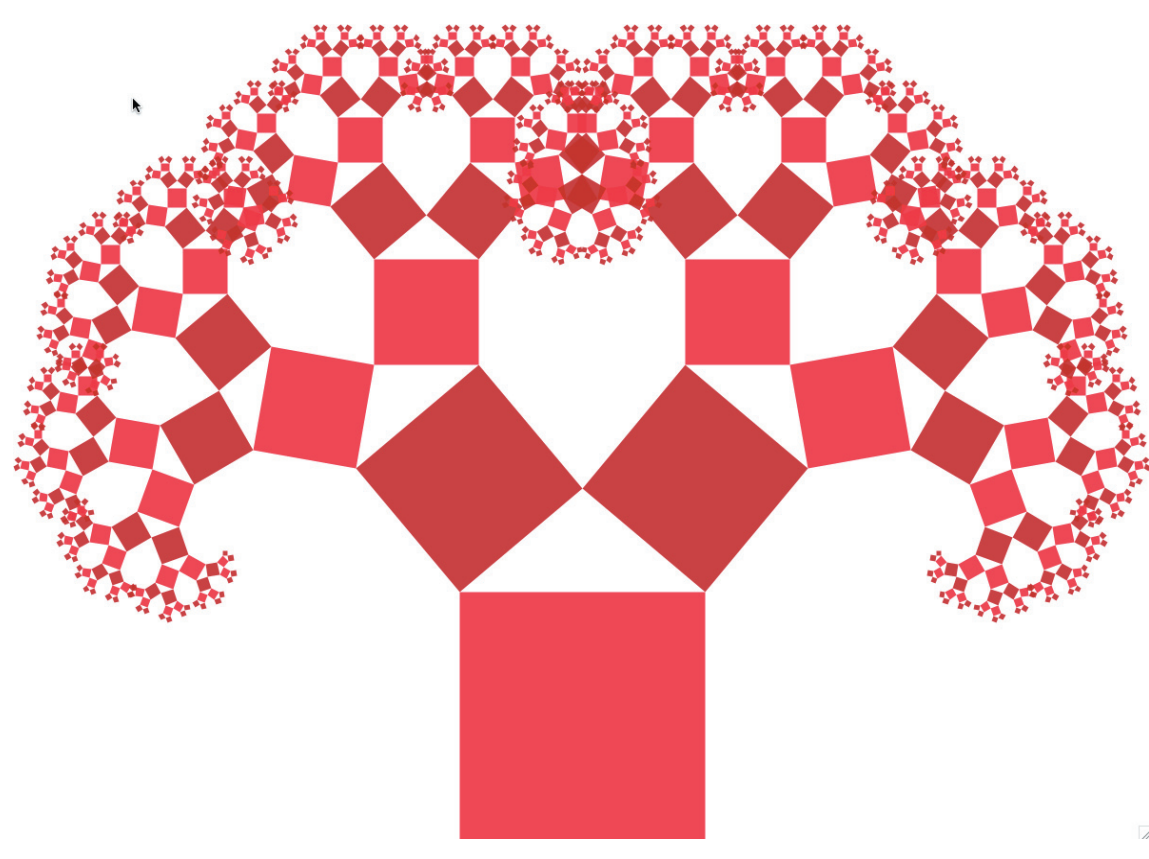


Decentralized Data Sharing for the Proteomics Community

Philip C. Andrews and Jayson A. Falkner
University of Michigan, Ann Arbor, MI, United States

Poster PDF and Additional Information:
<http://tranche.proteomecommons.org>



Tranche Project

Secure P2P for the Scientific Community

Introduction

The ProteomeCommons.org Tranche Project is designed to allow publication of all raw data relating to proteomics experiments. The system allows investigators to securely share files with collaborators, reviewers of manuscripts, and to publicly disseminate their data. Encryption and digital signatures allow authentication of data files, so provenance and integrity are unambiguous. The system provides for distributed downloads, background file compression, and allows file replication across the network to minimize the chance of file loss. While the system is intended to solve the raw data distribution problem, it is independent of file type so it can be used for peak lists, search results, images, code, and any other form of digital information.

