DOI 10.15826/B978-5-7996-3848-1.58

ОПТИМИЗАЦИЯ РАБОЧЕГО ПРОЦЕССА ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ В НАУЧНОЙ ЛАБОРАТОРИИ АСТРОХИМИЧЕСКИХ ИССЛЕДОВАНИЙ УрФУ

Р. С. Накибов, Г. С. Федосеев, В. М. Картеева, М. Г. Медведев, М. Э. Ожиганов, У. А. Сапунова, Э. Д. Кузнецов, А. И. Васюнин
Уральский федеральный университет

Повседневная работа лаборатории, в частности работа на инфракрасном Фурье-спектрометре, подразумевает генерацию большого количества сырых экспериментальных данных. Обработка, анализ данных и представление их в формате, доступном для интерпретации и коллективного обсуждения, является трудоемкой и многоступенчатой, но необходимой частью работы исследователя. В публикации описана проведенная работа по оптимизации и автоматизации типичных сценариев обработки данных в Научной лаборатории астрохимических исследований УрФУ с использованием языка программирования Python. Приводятся диаграмма рабочего процесса обработки экспериментальных данных, полученных с использованием инфракрасного Фурье-спектрометра, перечень написанных программ. Нами был написан набор приложений, за счет использования которого достигнуто существенное ускорение обработки данных и программно реализован стандарт каталогизации данных. Данный опыт может быть полезен при оптимизации работы других лабораторий.

EXPERIMENTAL DATA WORKFLOW OPTIMISATION IN THE URFU RESEARCH LABORATORY FOR ASTROCHEMISTRY

R. S. Nakibov, G. S. Fedoseev, V. M. Karteeva, M. G. Medvedev, M. E. Ozhiganov, U. A. Sapunova, E. D. Kuznetsov, A. I. Vasyunin
Ural Federal University

Day-to-day laboratory work, in particular, usage of the FT-IR spectrometer, generates vast amounts of raw experimental data. Processing and analysis of such data, its transformation into a format suited for interpretation and collective discussion is a tedious, multistage, yet necessary task. This report describes the work done to optimize and automate common data processing scenarios present in the work of Research Laboratory for Astrochemistry using the Python programming language. We provide a workflow chart for processing FT-IR spectrometer data and provide a list of developed apps. A set of applications was developed through the usage of which we achieved an increase of the speed of data processing and implemented a data cataloging standard on a software level. This experience can be used to improve the workflow of other laboratories.

Введение

В межзвездной среде, включая регионы активного звездообразования, обнаружено более 240 соединений [1]. Исследование состава, свойств межзвездной среды, характерных реакций служит построению исчерпывающей теории звездообразования и образования планетных систем. Помимо прочего интерес представляют сложные органические молекулы,

такие, как, например, бензол или метанол. Основными объектами, доступными для наблюдений, являются межзвездный газ, пыль и лед, состоящий из молекул, сконденсировавших на пылевые частицы — еще на стадии темного облака, например, большинство органических молекул оказываются вморожены в лед на их поверхности. Для идентификации новых соединений и выяснения такой информации, как состав окружения молекул или температура среды, в которой образуются льды, необходимо располагать информацией о полосах поглощения и эмиссии молекул, составляющих лед. На базе ИЕНиМ УрФУ развернута деятельность Научной лаборатории астрохимических исследований (НЛАИ УрФУ). Сотрудники лаборатории проводят астрономические наблюдения [2], ведут теоретическое сопровождение наблюдательных астрономических данных методами компьютерного моделирования [3]. Помимо этого создана экспериментальная установка, позволяющая выращивать аналоги космических льдов. На установке можно поддерживать условия, характерные для межзвездной среды: охлаждение главной камеры до 6.5 K, сверхвысокий вакуум вплоть до $5 \cdot 10^{-10}$ мбар.

