

GEO Uploader: Simplifying the data deposition in the GEO repository

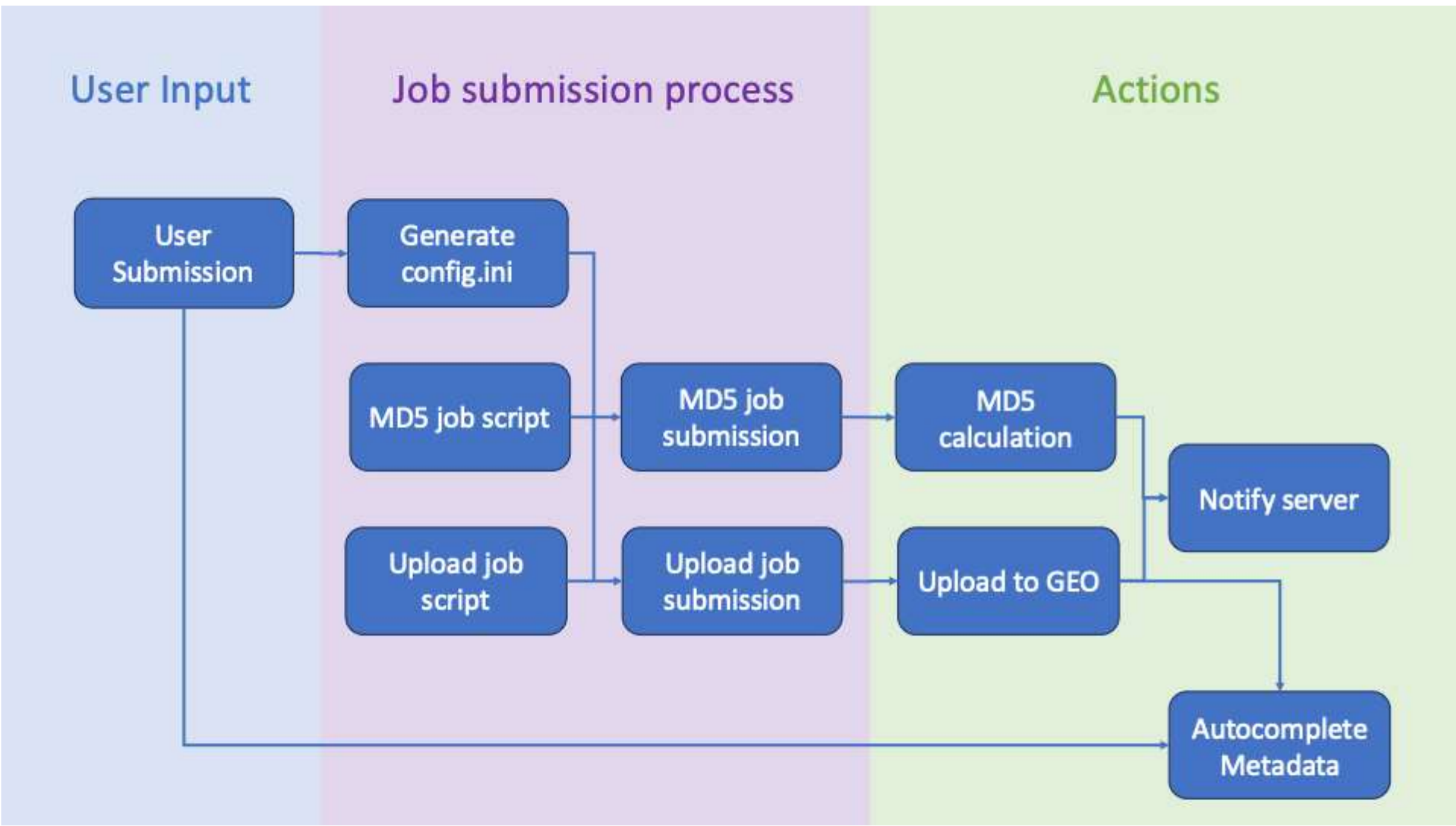
Ronald Domi^{1,2}, Falko Noé^{1,2}, Peter Leary^{1,2}, Hubert Rehrauer^{1,2}