Key Points

- Store all data associated with a manuscript, or study, on-line
- Download data to reproduce a study, find new results, test new tools, etc.
- Verifies that data hasn't changed since publication
- Data is replicated to prohibit loss of files
- No centralized repository is required
- Open-source and free to use

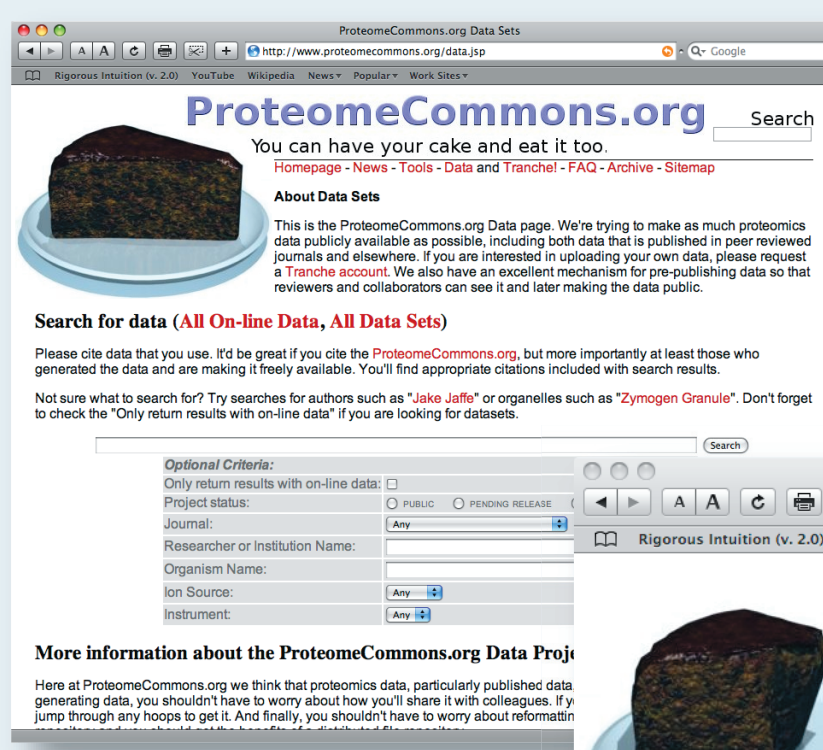
Availability

This project is on-line now and requires a web browser to use. All code is open-source and free to use both commercially and non-commercially. Documentation and code releases may be downloaded from the project homepage on ProteomeCommons.org. The project is coded in the Java programming language and it runs on Linux, OS X, Solaris, BSD, and Windows.

