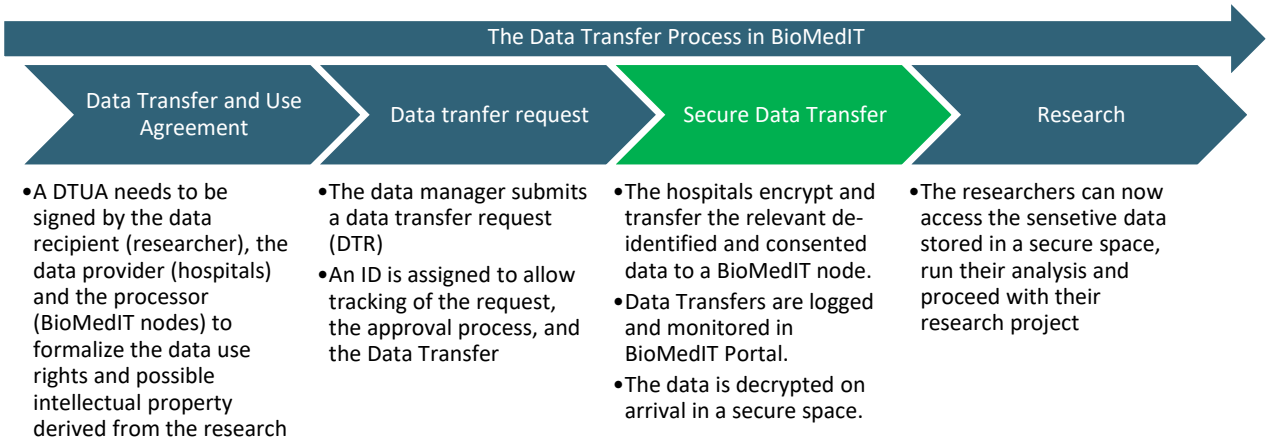
Fact-sheet 2021

BioMedIT's Secure Data Transfer

Given the sensitive nature of health-related information, research using patient data calls for high levels of security and data protection in ICT infrastructure and processes, and requires a corresponding level of expertise in handling sensitive data to fulfil all the stringent legal, regulatory and ethical requirements.

The Personalized Health Informatics (PHI) Group of SIB Swiss Institute of Bioinformatics together with the collaborating partners have setup the BioMedIT Network as an integral part of SPHN to provide all authorized researchers in Switzerland with easy access to collaborative analysis of confidential data without compromising data privacy.

To bring the data to BioMedIT in a secure way, BioMedIT set up the following process:

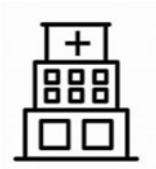


To support the full process of secure data transfer, the BioMedIT Interoperability Working Group (BIWG) developed and maintains **sett**, **S**ecure **E**ncryption and **T**ransfer **T**ool.

With both a graphical user interface (GUI) and a command line interface (CLI), sett has four modules: **PGP key Management**, **Packaging and Encryption**, **Transfer, and Unpacking and Decryption**.

1 Who can benefit from sett

- **Data providing institutions** can securely connect to the network to enable secured sharing of sensitive research data over the BioMedIT infrastructure.



- ✓ PGP key management
- ✓ Data packaging
- ✓ Data encryption
- ✓ Data transfer

- On the other end, **Data Managers of a research project** receive, decrypt, unpack, and process the data for research.

