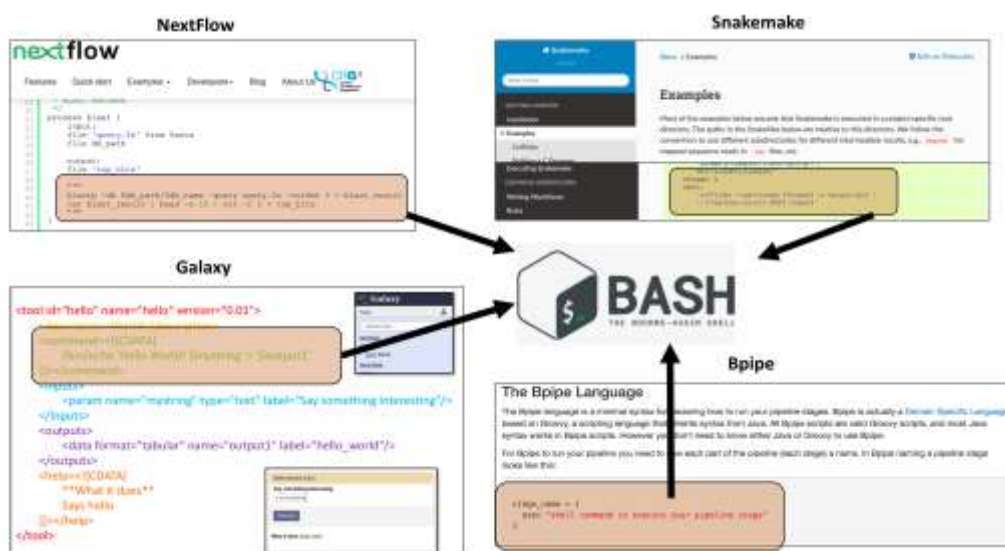


### 6.6.1 Εξαρτήσεις εργαλείων

Στην επιστημονική κοινότητα της βιοπληροφορικής, υπάρχουν εργαλεία τα οποία θεωρούνται “βασικά” και εργαλεία τα οποία είναι πιο πειραματικά, χρησιμοποιούνται σπάνια και βασίζονται στα “βασικά”. Επίσης υπάρχουν κοινές προγραμματιστικές βιβλιοθήκες (π.χ. για στατιστική, γραμμική άλγεβρα, αριθμητική ανάλυση) οι οποίες χρησιμοποιούνται από πληθώρα άλλων εργαλείων. Αυτό δημιουργεί μία αλληλεξάρτηση από εργαλεία και βιβλιοθήκες. Αν ένα ερευνητικό αντικείμενο το οποίο μοντελοποιεί και τελικά αναπαριστά ένα ερευνητικό εργαλείο, δεν περιέχει οδηγίες εγκατάστασης για τις βιβλιοθήκες από τις οποίες εξαρτάται, τότε είναι πρακτικά άχρηστο. Από την άλλη μεριά, αν κάθε μικρό αντικείμενο περιέχει πληροφορία για τις εν-δυνάμει δεκάδες εξαρτήσεις του, τότε, τόσο η εισαγωγή η συντήρηση, και η διαχείρισή του γίνεται πολύπλοκη και μη-πρακτική. Μία επιπλέον καινοτομία του OpenBio-C είναι ότι ο χρήστης μπορεί να δηλώνει ότι ένα εργαλείο έχει ως εξάρτηση ένα ή περισσότερα άλλα εργαλεία. Τα εργαλεία αυτά είναι ανεξάρτητα ερευνητικά αντικείμενα, υπάρχουν σε διαφορετικές εγγραφές στη βάση δεδομένων και έχουν τα δικά τους στατιστικά χρήσης και τις δικές τους συζητήσεις. Όταν ένα εργαλείο γίνεται fork, τότε “υιοθετεί” και τις εξαρτήσεις του αρχικού εργαλείου, ο χρήστης φυσικά μπορεί να τις μεταβάλλει. Οπότε κάθε εργαλείο “συνοδεύεται” από το **δέντρο εξαρτήσεων** του. Το δέντρο αυτό είναι «ορατό» και στο γραφικό περιβάλλον σύνθεσης ροών εργασίας.

### 6.6.2 Η δύναμη του BASH

Σήμερα υπάρχουν πάνω από 100 εργαλεία διαχείρισης επιστημονικών ροών εργασίας. Επίσης είναι πολύ κοινό κάθε εργαλείο να ορίζει και τη δική του γλώσσα περιγραφής ροών εργασίας (γνωστά και ως Domain Specific Languages). Κάθε νέα γλώσσα, όσο απλή και να είναι, απαιτεί από τους χρήστες προσπάθεια και χρόνο για την εκμάθησή της. Επιπλέον κάθε γλώσσα απαιτεί λογισμικό το οποίο ελέγχει το συντακτικό και τη σημασιολογική ορθότητα των ροών εργασίας που περιγράφονται σε αυτή. Τα παραπάνω προσθέτουν αναίτια πολυπλοκότητα στα συστήματα ροών εργασίας και καθιστούν την εξάπλωσή τους δύσκολη. Ως παράδειγμα παραθέτουμε ότι για το 2018, ο αριθμός των δημοσιευμένων εργασιών οι οποίες χρησιμοποίησαν το Galaxy ως σύστημα διαχείρισης ροών εργασίας ήταν περίπου 1,000<sup>67</sup>.



**Σχήμα 48.** Τέσσερα από τα πιο βασικά εργαλεία διαχείρισης ροών εργασίας (NextFlow, Galaxy, Snakemake, Bpipe) χρησιμοποιούν BASH για τη περιγραφή των επιμέρους βημάτων των ροών εργασιών.

Αν θεωρήσουμε ότι το Galaxy είναι το πιο διαδεδομένο περιβάλλον και ότι συνολικά δημοσιεύονται πάνω από 400.000 εργασίες στη περιοχή της βιολογίας, γενετικής και βιοπληροφορικής<sup>68</sup> τότε, παρατηρούμε ότι η συντριπτική πλειοψηφία των δημοσιευμένων αναλύσεων δεν χρησιμοποιεί κάποιο περιβάλλον διαχείρισης ροών εργασιών. Παρόλα αυτά κάθε εργασία η οποία απαιτεί τον συνδυασμό παραπάνω του ενός εργαλείου, ή απαιτεί ένα εργαλείο το οποίο είναι διαθέσιμο μόνο σε περιβάλλον Linux, τότε

<sup>67</sup> <https://galaxyproject.org/galaxy-project/statistics/>

<sup>68</sup> <http://bio.biologists.org/content/7/8/bio037325>

χρησιμοποιεί το περιβάλλον BASH. Το BASH είναι το πιο κοινό περιβάλλον διαχείρισης γραμμής εντολών σε περιβάλλον Linux και OSX (υπάρχουν και εκδόσεις για Windows). Αν και το BASH δεν είναι ένα περιβάλλον διαχείρισης ροών εργασιών είναι στην ουσία ένα περιβάλλον όπου μπορούν να περιγραφούν τα επιμέρους βήματα μίας ροής εργασίας. Επίσης το περισσότερο υπάρχοντα περιβάλλοντα για διαχείριση ροών εργασιών κωδικοποιούν ή περιγράφουν τα επιμέρους βήματα σε μορφή BASH (Σχήμα 48).

Παρατηρούμε δηλαδή ότι τα περισσότερα υπάρχοντα περιβάλλοντα παρέχουν μία διαστρωμάτωση όπου στο βασικό επίπεδο όλες οι λειτουργίες γίνονται σε BASH, ενώ σε ανώτερο επίπεδο ο χρήστης διαχειρίζεται τα BASH scripts σε κάποιο εύχρηστο περιβάλλον. Ως βασικό σχεδιαστικό κριτήριο των εργαλείων αυτών είναι να «αποκρύψουν» τον χρήστη από τη πολυπλοκότητα και δυσκολία του BASH. Συνήθως όμως οι χρήστες συνεχίζουν να γράφουν BASH και να χρησιμοποιούν τα περιβάλλοντα αυτά για διαχείριση και συνένωση των επιμέρους βημάτων.

Η καινοτομία του OpenBio-C είναι ότι φέρνει το BASH στην “επιφάνεια”. Ο χρήστης δεν χρειάζεται να μάθει καμία καινούργια γλώσσα προγραμματισμού ή εξειδικευμένη γλώσσα περιγραφής (DSL). Για την εγκατάσταση ενός εργαλείου ο χρήστης απλά εισάγει τα BASH commands τα οποία το εγκαθιστούν. Ομοίως όταν περιγράφει ένα βήμα μίας ροής εργασίας απλά εισάγει τις εντολές σε BASH οι οποίες εκτελούν αυτό το βήμα. Το OpenBio-C αναλαμβάνει να (1) δώσει στον χρήστη ένα φιλικό διαδικτυακό περιβάλλον για επεξεργασία αυτών των εντολών, (2) αποθηκεύει τις εντολές στη βάση δεδομένων και αποτελούν μια πληροφορία που *επισημαίνει* (annotate) και *τεκμηριώνει* το ερευνητικό αντικείμενο το οποίο μοντελοποιεί τα εργαλεία και τις ροές εργασίας και φυσικά (3) να τις εκτελεί σε κάποιο περιβάλλον που επιλέγει ο χρήστης.