Link: <http://tranche.proteomecommons.org>

Methods

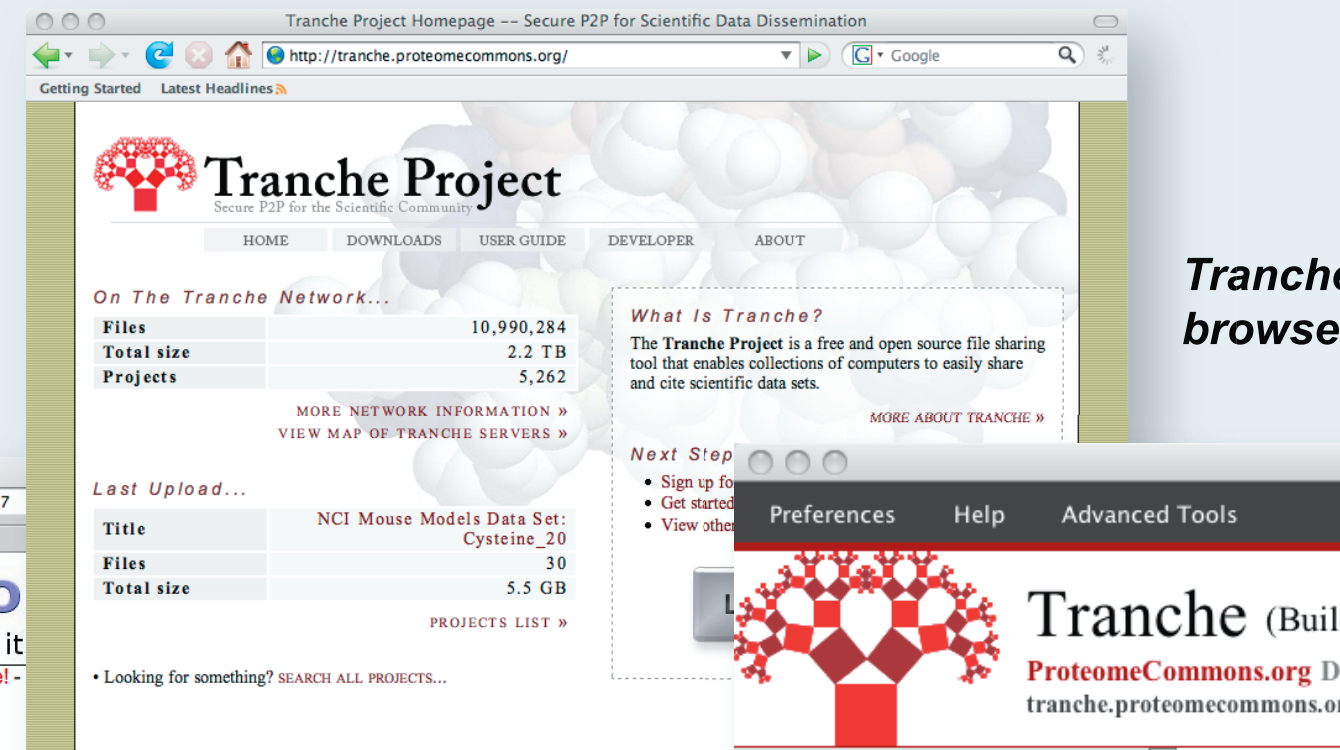
Data Search Page



Individual Data Information Pages

ProteomeCommons.org tracks information pages for individual data sets and provides a way to search for specific data sets.

Tranche Tool (Java WebStart)

