REGENSTRIEF X DATAVANT

N3C Linkage Honest Broker – Site Technical Engagement Packet

Introduction

Thank you for joining the N3C hashing community for COVID data linkage. N3C has established a Linkage Honest Broker (LHB) service which will enable privacy preserving record linkage across COVID-19 datasets developed in coordination with the N3C Phenotype and Data Acquisition Workstream.

Regenstrief Institute is the partnered Linkage Honest Broker. Regenstrief Institute is a dynamic, people-centered research organization driven by a mission to connect and innovate for better health. All people deserve the best quality care. That is why Regenstrief Institute conducts research and development at the intersection of clinical medicine, technology, academia, and industry.

Datavant's mission is to connect the world's health data to improve patient outcomes. Datavant believes in connecting healthcare data to eliminate the silos of healthcare information that hold back innovative medical research and improved patient care. Datavant helps data owners manage the privacy, security, compliance, and trust required to enable safe data sharing.

The N3C Data Enclave is a secure platform through which the harmonized clinical data provided by our contributing members is stored. The data itself can only be accessed through a secure cloud portal hosted by NCATS and cannot be downloaded or removed.

In addition to sending data to the N3C Data Enclave, sites participating in the hashing community will prepare an additional set of files that will be submitted directly to the LHB service at Regenstrief Institute. These additional files include hashed identifiers (referred to throughout this packet as ***tokens***), which correspond to a unique patient ID, as well as a Manifest file that includes metadata describing site-specific information.

The LHB service will enable privacy preserving record linkage (PPRL) across COVID-19 datasets developed in coordination with the N3C Phenotype and Data Acquisition Workstream. There are three main reasons why privacy preserving record linkage is key to this effort:

1. PPRL enables de-identified deduplication of patients across institutions to account for care fragmentation.
2. PPRL enables de-identified linking to multi-modal data, such as image data from various health system PACS systems.
3. PPRL enables de-identified cohort overlap discovery from other research studies. For example, we can understand the extent of overlap between the NIH All of Us cohort and the N3C cohort.

This packet is intended to offer ***technical documentation*** on onboarding with the LHB and submitting tokenized data to the LHB. It is divided into two main sections:

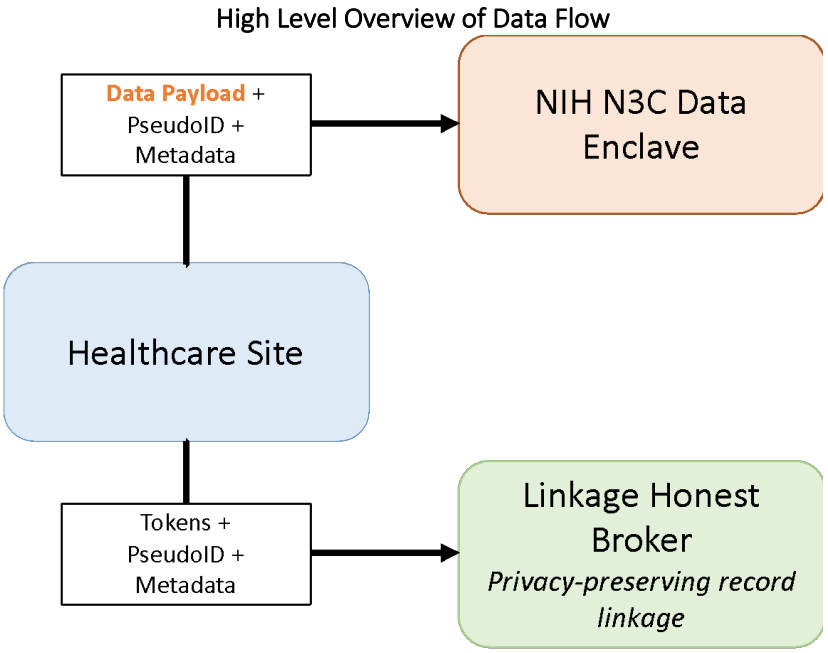
1. **Prepare and tokenize data** for your N3C phenotype cohort using Datavant. At the end of this section, sites should be able to understand and implement the following processes:

   a. implement Datavant's de-identification and tokenization software
   b. run Datavant software to generate tokens
   c. package the tokens into the specified file formats in preparation for submission to the LHB
2. **Onboard with the LHB and submit token files** prepared from the first section. Prior to onboarding with the LHB, participating sites must execute the Linkage Honest Broker Agreement. At the end of this section, sites should be able to understand and implement the following processes:
   a. register team members for SFTP credentials
   b. submit the token package to the LHB Secure Landing Zone