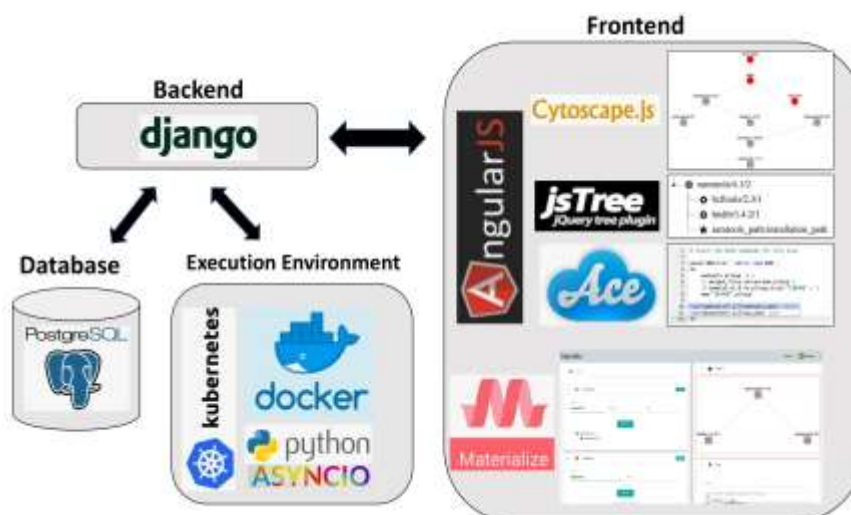
■ **Εισαγωγή και επεξεργασία BASH script.** Όπως έχουμε περιγράψει, η εισαγωγή εργαλείων και η περιγραφή των βημάτων στις ροές εργασίας γίνεται μέσω BASH scripts. Για να γίνει αυτό, πρέπει ο χρήστης να μπορεί να εισάγει αυτές τις εντολές σε ένα περιβάλλον το οποίο να παρέχει οπτικά βοηθήματα παρόμοια με αυτά των σύγχρονων ολοκληρωμένων περιβαλλόντων ανάπτυξης (IDEs). Για αυτό το σκοπό χρησιμοποιούμε τη javascript βιβλιοθήκη ace. Έχουμε προσαρμόσει τη βιβλιοθήκη ώστε να παρέχει αυτόματη ολοκλήρωση (autocomplete) κατά τη διάρκεια επεξεργασίας του κειμένου BASH. Για παράδειγμα όταν ένας χρήστης θέλει να χρησιμοποιήσει ένα εργαλείο από ένα βήμα μίας ροής εργασίας, αρκεί να πατήσει “tool” και αυτόματα εμφανίζεται μία λίστα με όλα τα διαθέσιμα εργαλεία που έχει η ροή.

■ **Δέντρα εξαρτήσεων και δέντρα forks.** Όπως έχουμε περιγράψει, οι χρήστες μπορούν να δηλώσουν ότι ένα εργαλείο απαιτεί (ή αλλιώς εξαρτάται) από κάποιο άλλο εργαλείο. Αυτό σημαίνει ότι κάθε εργαλείο συνοδεύεται από ένα δένδρο εξαρτήσεων. Παράλληλα ένας χρήστης μπορεί να κάνει fork οποιοδήποτε ερευνητικό αντικείμενο (εργαλεία, δεδομένα, ροές εργασίας). Έτσι, κάθε αντικείμενο ανήκει σε ένα δένδρο-fork. Ο πατέρας ενός κόμβου είναι το ερευνητικό αντικείμενο από το οποίο έγινε fork και τα παιδιά είναι τα αντικείμενα που δημιουργήθηκαν από fork αυτού του κόμβου. Και οι δύο μορφές δένδρων διαχειρίζονται από το frontend μέσω της βιβλιοθήκης jsTree.

### 6.6.3 Το σύστημα εκτέλεσης ροών εργασιών (Execution Environment)

Όταν ο χρήστης εισάγει ένα εργαλείο, δεδομένα ή ροή εργασίας, το OpenBio-C αναλαμβάνει να τα εκτελέσει πρωτίστως για να επιβεβαιώσει την ορθότητά τους. Για να γίνει αυτό, αρχικά, συλλέγονται οι εντολές BASH που αποτελούν το εργαλείο ή τη ροή εργασίας. Οι εντολές αυτές στέλνονται σε ένα ανεξάρτητο υποσύστημα το οποίο έχει τη δική του αρχιτεκτονική και «τρέχει» στο δικό του υπολογιστικό περιβάλλον. Το υποσύστημα αυτό λαμβάνει «αιτήσεις» από το σύστημα υποστήριξης (backend) του OpenBio-C για εκτέλεση κάποιου προγράμματος BASH. Όταν λάβει μία τέτοια αίτηση, ενεργοποιεί ένα εικονικό υπολογιστικό περιβάλλον, στο οποίο εκτελεί τις εντολές και όταν αυτές τερματίσουν (ή περάσει ένα προκαθορισμένο χρονικό όριο, ή συμβεί κάποιο λάθος κατά την εκτέλεση) επιστρέφει τα αποτελέσματα της εκτέλεσης στο backend. Το εικονικό υπολογιστικό περιβάλλον διαμορφώνεται μέσω του δημοφιλούς λογισμικού Docker. Είναι σημαντικό να τονίσουμε ότι η αρχιτεκτονική δομή αυτού του συστήματος βασίζεται στην “ασύγχρονη εκτέλεση”. Το backend μπορεί να στείλει χιλιάδες αιτήσεις σε ένα λεπτό, οι οποίες μπαίνουν σε μία ουρά προτεραιότητας. Στην συνέχεια, δευτερεύοντες διεργασίες «παίρνουν» παράλληλα αιτήσεις από την ουρά προτεραιότητας, τις εκτελούν και επιστρέφουν το αποτέλεσμα. Αυτό μας δίνει τη δυνατότητα να εκτελούμε

όλες τις “αιτήσεις” ανεξάρτητα από την υπάρχουσα υπολογιστική υποδομή. Επίσης το σύστημα αυτό είναι “αρθρωτό” και παρέχει αυτόματη κλιμάκωση (scaling) με τη προσθήκη επιπλέον υπολογιστικών πόρων. Προς το παρόν η υλοποίησή του βασίζεται στη βιβλιοθήκη **asyncio** της Python 3. Στο μέλλον όμως η υλοποίηση θα βασίζεται στο περιβάλλον διαχείρισης εικονικών υπολογιστικών περιβαλλόντων, **Kubernetes**<sup>69</sup>. Η όλη αρχιτεκτονική σύνθεσης ροών εργασίας στο OpenBio-C και η διασύνδεση με τις άλλες υπο-μονάδες της πλατφόρμας απεικονίζονται στο Σχήμα 49.

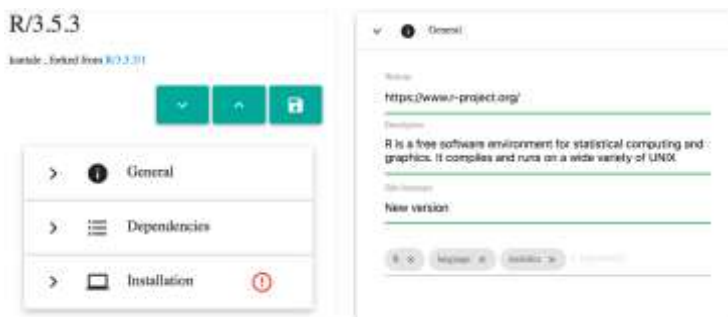


**Σχήμα 49.** Οι 4 κύριες υπο-μονάδες του OpenBio-C για τη σύνθεση ροών εργασίας και οι τεχνολογίες ανοικτού κώδικα που χρησιμοποιούμε σε αυτά.

#### 6.6.4 Εισαγωγή εργαλείων στο OpenBio-C

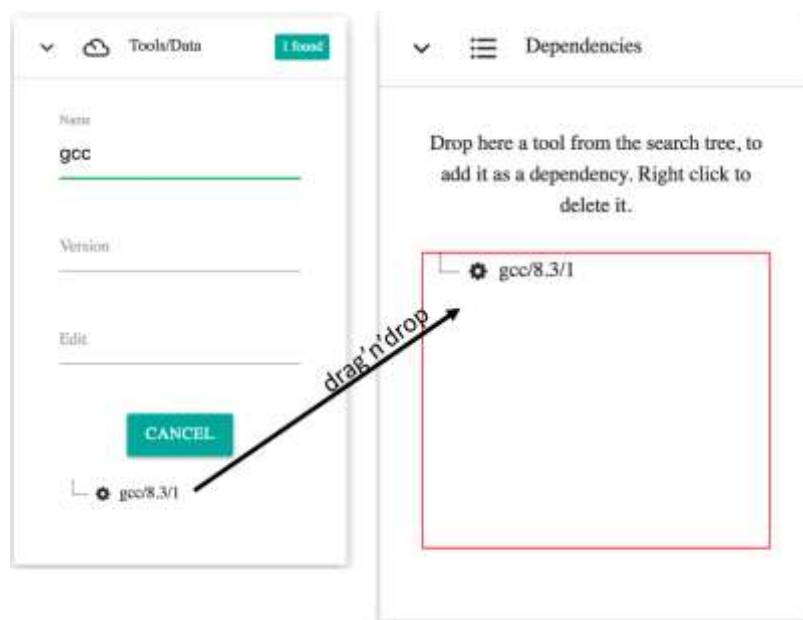
Κατά την εισαγωγή ενός νέου εργαλείου ο χρήστης καλείται να εισάγει τα ακόλουθα πεδία:

- Ένα site αναφοράς με αυτό το εργαλείο ή δεδομένα (Σχήμα 50).
- Μία περιγραφή σε ελεύθερο κείμενο (Σχήμα 50)
- Λέξεις κλειδιά σχετικά με αυτό το εργαλείο ή δεδομένα (Σχήμα 50)
- Εισαγωγή άλλων εργαλείων από τα οποία αυτό το εργαλείο εξαρτάται. Η εισαγωγή γίνεται με μεταφορά και απόθεση από το περιβάλλον αναζήτησης στο ειδικό δένδρο εξαρτήσεων. Οι εξαρτήσεις αυτές μπορούν να διαγραφούν με δεξί κλικ (Σχήμα 51).
- Τις εντολές BASH οι οποίες εγκαθιστούν αυτό το εργαλείο ή μεταφορτώνουν αυτά τα δεδομένα (Σχήμα 52).
- Τις εντολές BASH οι οποίες επιβεβαιώνουν τη σωστή εγκατάσταση του εργαλείου (Σχήμα 52).
- Οι μεταβλητές οι τιμές των οποίων γίνονται προσβάσιμες από οποιοδήποτε άλλο εργαλείο θα εξαρτάται από το εργαλείο που εισάγουμε. Επίσης αυτές οι μεταβλητές είναι ορατές από όλες τις ροές εργασίας οι οποίες χρησιμοποιούν αυτό το εργαλείο (Σχήμα 53).