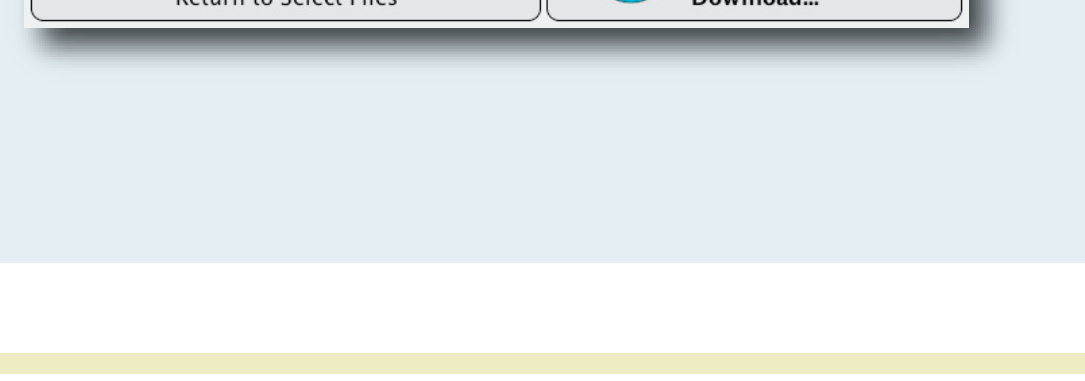
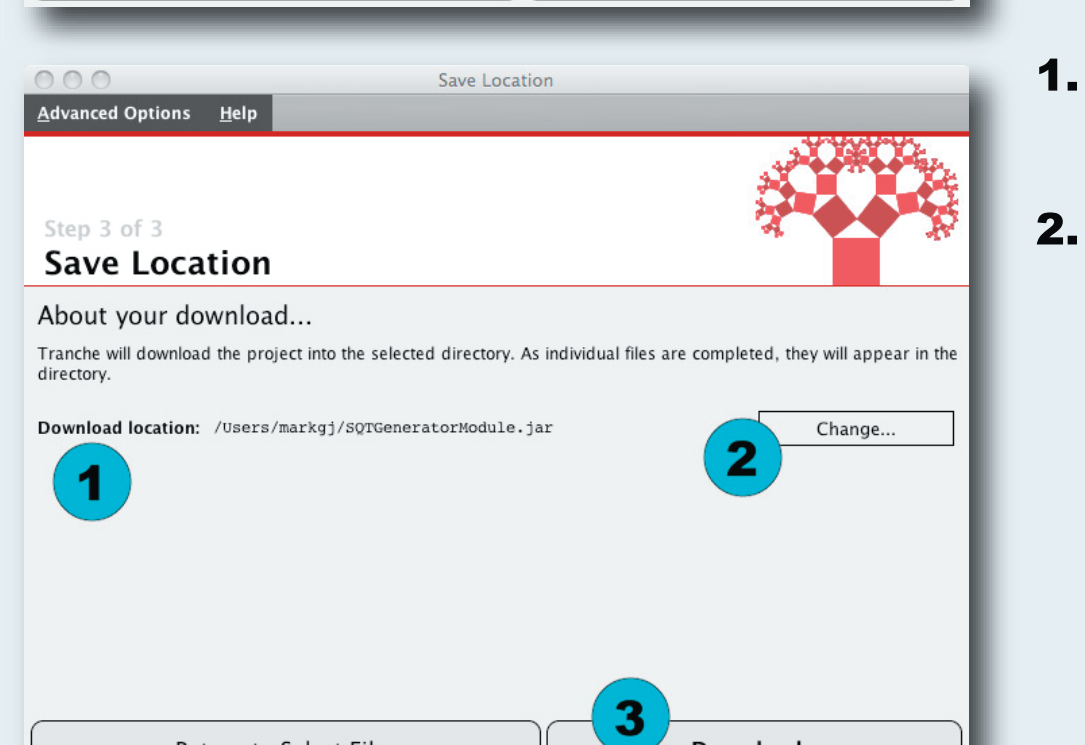