1. Functional Genomics Center Zurich, ETH Zurich and University of Zurich. 2. Swiss Institute of Bioinformatics

Introduction

We present GEO Uploader, a web-based tool that automates the entire GEO submission workflow through an intuitive interface. The application reduces submission initiation time from 2-3 hours to under 20 seconds by automating file uploads, MD5 calculation, and metadata template population. Key features include parallel processing of uploads and checksum calculations, automated error prevention through template-based metadata completion, real-time progress tracking, and support for complex submission structures. Deployment across 30+ users with 50+ upload sessions, including datasets exceeding hundreds of gigabytes, demonstrates practical utility and reliability in research environments.

Steps of the Upload Workflow



Upon user submission through the web interface, the system performs three parallel operations: upload job creation, MD5 checksum job creation, and metadata template updating. Using the generated configuration file containing the list of files to upload and FTP credentials the tool submits the tasks to the job scheduler for background processing. Two parallel Python modules execute simultaneously, file upload to GEO servers and MD5 checksum generation with output to md5sheet.tsv. Upon completion of both processes, the server receives notification and marks the submission as complete, allowing users to download the populated metadata template for final GEO submission.

Implementation & Source Code

GEO Uploader is implemented as a web application using the Python Flask framework, offering compatibility with Mac, Linux, and Windows systems. The application uses a modular architecture, separating backend services into core and external components. Core services manage the main application logic, while external services handle tasks like job submission, user authentication, and directory management. This separation allows the tool to adapt to different deployment scenarios: External services can integrate with existing facility infrastructure (such as institutional job schedulers and authentication systems) for centralized deployments or use default implementations for standalone installations on a client computer.

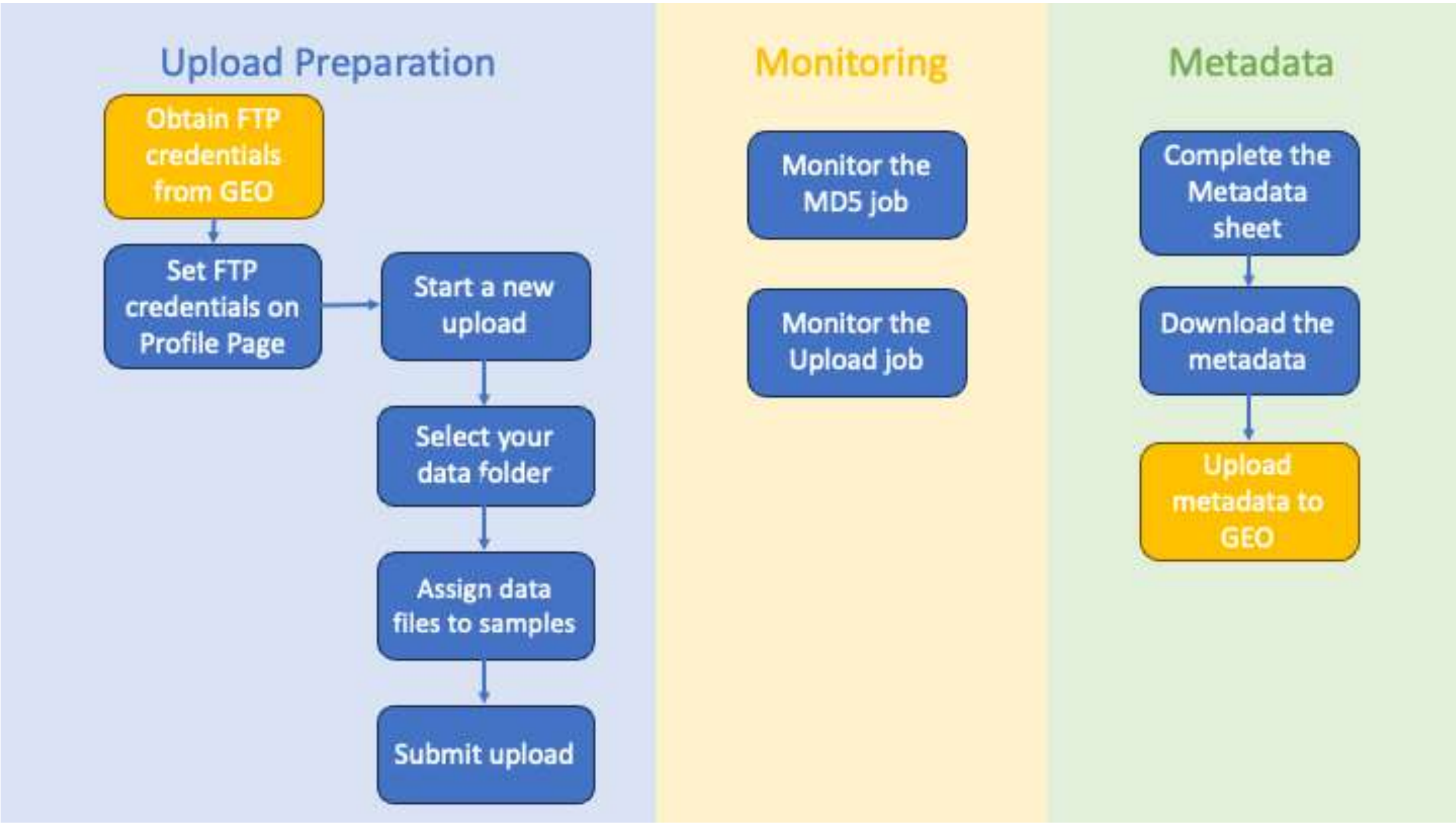
The frontend utilizes HTML with Jinja2 templating, Bootstrap for responsive design, and JavaScript for interactive elements. Installation and configuration details are maintained in the project's GitHub repository, with Conda or Mamba package managers as the sole requirement for setup.