**Σχήμα 50.** Το βασικό τμήμα όπου εισάγονται πληροφορίες για τα εργαλεία ή για τα δεδομένα. Αριστερά εμφανίζονται όλα τα υπο-μενού και δεξιά το υπο-μενού όπου εισάγουμε τις βασικές πληροφορίες (Ιστότοπος, περιγραφή και λέξεις κλειδιά)

<sup>69</sup> <https://kubernetes.io/>



Σχήμα 51. Οι εξαρτήσεις (dependencies) των εργαλείων γίνεται με μεταφορά και απόθεση από το περιβάλλον αναζήτησης.

## Installation Commands

```
1 wget https://cloud.r-project.org/
2 tar zxvf r-base_3.5.0.orig.tar.gz
3 cd r-base_3.5.0
4 ./configure
5 make
6 make install
7
```

## Validation Commands

```
1 R
2
3 exit $?
4
```

Σχήμα 52. Εισαγωγή των εντολών BASH για την εγκατάσταση (αριστερά) και την επιβεβαίωση της εγκατάστασης (δεξιά) ενός νέου εργαλείου. Το περιβάλλον εισαγωγής και επεξεργασίας BASH είναι η javascript βιβλιοθήκη ace, η οποία δίνει οπτικά βοηθήματα παρόμοια με σύγχρονα ολοκληρωμένα περιβάλλοντα προγραμματισμού (IDEs).

## Variables

| Name | Value      | Description     |
|------|------------|-----------------|
| R    | /usr/bin/R | executable path |

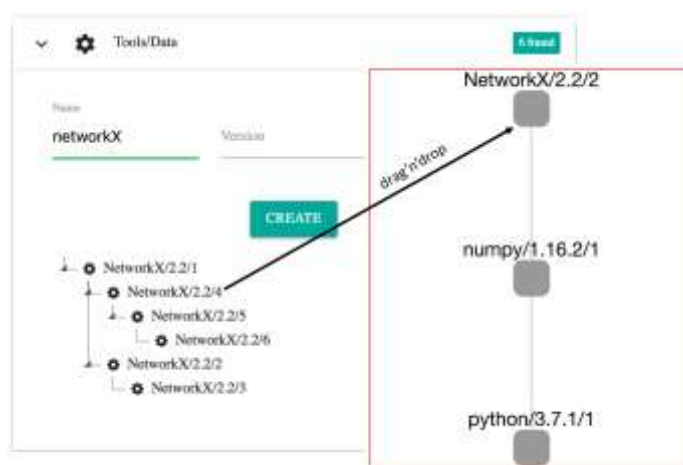
Σχήμα 53. Κάθε εργαλείο μπορεί να έχει μεταβλητές οι τιμές των οποίων γίνονται ορατές σε άλλα εργαλεία τα οποία εξατώνται από αυτό ή σε ροές εργασίας οι οποίες το χρησιμοποιούν. Μία κλασική χρήση αυτών των μεταβλητών είναι για να δηλώσουμε τον κατάλογο που είναι εγκαταστημένο το εργαλείο.

Για την εισαγωγή μίας νέας ροής εργασίας ο χρήστης καλείται να εισάγει έναν *ιστότοπο αναφοράς*, μία περιγραφή και λέξεις κλειδιά καθώς και στο περιβάλλον εισαγωγής εργαλείων. Στη συνέχεια ο χρήστης αλληλοεπιδρά με τον γράφο που απαρτίζει τη ροή εργασίας με τον ακόλουθο τρόπο:

- Για να εισάγει εργαλεία (μαζί με τις εξαρτήσεις τους) κάνει μεταφορά και απόθεση από το περιβάλλον αναζήτησης (**Error! Reference source not found.**).
- Για να εισάγει τα βήματα εκτέλεσης της BEPE, ο χρήστης εισάγει της εντολές που περιγράφουν κάθε βήμα σε μορφή BASH. Μέσα από αυτές τις εντολές μπορεί να:
  - ο Έχει πρόσβαση στις μεταβλητές των εργαλείων (**Error! Reference source not found.**).
  - ο Καλέσει ένα άλλο βήμα της BEPE μέσω της εντολής call. Για παράδειγμα, αν υποθέσουμε ότι υπάρχει ένα βήμα με το όνομα step\_1 μπορεί να γράψει: call(step\_1).



- ο Να έχει πρόσβαση στις παραμέτρους της BEPE.
- ο Να θέσει τιμή στις μεταβλητές που έχουν οριστεί ως «έξοδος» της BEPE και περιέχουν τα αποτελέσματα της επεξεργασίας.
- Σε όλα τα παραπάνω βήματα ο χρήστης καθοδηγείται μέσω αυτόματης συμπλήρωσης (autocompletion) ώστε να επιλέξει τα κατάλληλα εργαλεία, βήματα, ή παραμέτρους (**Error! Reference source not found.**).
- Το OpenBio-C επίσης θα υποστηρίζει και την εισαγωγή πολλαπλών BEPE σε μία υπάρχουσα BEPE όπως ακριβώς γίνεται και με τα εργαλεία, δηλαδή μέσω μεταφοράς και απόθεσης.
- Κάθε BEPE έχει μεταβλητές ή αλλιώς παραμέτρους. Όταν καλείται να εκτελεστεί μία BEPE τότε το σύστημα ή ο χρήστης που τη κάλεσε θα πρέπει να δώσει τιμές σε αυτές τις μεταβλητές. Για παράδειγμα μία BEPE που υπολογίζει ποιές μεταλλάξεις είναι στατιστικά σημαντικές, μπορεί να έχει μία παράμετρο που να ορίζει το όριο στατιστικής σημαντικότητας. Ομοίως μία BEPE μπορεί να έχει μεταβλητές οι οποίες να περιέχουν τα αποτελέσματα της επεξεργασίας. Αυτές οι δύο τύπου μεταβλητές ονομάζονται “inputs” και “outputs” αντίστοιχα και ορίζονται κατά τη διάρκεια δημιουργίας μίας BEPE (Σχήμα 54).



**Σχήμα 54.** Η εισαγωγή ενός εργαλείου (ή κάποιων δεδομένων) σε μία BEPE γίνεται μέσω μεταφοράς και απόθεσης στο διαδικτυακό γραφικό περιβάλλον που περιέχει τη γραφική αναπαράσταση της BEPE.

| Name                                       | Name                                       |
|--|--|
| alignment                                  | load_data                                  |
| 1 # Insert the BASH commands for this step | 1 # Insert the BASH commands for this step |
| 2  | 2  |
| 3 tool                                     | 3 call                                     |
| tool/samtools/4.1/2/samtools_path instal   | call(visualize) STEP                       |
| tool/NetworkX/2.2/2/nws_path path          | call(alignment) STEP                       |

**Σχήμα 55.** Όταν ο χρήστης εισάγει τις εντολές BASH για ένα καινούργιο βήμα σε μία BEPE, τότε υποβοηθάτε μέσω της αυτόματης συμπλήρωσης για να επιλέξει τα κατάλληλα εργαλεία ή να καλέσει τα κατάλληλα βήματα. Όταν πληκτρολογεί “tool” τότε εμφανίζεται μία λίστα με όλα τα εργαλεία και όλες τις μεταβλητές που έχουν αυτά εργαλεία (αριστερά). Όταν πληκτρολογεί “call”, τότε εμφανίζεται μία λίστα με όλα τα βήματα που μπορεί να καλέσει.

| Variables      |                                     |   |
|----------------|-------------------------------------|---|
| Name           | Description                         | Input/Output  |
| significance   | threshold of statistical significan | IN <input checked="" type="checkbox"/> OUT <input type="checkbox"/> |
| manhattan_plot | path of manhattan plot              | IN <input checked="" type="checkbox"/> OUT <input type="checkbox"/> |

**Σχήμα 56.** Ορισμός μεταβλητών μίας BEPE. Οι μεταβλητές αυτές είναι είτε οι τιμές των παρμέτρων εισόδου (inputs), είτε τα αποτελέσματα της επεξεργασίας (outputs).