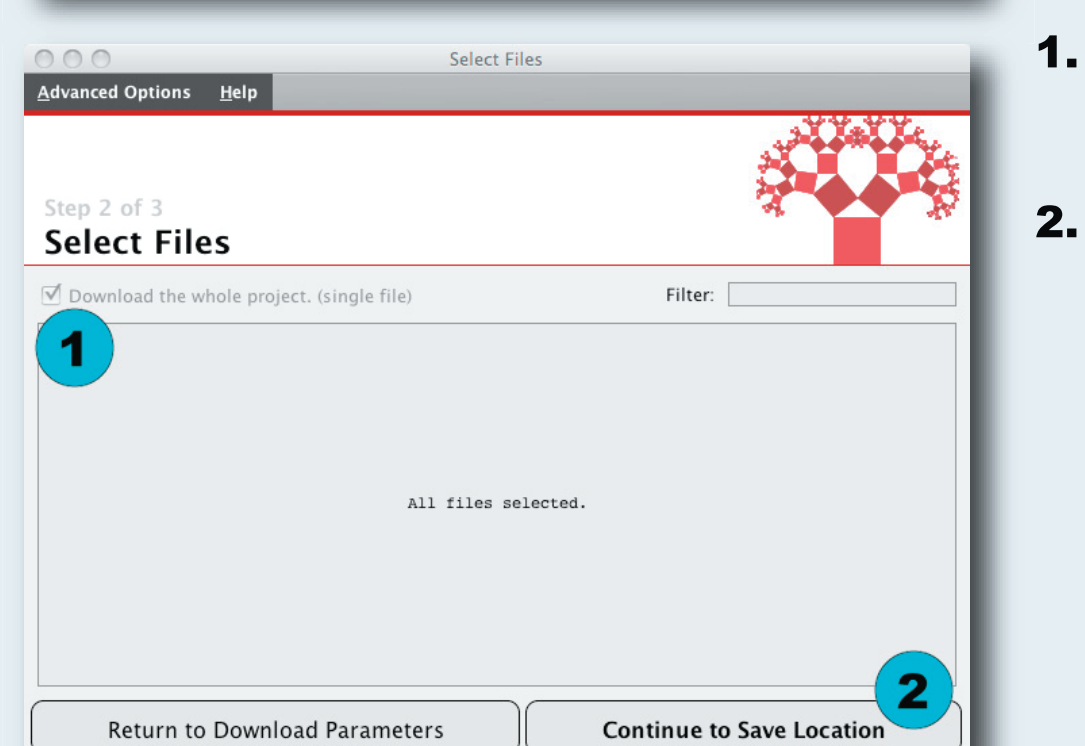
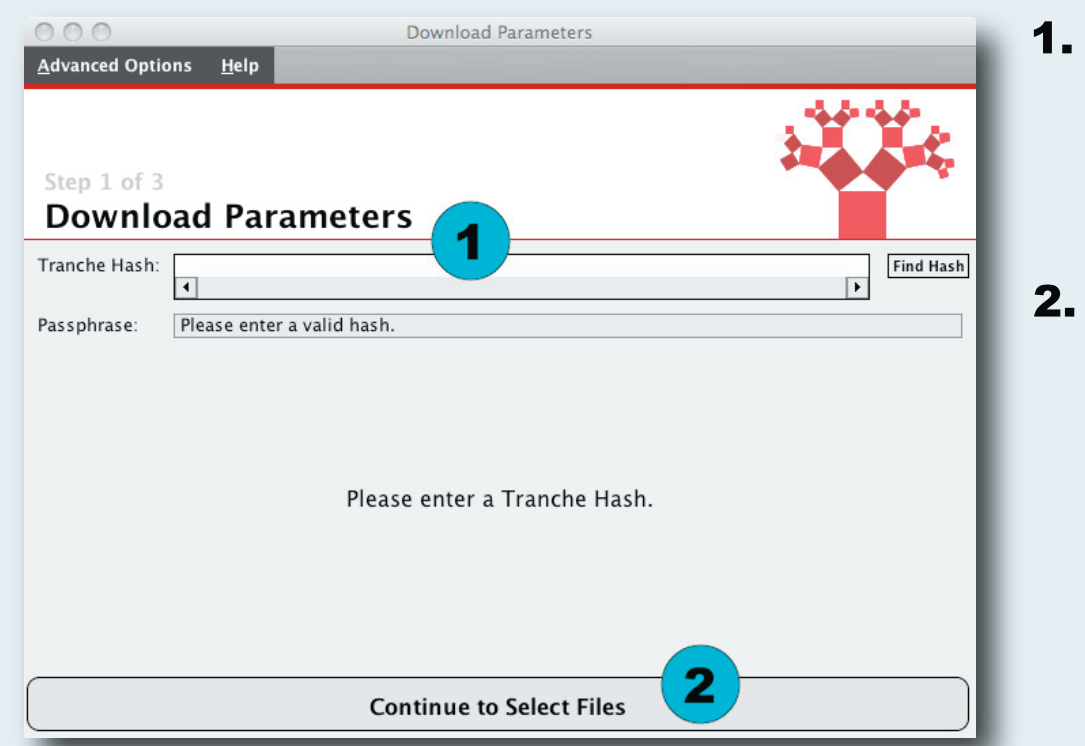


