

BIOLOGICAL SCIENCES SOFTWARE ENVIRONMENT

ПРОГРАМНЕ СЕРЕДОВИЩЕ БІОЛОГІЧНИХ НАУК

Igor Polsky¹

DOI: <https://doi.org/10.30525/978-9934-26-536-5-2>

Робота сучасного біолога залежить від багатьох цифрових систем. Збір та обробка даних спираються на програмні пакети, документація формується та публікується в цифровому форматі. На жаль, на даний момент технічні знання, необхідні для оцінки якості програмного забезпечення, пов'язані зі спеціалізованими навичками, недостатньо розповсюдженими серед біологів. Через це програмне середовище, з яким працюють біологи України, страждає від ряду недоліків.

По-перше, середовище страждає від фрагментації. Дослідницьке обладнання генерує дані в різних форматах, зазвичай у вигляді зображень або форматowanego тексту. Згенеровані дані піддаються статистичному аналізу за допомогою різних програмних пакетів: Statistica, Mathematica, Matlab, Origin та інші, спеціалізовані для конкретних галузей, на кшталт SPM, Neuron, EEGLAB тощо. Більшість з цих пакетів має власний формат даних, несумісний з іншими, навіть коли структура даних не є спеціалізованою.

Наприклад, розповсюдженим форматом даних є таблиця, де дані згруповані в іменовані стовпці. Таким форматом оперує Statistica, в нього ж часто збираються результати вимірів. Часто такі дані зберігаються та передаються як таблиці в форматі Microsoft Excel, оскільки він дозволяє групувати дані в стовпці та копіювати чисельні дані як текст. Буфери обміну сучасних операційних систем дозволяють копіювати та вставляти такі табличні дані, не втрачаючи структури. Однак через різницю в роботі Excel та Statistica, просте копіювання призводить до некоректних результатів, змушуючи оператора вручну формувати дані, витрачаючи час та ризикуючи зробити механічну помилку. Крім цього, автоматичне форматування даних в обох програмах може призводити до непередбачуваних викривлень та помилок в обрахунках.

Продовженням проблеми фрагментації є проблема недостатньої автоматизації. Майже будь-яка нетривіальна обробка даних залучає декілька різних програм. Спеціалізовані програми, на відміну від універсальних пакетів, мають стандартні формати (напр. FASTA), але

¹ Taras Shevchenko National University of Kyiv, Ukraine

страждають від незначних відмінностей в їх обробці. Наприклад, деякі програми здатні працювати з декількома послідовностями в форматі FASTA в одному файлі, а інші потребують завантаження окремих файлів. Значна частина цих програм доступна в першу чергу як веб-сторінки, що дозволяє швидко почати роботу, але перетворює роботу з послідовністю декількох програм досить громіздкою.

Іншою проблемою є залежність біологічних досліджень від пропріетарного програмного забезпечення. Комерційні ліцензії мають значну ціну та вимагають процесу аутентифікації, що часто стикається з технічними проблемами, витрачаючи час та зусилля операторів. Хоча процеси сертифікації дають міру впевненості в якості програмного забезпечення, жодна складна програма не обходиться без помилок. У разі непередбаченої помилки, оператори пропріетарних програм змушені спиратись на підтримку розробника, яка не завжди є доступною чи швидкою. Ті ж проблеми виникають, коли для виконання дослідження потрібна модифікація програм.

Публікації зазвичай відправляються в форматі Microsoft Word та мають відповідати стандартному для публікації форматуванню, що може різнитись від журналу до журналу. Цей формат погано підходить для цього завдання, оскільки він містить в собі повне форматування. Через це оператор часто змушений вручну налаштовувати редактор, аби отримати потрібне форматування.

Ці фактори призводять до того, що сучасні українські біологи витрачають багато часу та зусиль на монотонну роботу, продиктовану недоліками програмного забезпечення. Для вирішення цієї проблеми та формування кращого наукового простору слід сформувати нове програмне середовище.

В основі такого середовища можна покласти стандартизовані формати обміну даними на кшталт EDF. Вони мають бути достатньо простими та гнучкими, щоб розробники програмного забезпечення для нового середовища мали можливість легко реалізувати алгоритми читання та запису. Не для всіх форматів це можливо, але велика частина даних вже розповсюджується у вигляді форматowanego тексту, що дає надію на формування стандартного пакету форматів.

Продуктивне програмне середовище має містити мову програмування, яку можуть використовувати оператори. Оскільки операторами є біологи, мова програмування має бути легкою в освоєнні та використанні. Прикладами таких мов є Python та Lua. Сучасна популярність обох цих мов спирається на легкість використання людьми без технічної підготовки. Lua з самого початку розроблювали для вбудовування в інші

програми як інструменту автоматизації та інтеграції, що робить цю мову досить популярною для “скриптингу” та аматорського програмування.

Також потенціал мають візуальні середовища програмування. Програма тривимірного моделювання Blender містить декілька підсистем, які дозволяють користувачу генерувати прості програми, візуально поєднуючи “вузли”, що відображають етапи обробки моделі. Схожі системи можна розробити для обробки даних в біологічних дослідженнях.

Більшість частин програмного середовища біологів, доступних як веб-сервіси, вже мають можливість використання з інших програм за допомогою інтерфейсу програмування (API, application programming interface). Будь-яка програма, яка підтримує скриптинг за допомогою Lua або Python, може бути легко модифікована, щоб презентувати схожий інтерфейс, перетворюючи завдання швидкого виконання багатовісних процедур обробки даних в багатьох програмах з неможливого в майже тривіальне.

Покращення механізмів публікації можливе двома основними шляхами. По-перше, оскільки різниця між форматуванням в різних журналах є візуальною, а не семантичною, можна обрати формат, який передає лише семантику та структуру, та автоматично генерувати з нього форматований текст. Прикладом є markdown, родина розповсюджених текстових форматів, які використовують прості текстові директиви (наприклад, **курсивний** текст отримується за допомогою “зірочок” навколо фрази) для передачі структури документу: варіанти шрифту, заголовки, списки, посилання та зображення.

По-друге, сучасна наукова документація зазвичай розповсюджується та використовується на комп’ютерах. Це дозволяє відійти від використання статичного тексту та зображень, та додати інтерактивні елементи, що дозволяють швидше та більш зрозуміло передати складні дані. Наприклад, система Jupyter Notebooks здатна генерувати інтерактивні графіки, що реагують на команди користувача, фільтруючи відображені дані, змінюючи їх масштаб тощо. Також ця система дозволяє напряму маніпулювати даними за допомогою програм на мові Python. В разі необхідності, більшість реалізацій Jupyter здатні експортувати статичні HTML-файли, які можна використовувати для друку.

Модернізоване програмне середовище має спиратись на відкриті (Open Source) програми. На відміну від пропріетарних програм, які за своєю природою ізольовані, відкрита модель розробки програм заохочує співробітництво та обмін досвідом. Відкрита екосистема має набагато більший потенціал для розвитку та інтеграції, ніж замкнена.

На даний момент більша частина компонентів нового програмного середовища вже існують та мають певне розповсюдження – Jupyter та

Python активно використовуються для обробки даних серед технічно грамотних науковців. Головною запропонованою зміною є розширення викладання технічних навичок серед біологів, що дозволить зсунути фокус з архаїчних пропрієтарних монолітів на більш сучасні та гнучкі технології.