## 6.7 Το Συνεργατικό περιβάλλον του OpenBio-C

Αυτή τη στιγμή υπάρχουν διάφορες υποψήφιες βιβλιοθήκες για την υλοποίηση γράφων τύπου mind map που βοηθούν στην αναπαράσταση συζητήσεων από συνεργατικά περιβάλλοντα. Τέτοιες βιβλιοθήκες είναι: alchemy.js (<http://graphalchemist.github.io/Alchemy>) ; cytoscape.js (<http://js.cytoscape.org>) (Franz *et al.*, 2016) ; sigma.js (<http://sigmajs.org>) ; d3.js (<https://d3js.org>) ; vis.js (<http://visjs.org>) ; dracula.js (<https://www.graphdracula.net>)

Παρόλο που η βιβλιοθήκη d3.js είναι από τις πιο διάσημες, για τους λόγους που εξηγούμε παρακάτω αποφασίσαμε να χρησιμοποιήσουμε την cytoscape.js. Οι υπόλοιπες βιβλιοθήκες είτε είναι όχι τόσο ώριμες (π.χ. alchemy.js), είτε δεν επιτρέπουν την εισαγωγή ad-hoc λογικής στο σύστημα (π.χ. sigma.js).

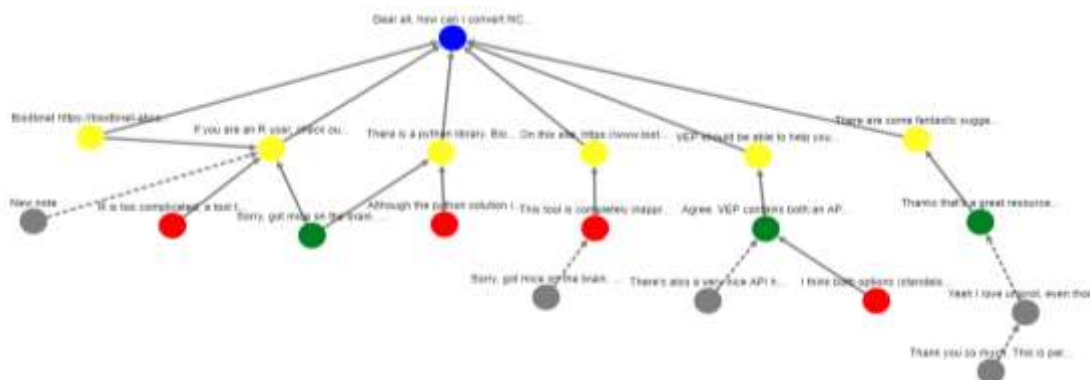
Η Cytoscape.js είναι μία βιβλιοθήκη ανοικτού λογισμικού (open-source). Χρησιμοποιείται, κυρίως, στην απεικόνιση διαδραστικών γράφων σε εφαρμογές Web. Θεωρείται ως μία μοντέρνα εναλλακτική λύση του βασισμένου σε Adobe Flash Cytoscape συστήματος. Ορισμένα από τα θετικά χαρακτηριστικά της, τα οποία ικανοποιούν τις απαιτήσεις του έργου, είναι τα ακόλουθα:

- Είναι μία αυτόνομη (standalone) JavaScript βιβλιοθήκη
- Δεν έχει εξαρτήσεις (dependencies) με άλλες βιβλιοθήκες
- Μπορεί να εκτελεστεί σε πρόγραμμα-πελάτη (client-side) αλλά και σε πρόγραμμα-διακομιστή (server-side)
- Επιτρέπει να αλλάξουμε το styling του συστήματος μέσω CSS εντολές.
- Επιτρέπει τη δημιουργία διάφορων τύπων γράφων (π.χ. κατευθυνόμενος, μη κατευθυνόμενος, σύνθετος, απλός)
- Παρέχει κλασσικούς αλγόριθμους γράφων (π.χ. συντομότερο μονοπάτι, ελάχιστο δέντρο επικάλυψης, κατάταξη)
- Παρέχει ενσωματωμένη υποστήριξη (built-in support) για συσκευές με οθόνη αφής (touch based devices)
- Επιτρέπει την δημιουργία animations
- Επιτρέπει την εισαγωγή και εξαγωγή (import & export) δεδομένων σε JSON μορφή
- Παρέχει Διεπαφή Προγραμματισμού Εφαρμογών (API) για τη δημιουργία επεκτάσεων (extensions)
- Ανταποκρίνεται σε ικανοποιητικό βαθμό σε περιβάλλοντα με υψηλή απαίτηση διαχείρισης δεδομένων (data-intensive environments)
- Υπάρχουν αρκετές υλοποιημένες επεκτάσεις (extensions) που παρέχουν χρήσιμα χαρακτηριστικά (features), όπως επεξεργασία κόμβων, επεξεργασία ακμών (Dogrusoz *et al.*, 2018).

### 6.7.1 Σχεδιαστικές αρχές & λειτουργικότητα του συνεργατικού της αρχικής υλοποίησης του συνεργατικού περιβάλλοντος του OpenBio-C

Το συνεργατικό περιβάλλον που υλοποιείται στα πλαίσια του έργου υποστηρίζει τα ακόλουθα βασικά χαρακτηριστικά:

- Εισαγωγή/Επεξεργασία/Διαγραφή κόμβων στον γράφο της συζήτησης / επιχειρηματολογίας
- Επεξεργασία του κειμένου ενός κόμβου του γράφου συζήτησης
- Παροχή σύντομων πληροφοριών μέσω ειδικού tooltip



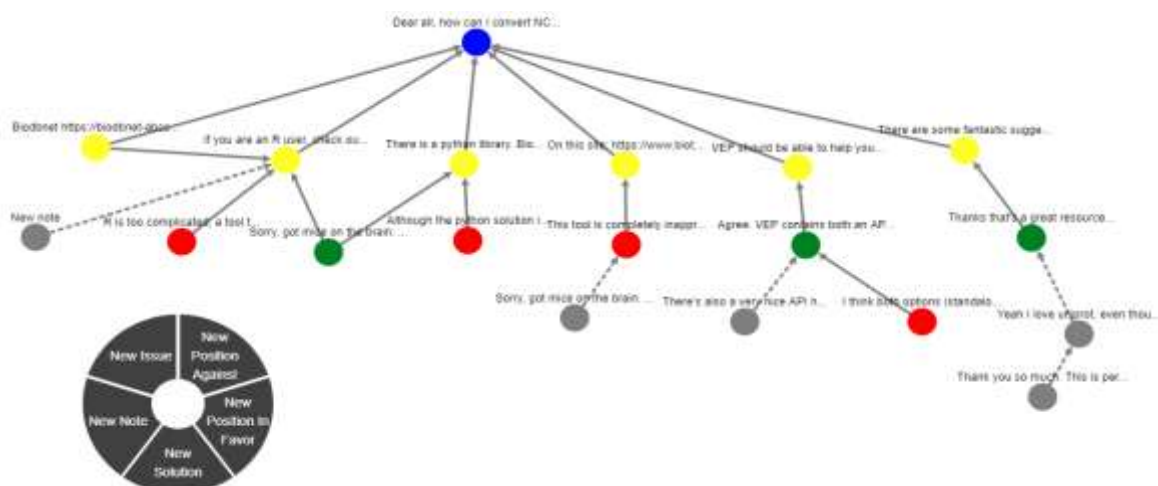
Σχήμα 57. Συνεργατικό περιβάλλον: αποτύπωση της συζήτησης και επιχειρηματολογίας στο OpenBioC σε μορφή γράφου.

Κάθε συζήτηση στο συνεργατικό περιβάλλον αναπαρίσταται σαν ένας **γράφος επιχειρηματολογίας** (βλέπε Σχήμα 57). Το μοντέλο επιχειρηματολογίας που χρησιμοποιείται είναι βασισμένο στο IBIS (Issue-Based Information System) (Kunz and Rittel, 1970) ακολουθώντας τις σχεδιαστικές αρχές που παρουσιάζονται στο (Scheuer *et al.*, 2010). Κάθε κόμβος στο γράφο μπορεί να ανήκει σε μία/έναν από τις/τους παρακάτω κατηγορίες/τύπους:

- Issue (● – αντιστοιχεί σε μία ερώτηση/πρόβλημα που έχει κάποιος χρήστης)
- Solution (● – μία προτεινόμενη λύση σε κάποιο πρόβλημα)
- Position in favor (● – υπερασπίζεται μία λύση που προτείνεται για κάποιο πρόβλημα)
- Position against (● – διαφωνεί με μία λύση που προτείνεται για κάποιο πρόβλημα)
- Note (● – ένα σχόλιο/έναν ουδέτερο κόμβος)

Για την υλοποίηση της back-end εφαρμογής χρησιμοποιήθηκε η γλώσσα Python και συγκεκριμένα το Framework Flask. Τα δεδομένα από τον γράφο μίας συζήτησης αποθηκεύονται, προσωρινά, σε JSON αρχεία.