Tranche Tool Launches from the browser page and is ready to use

Tranche (Build: 2173) ProteomeCommons.org Distributed File System

Intuitive Java application with full cross-platform compatibility

Download Data

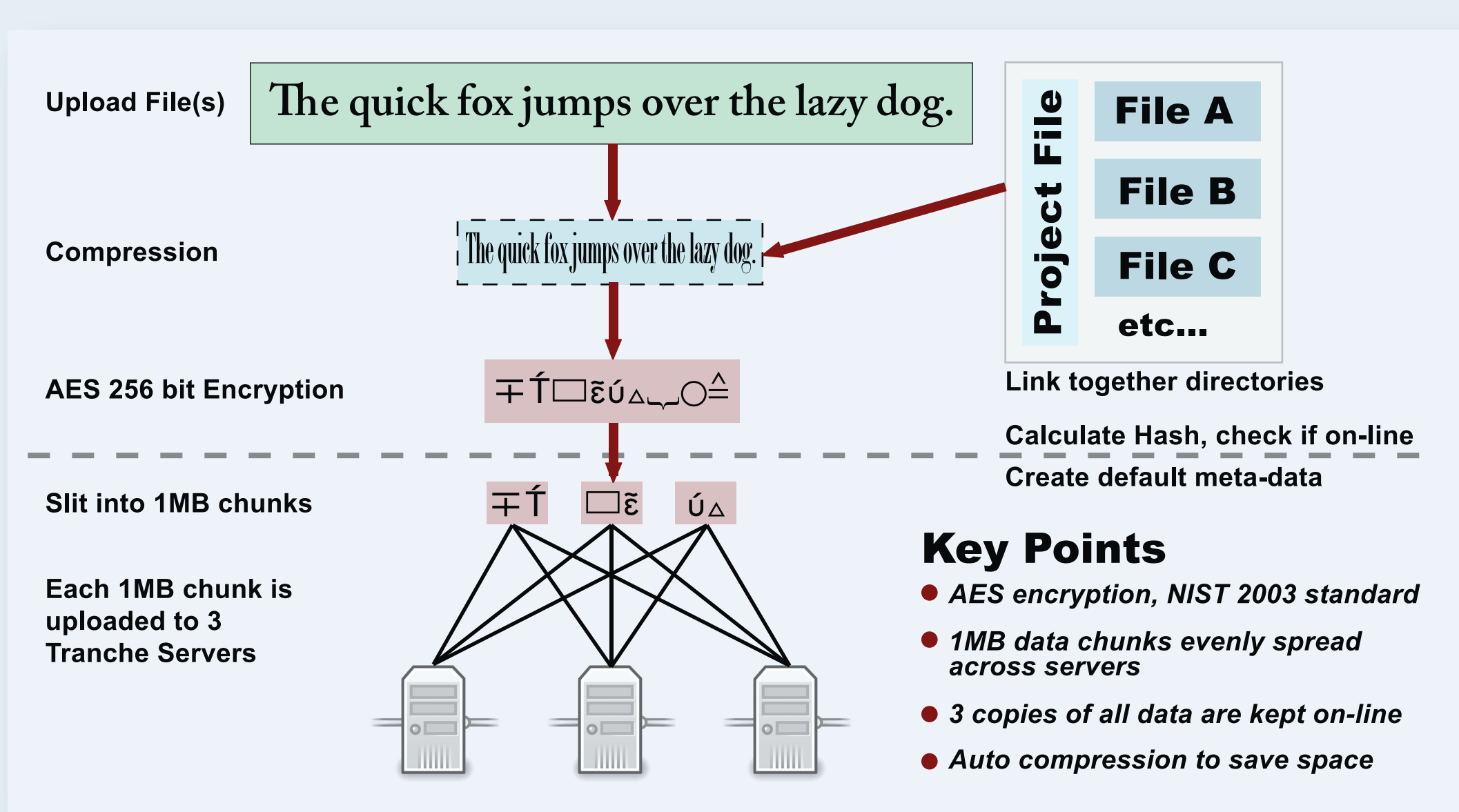


1. Input the Hash in the Download by Hash String dialog box. If the project is passphrase protected, you will need to enter the passphrase as well.
2. Continue to Selected Files when you have entered the required information.

1. Choose to download a complete project, or selected files from within a project.
2. Continue to Save Location when you have made your choices.

1. Choose the location path where you would like to save the files to.
2. When you have made your download location selection, click the Download... button to download the files to that location.

Tranche Upload Process



Encryption and Unique Identifiers for files

Individual Hashes (base 64 encoded for "Some data...")

MD5: KmwU1fnU9nk/Bt8Z+MA5gbdhXgSn1qVqx8VLgexZseuzXg=

SHA-1: 5gbdhXgSn1qVqx8VLgexZseuzXg=

SHA-256: iXxEIMiQ7KpNswWWSaItWhxdWFRZCA1sReTHKLNsW=

Length: AAAAAAAAAAw=

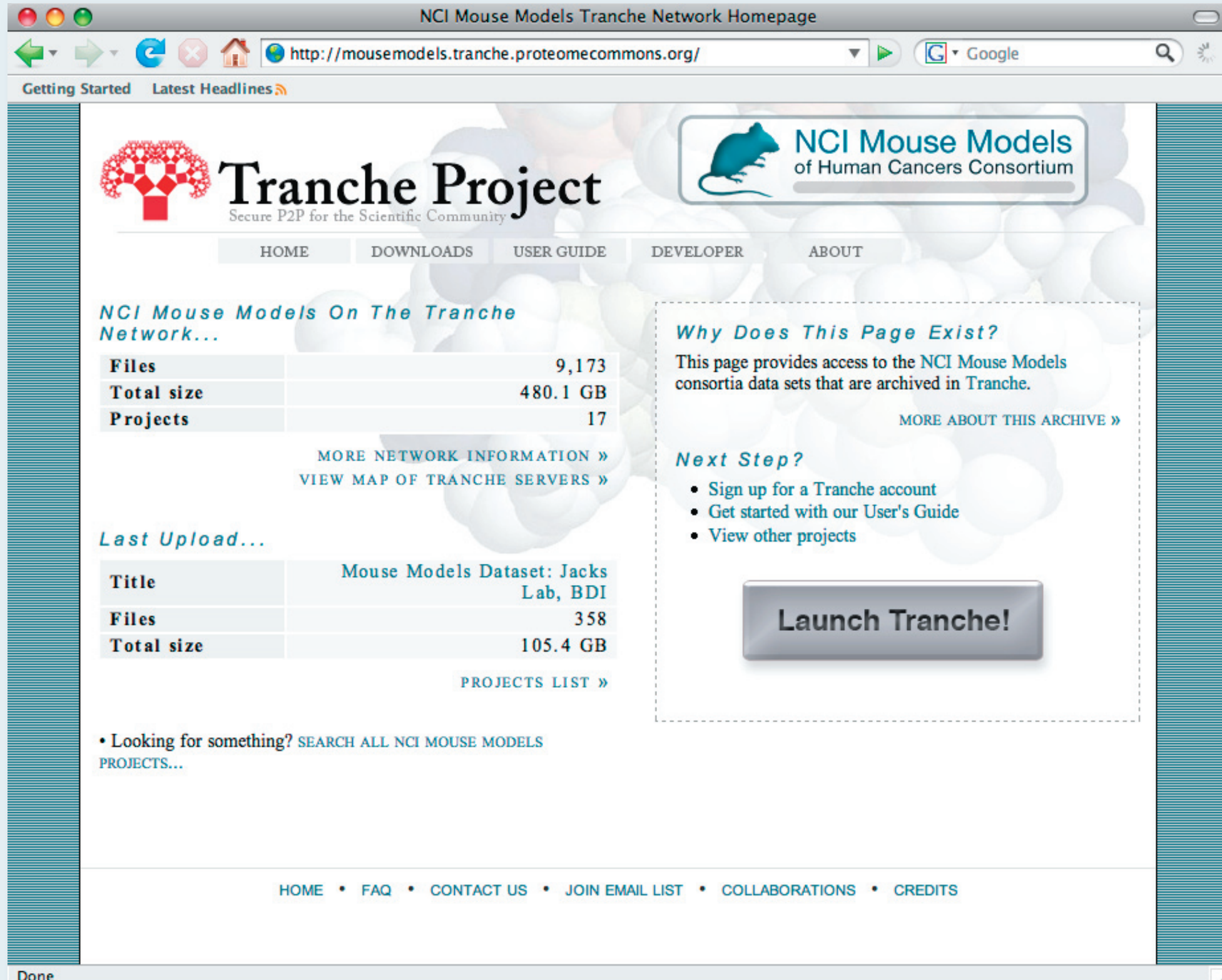
Aggregate Hash Identifier: KmwU1fnU9nk/Bt8Z+MA5gbdhXgSn1qVqx8VLgexZseuzXgiXxEIMiQ7KpNswWWSaItWhxdWFRZCA1sReTHKLNsWAAAAAAAAAAW=