Data Flow

There are three main parties involved in the data flow:

1. N3C Data Enclave: the Data Enclave is a secure platform where clinical data for participating sites is stored. The Data Enclave's technology partner is Palantir.
2. Linkage Honest Broker (LHB) at Regenstrief Institute: the LHB service performs privacy performing record linkage on de-identified data using Datavant software tools. Any files sent to the LHB are *not* intended to be shared with the Data Enclave.
3. Healthcare Site: healthcare sites perform de-identification on their data using Datavant software tools. De-identified data is sent via SFTP to the LHB. Healthcare sites also send data to the Data Enclave. For information on preparing your phenotype for the Data Enclave, please visit this link.



Data Governance Resources

As stated above, this packet is a resource of technical documentation for onboarding with the LHB and submitting tokenized data to the LHB.

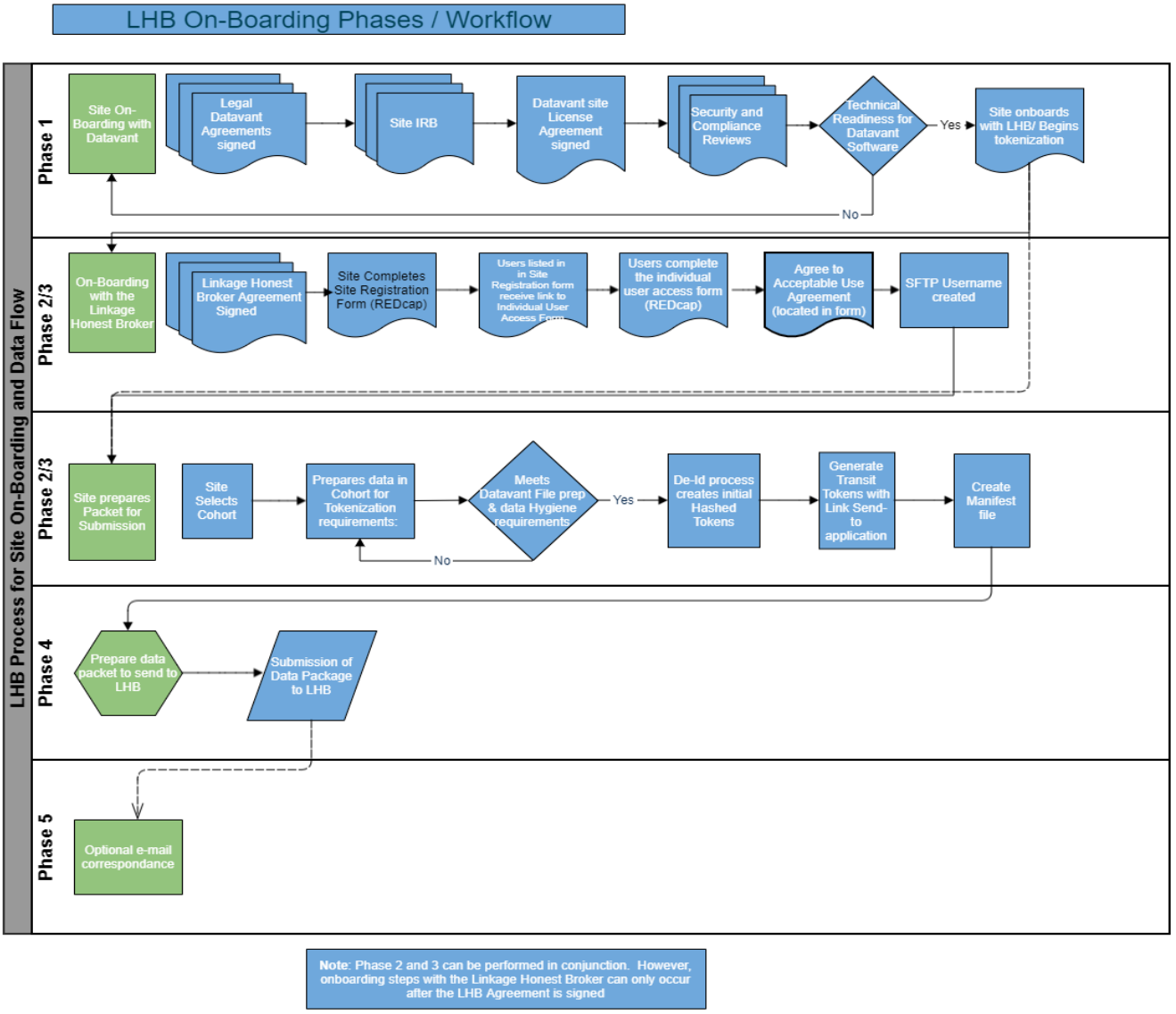
We recognize that many sites will have questions about data governance for tokenized data, such as:

- planned and potential use cases
- what use cases sites can opt into or out of

- operational firewalls between entities (e.g., where tokens will and will not be stored)

For more information on data governance, please refer to PPRL data governance information here.

LHB Onboarding Process



Version History

Version Number	Date	Summary of Changes
1	14 May 2021	Initial publication
2	27May2021	Updated Transit Token, Manifest file format to remove time and include date format. Updated .ZIP file naming convention to remove time and include date format Updated Upload to SFTP instructions to include additional details
3	10Aug2021	Updated Honest Data Broker (HDB) to Linkage Honest Broker (LHB) throughout document Updated diagrams/pictures to Linkage Honest Broker/LHB Update token and manifest file naming convention to add description. Update manifest file extension from. hdr to .csv Updated governance information link Updated dates on Datavant certifications
4	28Sep2021	Updated Datavant information to reflect v4 of software, released on August 25, 2021 Updated to remove reference to Service Desk Account Updated LHB Onboarding diagram to remove IU ID Added LHB Individual User Access form completed on to readiness checklist Updated LHB Registration Instructions Updated FAQ – added submission frequency question. Updated instructions for service desk tickets
5	27Oct2021	Added details on using v3 of Datavant software Updated LHB Onboarding Process flow diagram Added SSH key instructions document
6	12Nov2021	Added viral variant steps in Appendix L and to Site Readiness Checklist Updated column order for input files

Site Readiness Checklist

Please ensure you have completed the following activities in preparation for sending data to the LHB.

Legal, Compliance, and Governance:

- ❑ Linkage Honest Broker Agreement executed on _____
- ❑ Site IRB submitted (IRB details to be provided); approved on _____
- ❑ Datavant Site License Agreement executed on _____
- ❑ Security and compliance reviews initiated and approved on _____

Datavant Technical Readiness:

- ❑ Technical walkthrough completed with Datavant on _____
- ❑ Datavant Portal account provisioned on _____
- ❑ Datavant site established on _____
- ❑ Datavant software installed in production environment on _____
- ❑ Test run of Datavant software on synthetic data on _____
- ❑ Datavant software run on production data on _____

LHB Submission Readiness:

- ❑ LHB Site Registration form completed on _____
- ❑ LHB Individual User Access form completed on _____
- ❑ LHB SFTP Access Granted on _____

Viral Variant Readiness (if applicable, see Appendix L):

- ❑ All steps under **Datavant Technical Readiness** above completed on _____
- ❑ Tokens generated and submitted to LHB for tested patients on _____
- ❑ Viral variant summary data submitted to N3C Data Enclave on _____
- ❑ Viral variant sequence data submitted to NCBI on _____

Part 1. Prepare and Tokenize Your Data Using Datavant

Introduction

Datavant provides de-identification and tokenization software that enables HIPAA-compliant data connectivity. Datavant software tools are installed on premises, behind your firewall, and Datavant does not have access to your data or systems. Datavant is building a robust, secure ecosystem of data exchange and looks forward to working with you.