Работа посвящена оптимизации рабочего процесса обработки экспериментальных данных. Целью являлось построение эффективной системы управления потоком экспериментальных данных (пайплайна): сбор, каталогизация, анализ, конвертирование, представление. На языке программирования Python был написан ряд консольных и графических приложений, оптимизирующих типичные сценарии обработки спектроскопических данных.

Оптимизация рабочего процесса

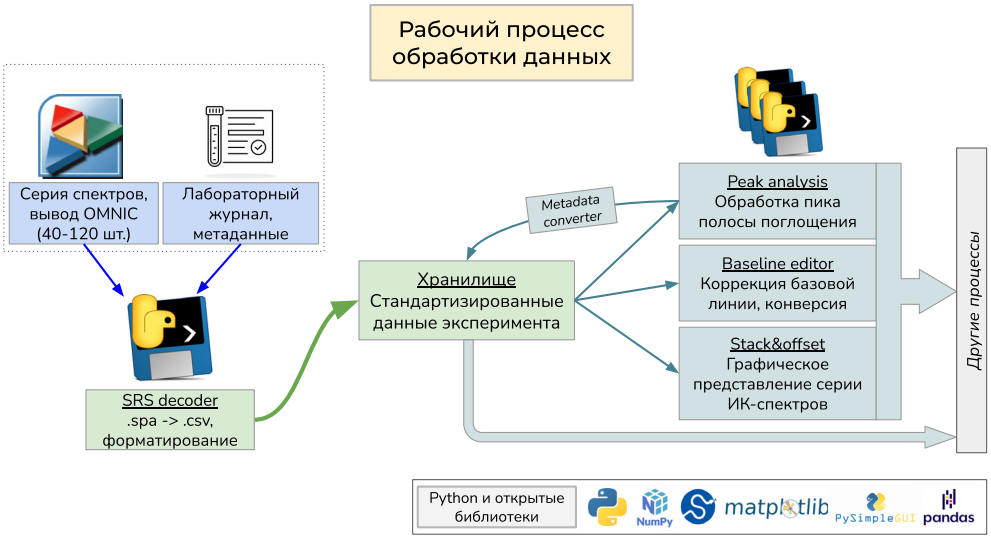
В ходе эксперимента на германиевое окно, закрепленное в главной камере, напыляется вещество, образуя лед. Контроль состава напыляемого вещества ведется при помощи масс-спектрометра. Образовавшийся лед сканируется ИК-спектрометром. Таким образом, источниками данных являются масс-спектрометр SRS RGA 200 и инфракрасный Фурье-спектрометр Thermo Scientific Nicolet iS50. Метаданные эксперимента, позволяющие, например, сопоставить время и температуру в камере при проведении термопрограммированной десорбции, хранятся в лабораторном журнале. Первичная обработка интерферограмм, снятых Фурье-спектрометром, осуществляется программой OMNIC.

Для изучения эволюции выращенного льда в течение продолжительного времени снимается последовательность ИК-спектров ($\sim$ 80 шт). Единичный спектр в формате *.spa* из серии из метаданных содержит только время с начала эксперимента, другие данные, такие как температура подложки или лучевая концентрация напыленного вещества, должны быть рассчитаны отдельно. В заводской программе не предусмотрена серийная обработка набора спектров, в связи с чем для типовых манипуляций над всей серией спектров необходима поштучная обработка. В ходе работы были автоматизированы задачи, связанные с поштучной обработкой, — насыщение спектров метаданными, исследование площади и формы пиков, оптимизировано время выполнения коррекции базовой линии, конвертации единиц измерения спектроскопических данных, построения вспомогательных графиков, на уровне программного обеспечения решена задача о каталогизации генерируемых данных.

`Python` был выбран в связи с его универсальностью, эффективностью, наличием большого количества открытых и активно поддерживаемых библиотек (`NumPy`, `SciPy`, `Matplotlib`, `Pandas`, `PySimpleGUI`) [4–6], а также обширного покрытия учебными материалами.