Tranche uses 3 standards to create a digital hash^{2,3}, the combination results in a "Big Hash." These hashes can be used to verify data has not changed since upload or publication. Unlike URL's, hashes will not suffer from "link rot."

Example Use Cases

Below are two examples of Tranche use. First, the National Cancer Institute (NCI) Mouse Proteomics Technologies Initiative (MPTI) data sets represents Tranche's use to archive and disseminate a large, multi-laboratory set of proteomics data. Second, the FASTA archive demonstrates Tranche's use as part of an ongoing archive of protein databases in FASTA format, including support for the PSI standardized format.

NCI Mouse Models (MPTI) Long Term Data Archive



- A large-scale NCI funded project to create and collect data from mouse models for the human cancers
- Tranche provides convenient access to all the collected data sets
- A collaboration page is also on-line to summarize key statistics and citations

FASTA Database Archive on ProteomeCommons.org



- A long-term archive of helpful protein databases in FASTA format
- Provides a uniform citation mechanism for specific versions of files
- Monthly updates are posted for each database

References:

1. JA Falkner, PJ Ulintz, PC Andrews "A Code and Data Archival and Dissemination Tool for the Proteomics Community" American Biotechnology Laboratory, 2006.
2. R. Rivest, "The MD5 MessageDigest Algorithm", Network Working Group, RFC 1321, <http://www.ietf.org/rfc/rfc1321.txt>
3. The NIST Secure Hash Standard (including SHA1, SHA256): <http://csrc.nist.gov/publications/fips/fips1802/fips1802withchangenotice.pdf>
4. Falkner JA, Hill JA, Andrews PC "Proteomics FASTA archive and reference resource." Proteomics. 2008.

Statistics and Usage

The Tranche project was started in 2006 and has grown to include millions of files and participating groups around the world. A complete list of collaborations that have been done with Tranche/ProteomeCommons.org can be found at:

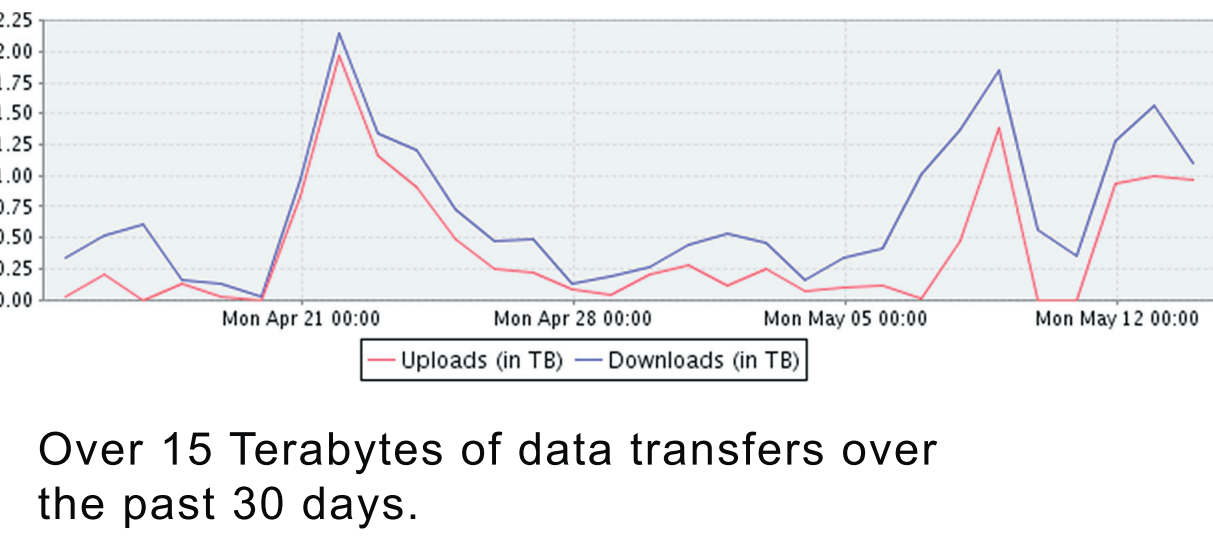
<http://tranche.proteomecommons.org/examples/>

Map of the Core Tranche Servers

Data On-line (April 2008)



Usage



Growth by year

2007
Data Sets On-line: 3,496
Total Size: 890 GB
Files Online: ~9 Million

2006
Data Sets On-line: 220
Total Size: 250 GB
Files Online: 220,000

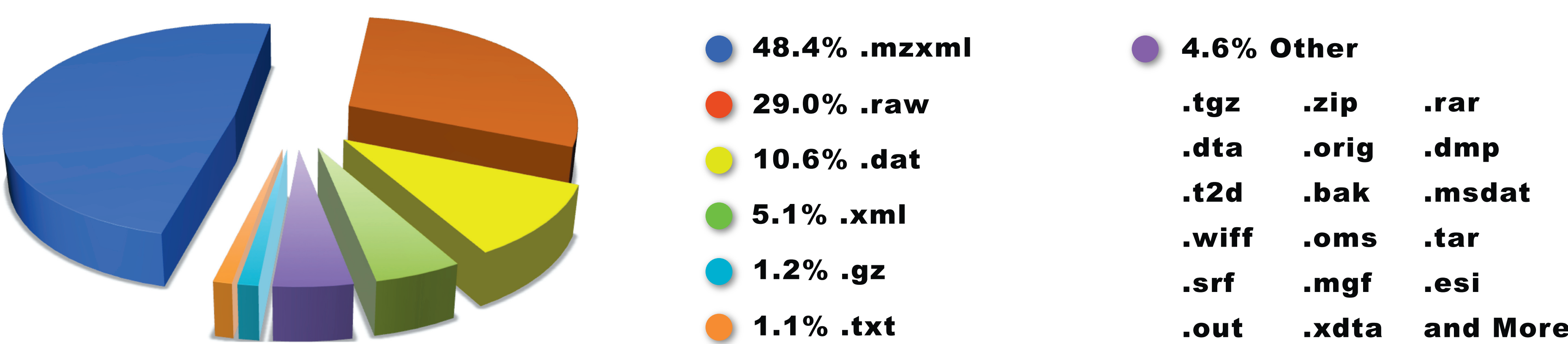
Files On-line.....11.2 Million

Data Sets.....5,435

Total Space Used.....3 TB

Available Space.....100 TB

Available File Types



Conclusion

It is possible to host the large amounts of raw data that are commonly associated with proteomics experiments, and it is possible to do it now. The ProteomeCommons.org Tranche Project provides an open-source, free-to-use tool that can readily handle large amounts of data and large numbers of users.

Along with being able to host all the raw data associated with an experiment is the need to annotate these data with information about the experiment. Tranche provides a mechanism for letting annotations be linked to raw data, which is intended for future use with upcoming proteomics annotation standards. This provides a practical option for researchers to meet recently proposed requirements for publishing data.