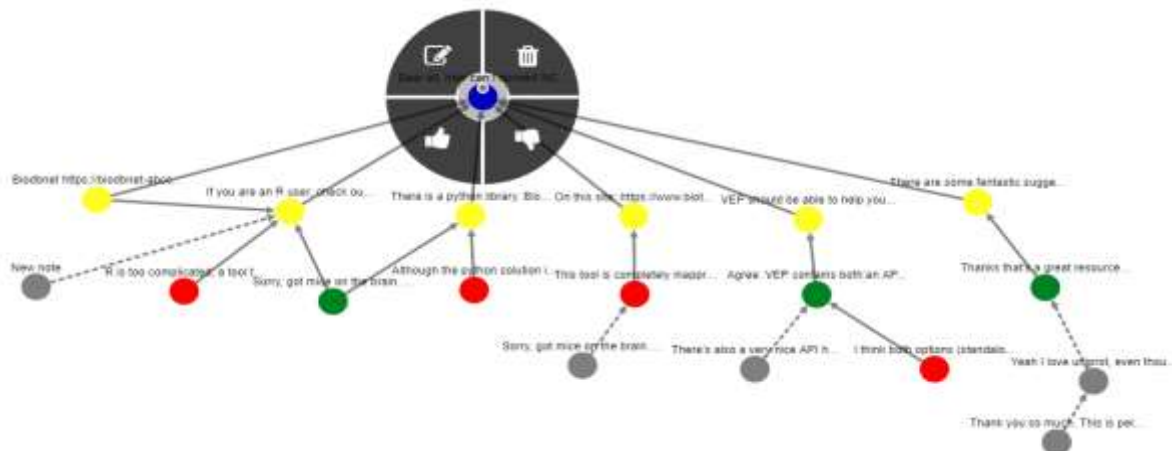
Στα σχήματα Σχήμα **58** έως Σχήμα **61** απεικονίζονται βασικές λειτουργίες του συνεργατικού περιβάλλοντος. Στο τρέχον παράδειγμα, χρησιμοποιήθηκαν δεδομένα από συζήτηση την οποία ανασύραμε από τον ιστότοπο του Reddit<sup>70</sup>. Πιο συγκεκριμένα, στο στιγμιότυπο που παρουσιάζεται στο Σχήμα 57 απεικονίζεται με μπλε κόμβο η βασική ερώτηση/ζήτημα (issue) ενός χρήστη, η οποία είναι: “How to convert NCBI RefSeq IDs to Gene names / symbols”. Παρατηρούμε πως έχουν δοθεί 6 λύσεις (solutions) στο συγκεκριμένο ζήτημα, οι οποίες απεικονίζονται με κόμβους κίτρινου χρώματος. Όπως είναι αναμενόμενο, υπάρχουν αντιπαραθέσεις μεταξύ των λύσεων και διαφορετικές απόψεις. Με άλλα λόγια, υπάρχουν επιχειρήματα υπέρ (positions in favor, πράσινοι κόμβοι) και επιχειρήματα κατά (positions against, κόκκινοι κόμβοι) κάποιων λύσεων. Υπάρχει και η δυνατότητα ανάρτησης κάποιων σχολίων/σημειώσεων (notes, γκρι κόμβοι) που αφορούν είτε σε λύσεις είτε σε επιχειρήματα. Στο Σχήμα 58 παρουσιάζεται το menu που βοηθάει τον χρήστη να δημιουργήσει έναν νέο κόμβο στον γράφο επιχειρηματολογίας. Αυτό το menu μπορεί να εμφανιστεί επιλέγοντας δεξί κλικ στο συνεργατικό περιβάλλον. Οι διαθέσιμες επιλογές είναι η δημιουργία νέου κόμβου ζητήματος (issue), η δημιουργία νέου κόμβου επιχειρήματος υπέρ και κατά μίας λύσης (position in favor, position against), η δημιουργία νέου κόμβου λύσης (solution) και η δημιουργία νέου κόμβου σημείωσης (note). Κατά την εισαγωγή ενός κόμβου, καταγράφεται ο δημιουργός και η ημερομηνία δημιουργίας του.



**Σχήμα 58.** Εισαγωγή νέου κόμβου.

Στο Σχήμα 59 παρουσιάζεται το menu που βοηθάει τον χρήστη στην επεξεργασία ενός κόμβου του γραφού. Σε κάθε κόμβο, ο χρήστης έχει την δυνατότητα να τον διαγράψει και να επεξεργαστεί το σχετικό με αυτόν κείμενο (στην περίπτωση που ο χρήστης είχε δημιουργήσει τον κόμβο). Επιπλέον, μπορεί να κάνει like ή dislike στον κόμβο αυτό.

<sup>70</sup> [www.reddit.com/r/bioinformatics/comments/8jcvql/how to convert ncbi refseq ids to gene names](https://www.reddit.com/r/bioinformatics/comments/8jcvql/how_to_convert_ncbi_refseq_ids_to_gene_names)



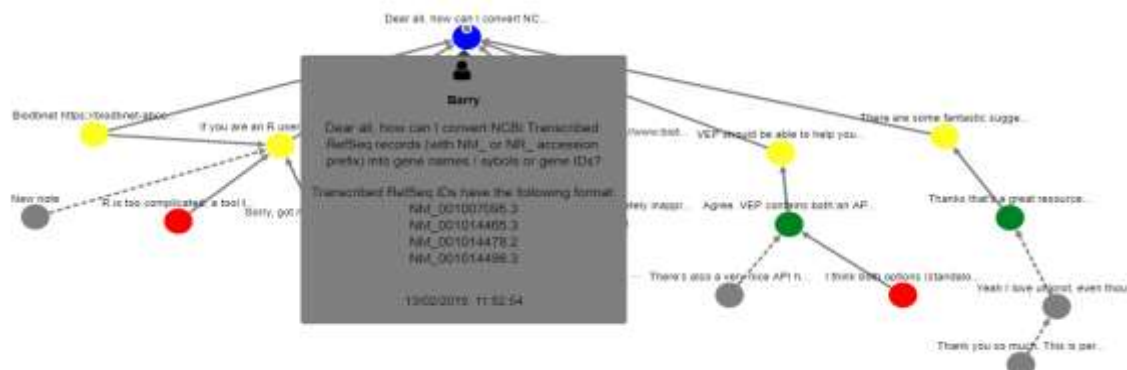
**Σχήμα 59.** Επεξεργασία κόμβου.

Κάθε κόμβος στον γράφο επιχειρηματολογίας διαθέτει μία *περιγραφή* (description). Ο δημιουργός του κόμβου έχει την δυνατότητα να επεξεργαστεί/διαγράψει/προσθέσει αυτήν την περιγραφή μέσω ενός ενσωματωμένου επεξεργαστή κειμένου που παρέχεται από το σύστημα. Ο επεξεργαστής κειμένου βοηθάει τον χρήστη παρέχοντάς του συντομεύσεις όπως την προσθήκη link στο κείμενο, την προσθήκη bold/italic χαρακτήρων, καθώς και την προσθήκη επικεφαλίδων αλλά και κομματιών κώδικα στο κείμενο. Οι προαναφερθείσες λειτουργίες/συντομεύσεις απεικονίζονται στην Σχήμα 60.



Σχήμα 60. Επεξεργαστής κειμένου του συνεργατικού περιβάλλοντος

Η ευχρηστία του συστήματος είναι ιδιαίτερα σημαντική. Έτσι, η τρέχουσα υλοποίηση του συνεργατικού περιβάλλοντος παρέχει στο χρήστη διάφορες συντομεύσεις (π.χ. αντιγραφή-επικόλληση ενός κόμβου). Μία από τις συντομεύσεις παρουσιάζεται στο Σχήμα 61. Ο χρήστης έχει την δυνατότητα να λάβει σύντομες πληροφορίες για έναν κόμβο, όπως τον δημιουργό του, την περιγραφή και την ημερομηνία δημιουργίας του. Το tooltip που παρουσιάζεται στο Σχήμα 61 παρέχει τις παραπάνω πληροφορίες. Το tooltip εμφανίζεται όταν ο χρήστης μετακινήσει τον δείκτη (ποντίκι) του υπολογιστή πάνω σε κάποιον συγκεκριμένο κόμβο (on hover event).



**Σχήμα 61.** Παροχή σύντομων πληροφοριών που αφορούν έναν κόμβο του γράφου



Τα χαρακτηριστικά του συνεργατικού συστήματος που αναμένεται να υλοποιηθούν και να βελτιωθούν κατά τη συνέχεια του έργου είναι:

- ✓ Αποθήκευση του γράφου σε graph database (Neo4J), (Miller, 2013)
- ✓ Δημιουργία μηχανής αναζήτησης χρησιμοποιώντας φίλτρα (search discourse graph by user, by node type)
- ✓ Εμφάνιση / κρύψιμο κόμβων
- ✓ Expand/Collapse σύνθετων κόμβων (compound node)
- ✓ Ορισμός κανόνων που αφορούν τις ακμές του γράφου
- ✓ Σημαντικότητα/Like/Dislike ενός κόμβου του γράφου
- ✓ Βελτίωση των γραφικών και της διεπαφής του γράφου με τον χρήστη
- ✓ Αλλαγή τύπου ενός γράφου (π.χ. από note σε solution)
- ✓ Containerization του κώδικα σε docker container

## 7 ΥΠΟΛΟΓΙΣΤΙΚΗ ΥΠΟΔΟΜΗ ΤΟΥ OPENBIO-C: ΑΡΧΙΚΕΣ ΛΥΣΕΙΣ & ΥΠΟΔΟΜΗ

Όλη η λειτουργικά που παρουσιάστηκε στα παραπάνω κεφάλαια υποστηρίζεται από τον επιχειρηματικό εταίρο του έργου Κωνσταντίνο Μουμούρη (ΚΜ). Συγκεκριμένα η εταιρεία ΚΜ έχει συμβάλει σημαντικά και έχει αφιερώσει πόρους στη παροχή των παρακάτω υπηρεσιών:

- Αγορά του domain [openbio.eu](http://openbio.eu)
- Αγορά και εγκατάσταση του SSH πιστοποιητικού του ιστότοπου του έργου.
- Διαχείριση της ονοματοδοσία (DNS) του ιστότοπου. Συγκεκριμένα ο ιστότοπος του OpenBio-C αποτελείται από τους παρακάτω υποτομείς:
  - a. [www.openbio.eu](http://www.openbio.eu). Ο κύριος ιστότοπος του συστήματος, διαθέσιμος προς όλους τους χρήστες. Περιέχει γενικότερες πληροφορίες, τους φορείς του έργου, τον φορέα χρηματοδότησης, νέα και φόρμα για επικοινωνία.
  - b. [staging.openbio.eu](http://staging.openbio.eu). Η δοκιμαστική έκδοση του συστήματος η οποία είναι προσιτή μόνο από τους προγραμματιστές και σχεδιαστές της πλατφόρμας OpenBio-C. Όλες οι νέες δυνατότητες της πλατφόρμας δοκιμάζονται αρχικά σε αυτόν τον υποτομέα πριν μεταφερθούν στον κύριο.
  - c. Σε μελλοντική έκδοση και αφού η ανάπτυξη του συστήματος φτάσει στο επίπεδο για δημόσια χρήση τότε αυτή θα ενσωματωθεί στον κύριο υποτομέα.
- Εγκατάσταση και υποστήριξη του βασικού διαδικτυακού Apache2 εξυπηρετητή (server) του OpenBio-C.
- Διαχείριση του εξυπηρετητή ηλεκτρονικού ταχυδρομείου.
- Επιπλέον, η ΚΜ υποστηρίζει το περιβάλλον δοκιμής και εκτέλεσης των ροών εργασίας. Το περιβάλλον αυτό αποτελείται από δύο εξυπηρετητές στους οποίους έχει εγκατασταθεί το λογισμικό για δημιουργία εικονικών υπολογιστικών περιβαλλόντων Docker. Το περιβάλλον αυτό βρίσκεται υπο ανάπτυξη.
- Τέλος, η ΚΜ έχει παράσχει πολύτιμες συμβουλευτικές υπηρεσίες όσον αφορά στη τεχνική διαμόρφωση του ιστότοπου, το ορθολογική χρήση και διαχείριση των υπολογιστικών πόρων, καθώς και την ασφάλεια του ιστότοπου.

Η υπολογιστική υποδομή του OpenBio-C θα αναπτυχθεί και θα οργανωθεί περαιτέρω τους επόμενους μήνες εξέλιξης του έργου με κύρια κατεύθυνση την εγκατάσταση και λειτουργία υπηρεσιών cloud για την εκτέλεση εικονοποιημένων επιστημονικών ροών εργασίας (virtualization services).

## 8 ΣΥΜΠΕΡΑΣΜΑΤΑ & ΜΕΛΛΟΝΤΙΚΕΣ ΔΡΑΣΕΙΣ

Στο παρόν παραδοτέο εστιάσαμε: (α) στην ενδελεχή μελέτη των τεχνολογιών αιχμής στο πεδίο των συστημάτων διαχείρισης βιοπληροφορικών ροών εργασίας; (β) στη καταγραφή των απαιτήσεων των εμπλεκόμενων επιστημονικών κοινοτήτων (ερευνητών μεταγονιδιωμιατικής βιοϊατρικής και βιοπληροφορικής) και (γ) στις προδιαγραφές ενός ολοκληρωμένου και φιλικού προς το χρήστη περιβάλλοντος σύνθεσης και εκτέλεσης βιοπληροφορικών επιστημονικών ροών εργασίας.



Όπως κατέδειξε η ανάλυση των απαντήσεων στο ερωτηματολόγιο του OpenBio-C, οι απαιτήσεις λειτουργικότητας και υπηρεσιών ενός τέτοιου περιβάλλοντος είναι ιδιαίτερα διακριτές. Το υψηλό μορφωτικό επίπεδο των δυνητικών χρηστών και η εξοικείωση τους με σχετικά συστατικά λογισμικού διαχείρισης και επεξεργασίας βιο-δεδομένων συνθέτουν ένα ιδιαίτερα απαιτητικό πλαίσιο προδιαγραφών.

- Από τη μία πλευρά, η ενσωμάτωση, η διασύνδεση και η διαλειτουργικότητα καθιερωμένων εργαλείων/συστημάτων ανάλυσης ετερογενών μεταγονιδιωματικών βιο-δεδομένων και από την άλλη, η ανάγκη υπηρεσιών δικτύωσης και συνεργατικότητας μεταξύ των εμπλεκόμενων ερευνητικών ομάδων και ανθρώπων, προδιαγράφουν τις βασικές κατευθύνσεις σχεδίασης και ανάπτυξης του OpenBio-C.
- Μηχανισμοί τεκμηρίωσης (annotation), συνεχούς επικύρωσης (validation) αποδοτικής εκτέλεσης και διαμοιρασμού (sharing) επιστημονικών ροών εργασίας αποτελούν απαραίτητες προϋποθέσεις για την επίτευξη αναπαραγωγίμων (reproducible) επιστημονικών ευρημάτων. Η σχεδίαση και η αρχική υλοποίηση βασικών συστατικών του OpenBio-C προσπαθούν να τηρήσουν αυτές τις προδιαγραφές.

Με βάση τα προαναφερθέντα, μπορούμε να θέσουμε και να συνοψίσουμε το γενικότερο πλαίσιο στοχεύσεων του περιβάλλοντος OpenBio-C, το οποίο προδιαγράφει και τις σχετικές δρ'στηριότητες και δράσεις του έργου.

- ❖ **Υποστήριξη και ενίσχυση μεθοδολογιών “ανοιχτής επιστήμης”.** Οι θεμελιώδεις επιδιώξεις του συνεργατικού περιβάλλοντος διαχείρισης, σύνθεσης και εκτέλεσης βιοπληροφορικών επιστημονικών ροών εργασίας OpenBio-C είναι: (i) η ενθάρρυνση και η δημιουργία κινήτρων σε ερευνητές της μεταγονιδιωματικής βιοϊατρικής να ακολουθήσουν μεθοδολογίες ανοιχτής-επιστήμης έτσι ώστε να διασφαλιστεί η ποιότητα, ο αντίκτυπος και η ακεραιότητα της σχετικής έρευνας έτσι ώστε να υποστηριχθεί η αξιοποίηση αξιόπιστων και επικυρωμένων αποτελεσμάτων; (ii) η απαλλαγή των επιστημόνων βιοπληροφορικής από “κλειστά” περιβάλλοντα ανάλυσης δεδομένων προσφέροντας μια πλατφόρμα που υποστηρίζει, ενισχύει και ανταμείβει τη συνεργατική και ποιοτική έρευνα, και (iii) η υποστήριξη της “ανοιχτής” πρόσβασης αλλά και της δυνατότητας επαναχρησιμοποίησης των αναγκαίων, πιο πρόσφορων και επαρκώς αξιολογημένων πηγών δεδομένων, μεθόδων και εργαλείων ανάλυσης τους και υπολογιστικών περιβαλλόντων υψηλής απόδοσης. Με το βλέμμα όχι μόνο στο εσωτερικό αλλά στο διεθνές περιβάλλον, αποβλέπουμε στο συντονισμό του OpenBio-C με το *European Open Science Cloud* (EOSC)<sup>71</sup>.
- ❖ **Παροχή ολοκληρωμένων και διασυνδεδεμένων υπηρεσιών βιοπληροφορικής.** Με στόχο την υποστήριξη της αναπαραγωγίμης (reproducible) επιστήμης, το OpenBio-C αξιοποιεί υπάρχουσες τεχνολογίες αιχμής αποβλέποντας στην ανάπτυξη ενός καινοτόμου περιβάλλοντος ομαδικής και χωρίς αποκλεισμούς (inclusive) εργασίας για τη διαχείριση και ανακάλυψη αξιολογημένων και επικυρωμένων γνώσεων στο πεδίο της μεταγονιδιωματικής βιοϊατρικής έρευνας. Σε αυτό το πλαίσιο, οι σχεδιαστικές αρχές αλλά και οι λύσεις που έχουν δοθεί στην αρχική ανάπτυξη του περιβάλλοντος επιτρέπουν την εύκολη σύνθεση – μέσω ενός διαλειτουργικού (μεταξύ των χρησιμοποιούμενων αναλυτικών εργαλείων) γραφικού περιβάλλοντος, και την εκτέλεση βιοπληροφορικών επιστημονικών ροών εργασίας. Η εκμετάλλευση καθιερωμένων οντολογιών και σχετικών σημασιολογικών περιγραφών των εμπλεκόμενων σε μια ροή ερευνητικών-αντικειμένων, επιτρέπουν στο OpenBio-C να καταστεί ένα περιβάλλον όπου όχι μόνο τα χρησιμοποιούμενα αντικείμενα αλλά και οι ακολουθούμενες ροές εργασιών, οι εμπλεκόμενοι ερευνητές και ομάδες καθώς και τα παραγόμενα αποτελέσματα να είναι ανιχνεύσιμα (findable), προσβάσιμα (accessible), διαλειτουργικά (interoperable) και επαναχρησιμοποιήσιμα (reusable). Έτσι, το OpenBio-C θα μπορέσει να καταστεί ένα από τα παραδείγματα υλοποίησης FAIR-συμβατών συστημάτων – Findable, Accessible, Interoperable, Re-usable (Wilkinson *et al.*, 2016). Ταυτόχρονα, η παροχή υπηρεσιών παραγωγής εικονοποιημένων ροών εργασίας (visualized workflows) για την εκμετάλλευση κατανεμημένων cloud συστημάτων και υπηρεσιών, δημιουργεί ένα ολοκληρωμένο περιβάλλον με βάση το οποίο οι παραγόμενες λύσεις και αποτελέσματα μπορούν εύκολα να διασυνδεθούν με εξωτερικά/εξωγενή περιβάλλοντα.