The GEO Uploader is available at Github:

<https://github.com/fgcz/geo-uploader>



User Workflow



Complete user workflow for GEO submission. Users must first obtain their personal GEO FTP credentials by following the GEO Uploader instructions and saving them to their Profile Page. Once credentials are saved, users can initiate a new submission by selecting files for upload and submitting the form, which then triggers background processing jobs. All job progress can be tracked through individual dashboard pages. Users can begin preparing their metadata sheet immediately, then download and submit the completed metadata to GEO once the MD5 verification job has finished.

Core Facility Integration

GEO Uploader is designed for end-users as well as for deployment in core facilities. End users run the uploader as a localhost web application that supports uploading of local files. When deployed as a centralized service for institutional access, the uploader can be adapted and integrated with user authentication systems and file management systems. The authentication system ensures complete session isolation, preventing cross-user data access while maintaining the collaborative benefits of shared infrastructure. The GEO Uploader is in operational use at the Functional Genomics Center Zurich and integrates with:

- Operational GEO Uploader instance:
<https://geo-uploader.fgcz.uzh.ch>
- SUSHI data analysis interface:
<https://fgcz-sushi.uzh.ch/>
- B-Fabric project and data management framework:
<https://fgcz-bfabric.uzh.ch>

Since deployment, the tool has processed over 50 upload sessions across more than 30 users including non-bioinformaticians, handling datasets ranging from standard RNA-seq studies to large-scale submissions exceeding 500GB. The multi-user deployment with session isolation has proven effective for institutional use, allowing multiple researchers to work simultaneously without interference.

Discussion

GEO Uploader automates the complex and error-prone process of data submission to the Gene Expression Omnibus repository. By reducing submission preparation time significantly and eliminating common data entry errors through automated file handling and metadata population, the tool significantly lowers the barrier to data sharing compliance required by major scientific journals.

The web-based interface successfully supports GEO submissions, enabling researchers without specialized bioinformatics expertise to complete uploads independently. The automation of technical components, including MD5 checksum calculation, FTP upload management, and metadata template completion provides an almost end to end support on this upload. This capability is particularly valuable for the "long tail" of submissions involving heterogeneous file structures that were previously impractical to manage manually.

GEO Uploader reduces technical barriers for uploading functional genomics data to GEO, and therefore supports the broader goals of reproducible research and data reuse in the genomics community.