Представим список написанных приложений с перечислением основной функциональности. Диаграмма обновленного рабочего процесса представлена на рисунке.

— SRS decoder: серийная конвертация *.spa* $\rightarrow$ *.csv*, насыщение выходных файлов метаданными эксперимента (температура депозиции, время скана, скорость прогрева...) с минимальным вводом данных пользователем, каталогизация данных эксперимента.



Обновленный рабочий процесс обработки спектроскопических данных

— Peak analysis: исследование (по набору *.csv* файлов) зависимости площади, интенсивности и положения пика полосы поглощения от времени, графическое и табличное представление результатов для дальнейшего использования.

— Baseline editor: несколько режимов редактирования базовой линии для астрономических и лабораторных спектров, автоматический подбор и вычитание спектра водяного пара, конвертация спектров между различными единицами измерения.

— Stack&offset: графическое отображение произвольного числа ИК-спектров с регулировкой сдвига и числа отображаемых линий, быстрая подготовка черновых рисунков.

— Metadata converter: изменение метаданных серии спектров.

Разработанный пакет приложений позволил кратно сократить время базовой обработки экспериментальных данных.

Заключение

В ходе проведенной работы нам удалось идентифицировать такие участки процесса обработки данных, на которых необходимо повторение однотипных действий для большого числа файлов, а также задачи, требующие участия пользователя, для которых можно существенно сократить время исполнения, например, построение вспомогательных графиков для проверки линейности напыления вещества. С целью автоматизации процессов и экономии времени был написан набор приложений, которые исполняют рутинную работу быстрее человека, при этом обеспечивая стандартизацию получаемых результатов. Необходимыми требованиями к приложениям являлись скорость разработки, надежность, эффективность и удобство использования конечными пользователями — сотрудниками лаборатории. Подготовлены решения, обеспечивающие сбор, каталогизацию, анализ, конвертирование и представление данных.

Несмотря на наличие большого количества готовых решений для хранения, обработки и графического представления информации, выделение времени на написание более узконаправленных и эффективных приложений под наиболее частые задачи лаборатории является более эффективным решением в долгосрочной перспективе. Помимо прочего такой подход — построение пайплайна обработки данных — позволяет выработать и закрепить стандарты обработки и каталогизации, обеспечить единообразие оформления результатов: графиков и таблиц, получаемых различными исследователями, а также упростить доступ к данным, в том числе и для новых сотрудников, магистров и аспирантов.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации, тема FEUZ-2020-0038.

Библиографические ссылки

[1] *McGuire B. A.* 2018 Census of Interstellar, Circumstellar, Extragalactic, Protoplanetary Disk, and Exoplanetary Molecules // The Astrophysical Journal Supplement Series. — 2018. — Vol. 239, № 2. — P. 17.

[2] *Punanova A., Caselli P., Feng S. et al.* Seeds of Life in Space (SOLIS). III. Zooming Into the Methanol Peak of the Prestellar Core L1544 // Astrophys. J. — 2018. — Vol. 855, № 2. — P. 112. 1802.00859.

[3] *Vasyunin A. I., Caselli P., Dulieu F., Jiménez-Serra I.* Formation of Complex Molecules in Prestellar Cores: A Multilayer Approach // Astrophys. J. — 2017. — Vol. 842, № 1. — P. 33. 1705.04747.

[4] *Harris Ch. R., Millman K. J., van der Walt S. J. et al.* Array programming with NumPy // Nature. — 2020. — Vol. 585, № 7825. — P. 357—362.

[5] *Virtanen P., Gommers R., Oliphant T. E. et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python // Nature Methods. — 2020. — Vol. 17. — P. 261—272.

[6] *Hunter J. D.* Matplotlib: A 2D graphics environment // Computing in Science & Engineering. — 2007. — Vol. 9, № 3. — P. 90—95.