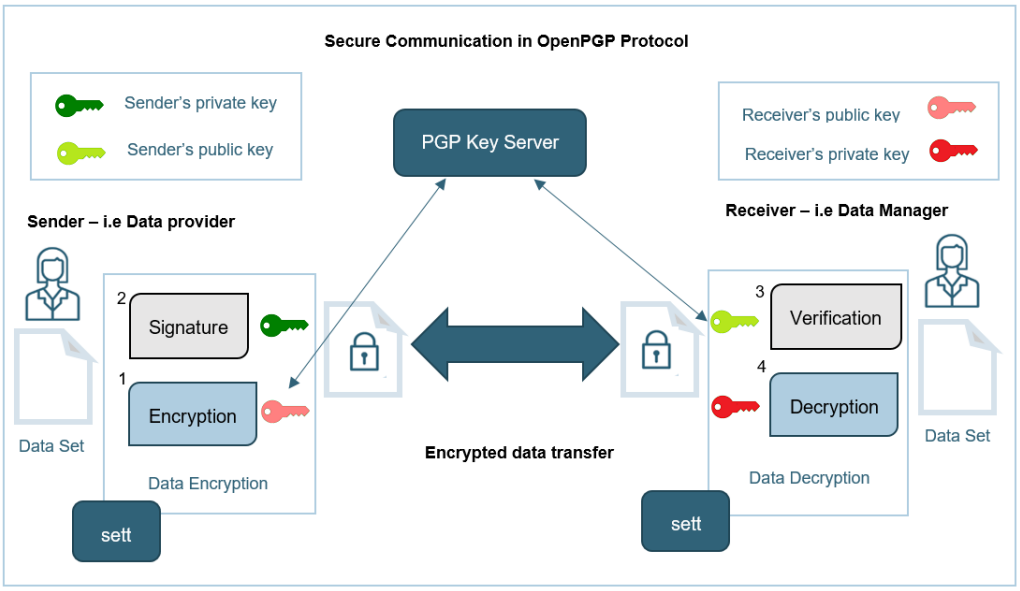
- ✓ PGP key management,
- ✓ Data decryption
- ✓ Data unpacking

SIB Swiss Institute of Bioinformatics
Personalized Health Informatics Group (PHI)
Elisabethenstrasse 43
CH-4051 Basel

2 How secure communication is implemented

sett uses OpenPGP[1], a method for secure communication between two or more users.
In this system, each user has a pair of unique keys consisting of a private key and a public key linked in a way that:

- data encrypted with a given public key can only be decrypted with the matching private key.
- data signed with a given private key will only be recognized by the matching public key.



1. The sender **encrypts** the data set using the recipient's public key
2. The sender **signs** the data set with his own private key
3. The recipient **verifies** the authenticity of the signature using the public key of the sender
4. The recipient **decrypts** the data set using his own private key

3 sett's modules - Covering the secure communication process end to end

Through an intuitive interface or command line, sett streamlines the process of data packaging, encryption, signing and data transfer. sett allows a number of options to be customized and pre-defined. For instance, you may change the default compression level, or define a default data sender.

PGP Key Management

- sett supports the generation of new PGP public and private keys.
- Includes options to download and upload keys to/from a pre-defined key server.
- Generates PGP key revocation certificates, required in case the user forgets or have his private key and password stolen/compromised.

Data Packaging and Encryption

- sett allows the encryption of any combination of individual files and directories.
- Supports multi-recipient data encryption, allowing the encrypted file to be decrypted by multiple recipients.
- The encrypted data set is bundled with a metadata file, containing information about who is sending the file and to whom it should be delivered.

Data Transfer

- sett can transfer data using *sftp*[2] or *liquid_files*[3] protocols, depending on the server to which the data should be sent.
- To avoid retyping connection settings for every transfer, it is possible to store predefined connection profiles.

Data Decryption and Unpacking

- sett GUI allows the decryption and decompression of files in a single step.

1 OpenPGP: email encryption standard

2 sftp: network protocol that provides file access, file transfer, and file management over any reliable data stream

3 liquid_files: virtual appliance that helps companies and organizations send, receive & share large files, fast & securely

BioMedIT key server

To provide a central server where public PGP keys for the BioMedIT data transfers can be stored, BioMedIT provides an key server based on SKS-Keyserver technology[4] hosted at the SIB Swiss Institute of Bioinformatics in Lausanne. To increase the level of trust, public keys are verified and approved ("signed") by the Data Coordination Center.

4 Integrated control-checks with the BioMedIT network

In addition to data integrity checks, when sett is used in the BioMedIT network, additional constrains apply to facilitate and ensure SPHN process requirements are met.

Certification status

- ✓ sett implements a security feature where all public keys used within sett must be signed/certified by a "central authority" key i.e. Data Coordination Center in the BioMedIT network.

Authorization status

- ✓ if sett is used in the BioMedIT network, it can verify if a Data Transfer request is in place, valid and authorized by the Data Coordination Center.
- ✓ Recipients must be officially approved Data Managers of a project in BioMedIT.

Data integrity

- ✓ sett ensures the integrity of the transferred files by computing checksums on each file that is packaged, and adding this information to the encrypted data.
- ✓ The integrity of each file is verified automatically upon decryption of the file by sett, providing the guarantee that all files were transferred flawlessly.