<sup>71</sup> [ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud](https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud)

- ❖ **Υποστήριξη συνεργειών και συνεργασίας.** Το OpenBio-C δεν αποτελεί μια ακόμη πλατφόρμα σύνθεσης και εκτέλεσης επιστημονικών ροών εργασίας. Αντίθετα, βασίζεται στη *συνέργεια* των εμπλεκόμενων επιστημόνων, με ιδιαίτερη αναφορά σε βιοπληροφορικούς και ερευνητές μεταγονιδιωμιατικής βιοϊατρικής, αποσκοπώντας στη *παροχή 'νοήμων' υπηρεσιών συνεργατικότητας για τον εντοπισμό των πιο πρόσφορων πόρων, πηγών και λύσεων κατά τη συλλογή και ανάλυση δεδομένων αλλά και κατά την ερμηνεία και εκμετάλλευση των παραγόμενων αποτελεσμάτων*. Η προσπάθεια και η στόχευση είναι να εξασφαλιστεί η χρηστικότητα και η αποδοχή του περιβάλλοντος και των υπηρεσιών που παρέχει μέσω της επικύρωσης τους σε πραγματικά σενάρια συνεργατικής ανάλυσης μεταγονιδιωμιακών βιοϊατρικών δεδομένων. Στο πλαίσιο αυτό, η ενσωμάτωση στο OpenBio-C καθιερωμένων πόρων στο πεδίο της μεταγονιδιωμιατικής – από επικυρωμένα μεταγονιδιωμιατικές πηγές δεδομένων και γνώσεων έως αξιόπιστες, αξιολογημένες και ευρέως διαδεδομένες τεχνικές ανάλυσης, αλλά και η παροχή ενός εύχρηστου περιβάλλοντος επιχειρηματολόγησης επί των χρησιμοποιούμενων πόρων και παραγόμενων αποτελεσμάτων. Οι προαναφερθείσες υπηρεσίες και λειτουργικότητα επιτρέπει στο OpenBio-C στο να δημιουργήσει ένα *οικοσύστημα επιστημονικής έρευνας*, βασισμένης σε ανοιχτά πρότυπα και σε εύκολα ενσωματώσιμα εργαλεία και τεχνικές ανάλυσης.
- ❖ **Ενίσχυση μεταφραστικής έρευνας.** Είναι αναγκαίο, κυρίως για την αποδοχή του περιβάλλοντος OpenBio-C, η σχεδίαση συγκεκριμένων *σεναρίων χρήσης* τους σε πραγματικά ερωτήματα μεταγονιδιωμιατικής βιοϊατρικής έρευνας καθώς και στην εξορθολογισμένη χρήση των απαραίτητων *γνωσιακών πόρων*. Αυτό θα βοηθήσει όχι μόνο τη προσέλκυση επιστημόνων και ερευνητών του πεδίου στο να χρησιμοποιήσουν το περιβάλλον αλλά και στην *υποστήριξη μεθοδολογιών μεταφραστικής έρευνας* με βάση την υλοποίηση, εύκολη προσαρμογή και τεκμηρίωση ροών εργασίας οι οποίες ξεκινούν από το ερώτημα και καταλήγουν σε αναπαραγώγιμα και έτσι τεκμηριωμένα αποτελέσματα τα οποία έχουν επαρκή υποστήριξη για τη κλινική τους αξιοποίηση.
- ❖ **Αξιοποιήσιμη και βιώσιμη εφαρμογή.** Υιοθετώντας προτάσεις και οδηγίες της Ευρωπαϊκής Ένωσης (ΕΕ) για τη δημιουργία, την ολοκλήρωση και τη παροχή υπηρεσιών e-science, προτιθέμεθα να αξιοποιήσουμε το OpenBio-C στα πλαίσια σχετικών Ευρωπαϊκών ερευνητικών ηλεκτρονικών υποδομών και να ενσωματώσουμε το περιβάλλον στο υπό-ανάπτυξη ευρωπαϊκό *Open Science Cloud*<sup>72</sup>. Σε αυτή τη κατεύθυνση, και με στόχευση τη μακροπρόθεσμη βιωσιμότητα του OpenBio-C, ιδιαίτερες δράσεις θα αφιερωθούν στη *διακυβέρνηση* και στο συντονισμό των δράσεων του έργου και στη διάδοση των υπηρεσιών του σε σχετικές Ευρωπαϊκές ερευνητικές υποδομές στο πεδίο της επιστήμης *βιο-δεδομένων* (life sciences data), όπως οι υποδομές *ELIXIR*<sup>73</sup> και *BBMRI*<sup>74</sup>, συμβάλλοντας στη διαλειτουργικότητα των εμπλεκόμενων συστατικών λογισμικού και βιο-δεδομένων (π.χ., πρωτοβουλία της ΕΕ *Turning FAIR into reality*<sup>75</sup>). Αναγνωρίζοντας ότι η ανάπτυξη και η διαχείριση ερευνητικών υποδομών δεν αποτελεί μόνο μια τεχνολογική διαδικασία αλλά διακρίνεται και από κοινωνικές και οργανωτικές διαστάσεις, στην επόμενη φάση εξέλιξης του έργου θα εξετάσουμε και θα μελετήσουμε λύσεις αξιοποίησης του περιβάλλοντος ως μιας ερευνητικής υποδομής η οποία μπορεί να προσφέρει διαδικτυακές υπηρεσίες βιοπληροφορικής '*ανοιχτής πρόσβασης*' όχι μόνο στην ακαδημαϊκή αλλά και στην ιδιωτική επιχειρηματική κοινότητα.
- ❖ **Ζητήματα βιοηθικής και ιδιωτικότητας.** Τέλος, αν και δεν άπτεται άμεσα με τις λειτουργίες ενός ολοκληρωμένου περιβάλλοντος βιοπληροφορικής, θα εξεταστούν όλα τα σχετικά θέματα *βιοηθικής* τα οποία εμπλέκονται στην ανάλυση βιο-δεδομένων. Σκοπός είναι να παρασχεθούν εκείνες οι υπηρεσίες οι οποίες διασφαλίζουν την *ασφάλεια* και την *ιδιωτικότητα* προσωπικών δεδομένων και εξασφαλίζουν τη συμβατότητα των σχετικών λειτουργιών με το *Γενικό Κανονισμό Προστασίας Δεδομένων* της ΕΕ<sup>76</sup> (όταν εμπλέκονται τέτοιας φύσης δεδομένα) αλλά και η *εμπειριστατωμένη ερμηνεία* και κατάλληλη διασπορά των παραγόμενων αποτελεσμάτων.

<sup>72</sup> [cordis.europa.eu/project/rcn/216096/factsheet/en](https://cordis.europa.eu/project/rcn/216096/factsheet/en)

<sup>73</sup> [www.elixir-europe.org](http://www.elixir-europe.org)

<sup>74</sup> [www.bbmr-eric.eu](http://www.bbmr-eric.eu)

<sup>75</sup> [publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283](https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283)

<sup>76</sup> [www.ekdd.gr/ekdda/images/seminaria/GDPR.pdf](http://www.ekdd.gr/ekdda/images/seminaria/GDPR.pdf)

## ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

- van der Aalst, W.M.P. and ter Hofstede, A.H.M. (2005) YAWL: yet another workflow language. *Inf. Syst.*, **30**, 245–275.
- Abouelhoda, M. *et al.* (2012) Tavaxy: Integrating Taverna and Galaxy workflows with cloud computing support. *BMC Bioinformatics*, **13**, 77.
- Adzhubei, I. *et al.* (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, **Chapter 7**, Unit7.20–Unit7.20.
- Anderson, M.S. *et al.* (2007) Normative dissonance in science: results from a national survey of u.s. Scientists. *J. Empir. Res. Hum. Res. Ethics*, **2**, 3–14.
- Athey, B.D. *et al.* (2013) TranSMART: An open source and community-driven informatics and data sharing platform for clinical and translational research. *Proc. AMIA Jt. Summits Transl. Sci.* **2013**, **2013**, 6–8.
- Badia, R.M. *et al.* (2017) Workflows for Science: a Challenge when Facing the Convergence of HPC and Big Data. *Supercomput. Front. Innov.*, **4**.
- Bandrowski, A.E. *et al.* (2014) OMICtools: an informative directory for multi-omic data analysis. *Database*, **2014**.
- Boettiger, C. (2015) An introduction to Docker for reproducible research. *ACM SIGOPS Oper. Syst. Rev.*, **49**, 71–79.
- Brazas, M.D. *et al.* (2010) Providing web servers and training in Bioinformatics: 2010 update on the Bioinformatics Links Directory. *Nucleic Acids Res.*, **38**, W3–W6.
- Byelas, H. *et al.* (2011) Towards a MOLGENIS Based Computational Framework. In, *2011 19th International Euromicro Conference on Parallel, Distributed and Network-Based Processing.*, pp. 331–338.
- Byelas, H. V *et al.* (2012) Introducing data provenance and error handling for NGS workflows within the molgenis computational framework. In, *BIOINFORMATICS 2012 - Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms.*, pp. 42–50.
- Byrne, M. *et al.* (2012) VarioML framework for comprehensive variation data representation and exchange. *BMC Bioinformatics*, **13**, 254.
- Carusi, A. and Reimer, T. (2010) Virtual Research Environment Collaborative Landscape Study. *Differences*, **94**, 106.
- Chen, Y. long and Yang, K. hu (2009) Avoidable waste in the production and reporting of evidence. *Lancet*, **374**, 786.
- Cingolani, P. *et al.* (2015) BigDataScript: a scripting language for data pipelines. *Bioinformatics*, **31**, 10–16.
- Craswell, N. *et al.* (2018) Neural information retrieval: introduction to the special issue. *Inf. Retr. J.*, **21**, 107–110.
- DePristo, M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491.
- Dogrusoz, U. *et al.* (2018) Efficient methods and readily customizable libraries for managing complexity of large networks. *PLoS One*, **13**, e0197238–e0197238.
- van Eemeren, F.H. *et al.* (1996) Fundamentals of Argumentation Theory: A Handbook of Historical Backgrounds and Contemporary Developments Lawrence Erlbaum Associates.
- Fisch, K.M. *et al.* (2015) Omics Pipe: a community-based framework for reproducible multi-omics data analysis. *Bioinformatics*, **31**, 1724–1728.
- Franz, M. *et al.* (2016) Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, **32**, 309–311.
- Galperin, M.Y. *et al.* (2017) The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. *Nucleic Acids Res.*, **45**, D1–D11.
- Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
- Giardine, B. *et al.* (2005) Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
- Goble, C.A. *et al.* (2010) myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res.*, **38**, W677–W682.
- Goodman, A. *et al.* (2014) Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Comput Biol*, **10**, e1003542+.
- Guimera, R. (2012) bcbio-nextgen: Automated, distributed next-gen sequencing pipeline. *EMBnet.journal*, **17**, 30.
- Harris, N.L. *et al.* (2016) The 2015 Bioinformatics Open Source Conference (BOSC 2015). *PLOS Comput. Biol.*, **12**, e1004691.
- He, K.Y. *et al.* (2017) Big Data Analytics for Genomic Medicine. *Int. J. Mol. Sci.*, **18**, 412.
- Hettne, K. *et al.* (2012) Workflow Forever: Semantic Web Semantic Models and Tools for Preserving and Digitally Publishing Computational Experiments. In, *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences, SWAT4LS '11*. ACM, New York, NY, USA, pp. 36–37.
- Hettne, K.M. *et al.* (2014) Structuring research methods and data with the research object model: genomics workflows

- as a case study. *J. Biomed. Semantics*, **5**, 41.
- Ioannidis, J.P.A. (2005) Why most published research findings are false. *PLoS Med.*, **2**, 0696–0701.
- Ison, J. *et al.* (2013) EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, **29**, 1325–1332.
- Ison, J. *et al.* (2016) Tools and data services registry: A community effort to document bioinformatics resources. *Nucleic Acids Res.*, **44**, D38–D47.
- Juve, G. and Deelman, E. (2011) Scientific Workflows in the Cloud - Grids, Clouds and Virtualization. In, Cafaro, M. and Aloisio, G. (eds). Springer London, London, pp. 71–91.
- Kallio, M.A. *et al.* (2011) Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*, **12**, 507.
- Kanterakis, A. *et al.* (2015) Molgenis-impute: imputation pipeline in a box. *BMC Res. Notes*, **8**, 359.
- Karacapilidis, N. *et al.* (2012) Mastering data-intensive collaboration through the synergy of human and machine reasoning. In, *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW.*, pp. 21–22.
- Karacapilidis, N. (2014) Mastering Data-Intensive Collaboration and Decision Making: Research and Practical Applications in the Dicode Project Springer Publishing Company, Incorporated.
- Karacapilidis, N. *et al.* (2009) Tackling Cognitively-Complex Collaboration with CoPe\_it! *Int. J. Web-Based Learn. Teach. Technol.*, **4**, 22–38.
- Karacapilidis, N. and Potamias, G. (2016) Towards a Sustainable Solution for Collaborative Healthcare Research BT - Innovation in Medicine and Healthcare 2016. In, Chen, Y.-W. *et al.* (eds). Springer International Publishing, Cham, pp. 159–170.
- Karacapilidis, N. and *et al.* CoPe\_it! - Supporting collaboration, enhancing learning. In, Ditsa, G. *et al.* (eds), *Proceedings of the 2009 International Conference on Information Resources Management (Conf-IRM 2009)*.
- Karim, M.R. *et al.* (2017) Improving data workflow systems with cloud services and use of open data for bioinformatics research. *Brief. Bioinform.*, **19**, 1035–1050.
- Krieger, M.T. *et al.* (2017) Building an open source cloud environment with auto-scaling resources for executing bioinformatics and biomedical workflows. *Futur. Gener. Comput. Syst.*, **67**, 329–340.
- Kulkarni, N. *et al.* (2018) Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics*, **19**, 349.
- Kunz, W. and Rittel, H.W.J. (1970) Issues as Elements of Information Systems Institute of Urban and Regional Development, University of California.
- Leipzig, J. (2016) A review of bioinformatic pipeline frameworks. *Brief. Bioinform.*, **18**, 530–536.
- Lek, M. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
- Leung, W. *et al.* (2015) Drosophila Muller F Elements Maintain a Distinct Set of Genomic Properties Over 40 Million Years of Evolution. *G3 (Bethesda)*, **5**, 1–39.
- Levin, L.A. and Danesh-Meyer, H. V. (2010) Lost in translation: Bumps in the road between bench and bedside. *JAMA - J. Am. Med. Assoc.*, **303**, 1533–1534.
- Li, P.-E. *et al.* (2017) Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Res.*, **45**, 67–80.
- Liddo, A. De and Shum, S.B. (2010) Cohere: A prototype for contested collective intelligence. In, *ACM Computer Supported Cooperative Work (CSCW 2010) - Workshop: Collective Intelligence In Organizations - Toward a Research Agenda*.
- Malone, J. *et al.* (2016) Ten Simple Rules for Selecting a Bio-ontology. *PLoS Comput. Biol.*, **12**, e1004743–e1004743.
- Malone, J. *et al.* (2014) The Software Ontology (SWO): a resource for reproducibility in biomedical data analysis, curation and digital preservation. *J. Biomed. Semantics*, **5**, 25.
- Marx, V. (2013) Biology: The big challenges of big data. *Nature*, **498**, 255–260.
- Miller, J.J. (2013) Graph Database Applications and Concepts with Neo4j.
- Mimori, T. *et al.* (2013) iSPV: an integrated structural variant calling pipeline from high-throughput sequencing data. *BMC Syst. Biol.*, **7 Suppl 6**, S8–S8.
- Mitra, B. and Craswell, N. (2017) Neural Models for Information Retrieval. *CoRR*, **abs/1705.01509**.
- Moreews, F. *et al.* (2015) A curated Domain centric shared Docker registry linked to the Galaxy toolshed. In, *Galaxy Community Conference 2015*. Norwich, United Kingdom.
- Mu, J.C. *et al.* (2015) VarSim: a high-fidelity simulation and validation framework for high-throughput genome



- sequencing with cancer applications. *Bioinformatics*, **31**, 1469–1471.
- Munafò, M.R. *et al.* (2017) A manifesto for reproducible science. *Nat. Hum. Behav.*, **1**, 21.
- Paila, U. *et al.* (2013) GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Comput. Biol.*, **9**, e1003153–e1003153.
- Pang, C. *et al.* (2015) BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. *J. Am. Med. Inform. Assoc.*, **22**, 65–75.
- Perkel, J.M. (2017) How bioinformatics tools are bringing genetic analysis to the masses. *Nature*, **543**, 137–138.
- Poste, G. (2011) Bring on the biomarkers. *Nature*, **469**, 156–157.
- Purcell, S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
- Robinson, P.N. and Mundlos, S. (2010) The Human Phenotype Ontology. *Clin. Genet.*, **77**, 525–534.
- Sadedin, S.P. *et al.* (2012) Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics*, **28**, 1525–1526.
- Samocha, K.E. *et al.* (2014) A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.*, **46**, 944–950.
- Scheuer, O. *et al.* (2010) Computer-supported argumentation: A review of the state of the art. *Int. J. Comput. Collab. Learn.*, **5**, 43–102.
- Schneider, G. *et al.* (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.*, **40**, W452–W457.
- Shah, N. *et al.* (2016) A crowdsourcing approach for reusing and meta-analyzing gene expression data. *Nat. Biotechnol.*, **34**, 803–806.
- Smiley, D. *et al.* (2015) Apache Solr Enterprise Search Server - Third Edition Packt Publishing.
- Smith, C. and Sotala, K. (2011) Knowledge, networks and nations Global scientific collaboration in the 21st century.
- Spjuth, O. *et al.* (2015) Experiences with workflows for automating data-intensive bioinformatics. *Biol. Direct*, **10**, 1–12.
- Spjuth, O. *et al.* (2016) Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research. *Eur. J. Hum. Genet.*, **24**, 521–528.
- Swertz, M.A. *et al.* (2010) The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics*, **11**, S12.
- The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
- Tomczak, K. *et al.* (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.*, **19**, A68–A77.
- Di Tommaso, P. *et al.* (2015) The impact of Docker containers on the performance of genomic pipelines. *PeerJ*, **3**, e1273–e1273.
- Torres-Garcia, W. *et al.* (2014) PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics*, **30**, 2224–2226.
- Tzarakis, M. *et al.* (2014) The Dicode Collaboration and Decision Making Support Services BT - Mastering Data-Intensive Collaboration and Decision Making: Research and practical applications in the Dicode project. In, Karacapilidis, N. (ed). Springer International Publishing, Cham, pp. 119–139.
- Uzuner, Ö. *et al.* (2011) 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inform. Assoc.*, **18**, 552–556.
- Vuong, H. *et al.* (2015) AVIA v2.0: annotation, visualization and impact analysis of genomic variants and genes. *Bioinformatics*, **31**, 2748–2750.
- Wang, K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
- Wilkinson, M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
- Wolstencroft, K. *et al.* (2013) The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res.*, **41**, W557–W561.
- Χριστοδούλου, Σ. (2015) Καινοτόμες Υπηρεσίες Υποστήριξης Συνεργασίας και Ομαδικής Λήψης Αποφάσεων σε Περιβάλλοντα Υπερφόρτωσης Πληροφοριών και Γνωστικής Πολυπλοκότητας. In, *Διακριτική Διατριβή, Πανεπιστήμιο Πατρών*.