Logging of Data transfers

- ✓ All data transfers are logged, the logs can be viewed on the BioMedIT portal

5 sett benchmarks

A customizable script is available to run sett benchmarks with a number of different factor levels, like data size, data type, sett commands to run, and more. A complete guide to setup and run benchmarks can be found here: sett benchmarks.

Note: The results presented below were obtained by using input data in the form of text files containing **simulated genomic data in FASTA[5] format**. Data compression efficiency, both in the amount of compression and the time needed for compression, can significantly vary depending on the type of input data, and therefore the sett performance values illustrated in this section are not representative of all data types.

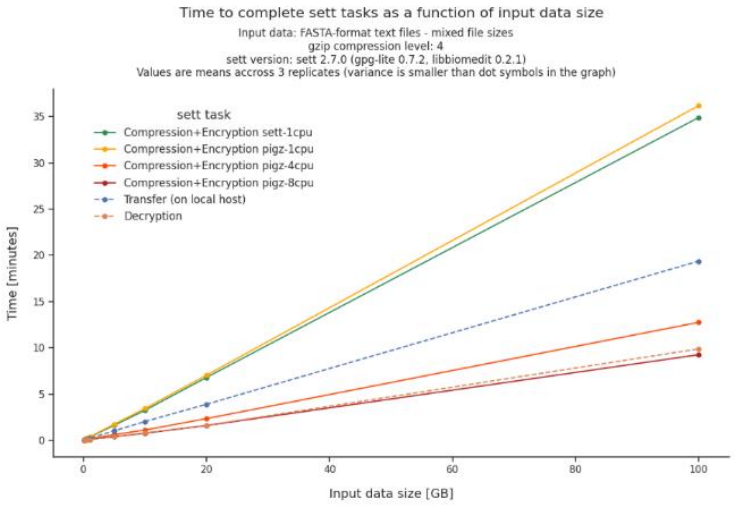
4 SKS-Keyserver technology, provides cryptographic privacy and authentication for data communication.

5 FASTA: text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes.

5.1 Overall sett performance

For data packaging (compression + encryption), **sett is able to process data at a speed of about 50 MB/second** (3 GB/minute).

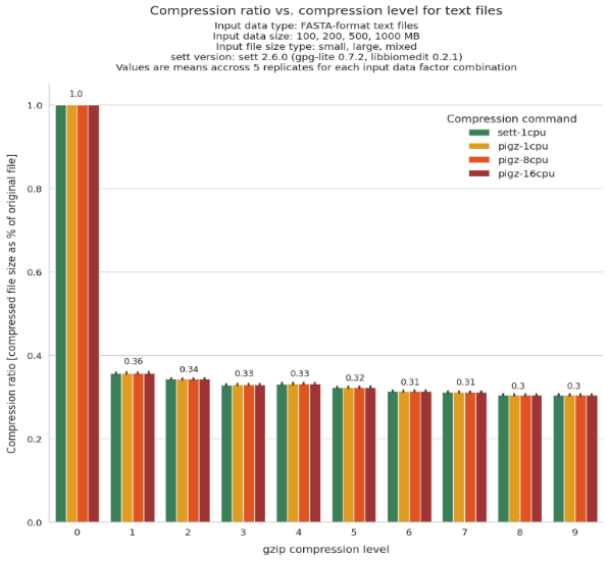
- ✓ Input data sizes up to 100 GB were tested, and the time to complete the different tasks scales linearly with input data size.
- ✓ The benchmark was run on a machine with an AMD Ryzen 7 3700X processor.



5.2 Compression level comparison

This graph shows **how much data compression can be expected from each compression level when packaging text files with genomic data.**

- ✓ All shown values are averages obtained across different input data sizes (ranging from 100 MB to 1 GB) and file size compositions (i.e. lots of small files vs a few large files).
- ✓ Values give the size of the *sett* output file as a fraction of the input data size.



6 Where can I find more information about sett?

Contact BioMedIT at dcc@sib.swiss, a member of the BioMedIT Interoperability Working Group, or go to https://sett.readthedocs.io, here user documentation can be accessed.

The codebase and data packaging specification can be accessed under https://gitlab.com/biomedit/sett.

7 Problems with sett?

Send an email to biomedit@sib.swiss to open a ticket with your problem.

For information about the BioMedIT Project, go to BioMedIT Project - SPHN.