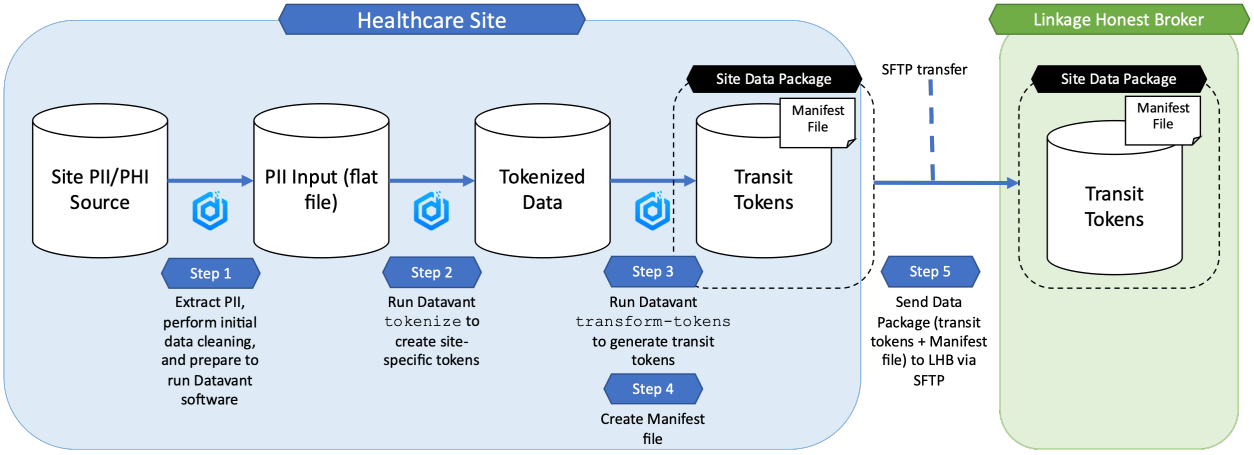
This section outlines the steps involved to use Datavant to de-identify, tokenize, and prepare your data to be sent to the Linkage Honest Broker

Datavant software has two modes:

- **Tokenize:** used when tokenizing identified data and creating site-specific Datavant tokens
- **Transform tokens:** used when transforming tokens between site-specific encryption keys

Sites will use both modes to tokenize their identified data and then prepare those tokens for submission to the LHB.

Overview of Data Flow from Healthcare Site to LHB



Section Contents:

Datavant compliance and security overview

To ensure a successful relationship and assist your technical diligence of Datavant, this section highlights key security features of Datavant and its software solutions, as well as answers to frequently asked questions during information security reviews.

Technical Diligence Documents Overview

Review these materials with your security and compliance teams. Let Datavant know if there are any questions.

- SOC 2 Type 1 Report; SOC 2 Type 2 Report: these reports demonstrate Datavant's compliance with SOC 2, which certifies that Datavant's systems and operations meet trust principles of security, availability, processing integrity, confidentiality, and privacy. **Available only upon request.**
- Security General Overview: overview of security topics often part of technical diligence not mentioned in our other documents.
- Security Policy Master List: master list of our current Datavant policies and latest revision dates.
    - **Sanitized Policies:** as a rule, we do not disclose our full company policies, however sanitized versions redacted to include outline, contents, summaries, version may be **made available upon request**.
- SDLC Overview: overview of our Software Development Lifecycle.
- Penetration Testing and Security Analysis Overview: overview of penetration testing and static analysis procedures on our code.

Datavant Software Privacy Certifications and Assessments

- Cryptographic Certification: Datavant De-identification: This document is a certification of the Datavant de-identification engine and tokens, demonstrating that they support HIPAA compliance and that Datavant tokens are cryptographically secure and anonymous. Note: This document is not a certification that any particular data set created from the Datavant engine is HIPAA compliant - the specific de-identification rules used on your data set must be certified separately by an expert.
- Certification of Datavant's Trusted Third Party Solution: This document is a certification of Datavant's Trusted Third Party security architecture, which is the technical and procedural framework used by Datavant to protect customer secrets (master salt/seed and customer encryption keys).

Datavant software configuration and token generation

The Datavant application in `tokenize` mode takes input PII and then, based on a specified configuration, creates de-identified, site-specific, and HIPAA-compliant Datavant tokens, and writes an output.

This configuration template specifies the following:

- columns and column orders to expect from the input file
- whether a header row is present in the input
- delimiter found within the input (e.g., pipe, comma, tab)
- what Datavant token designs to create
- what to write to the output file (i.e., what operations to perform, if any, on input fields)

The configuration template therefore controls what the *output* of the Datavant `tokenize` will be.

Every site will be provisioned a configuration template called `n3c_tokens`, which is passed as an input when running Datavant. Refer to Appendix H for the configuration template and list of tokens created when running the application.

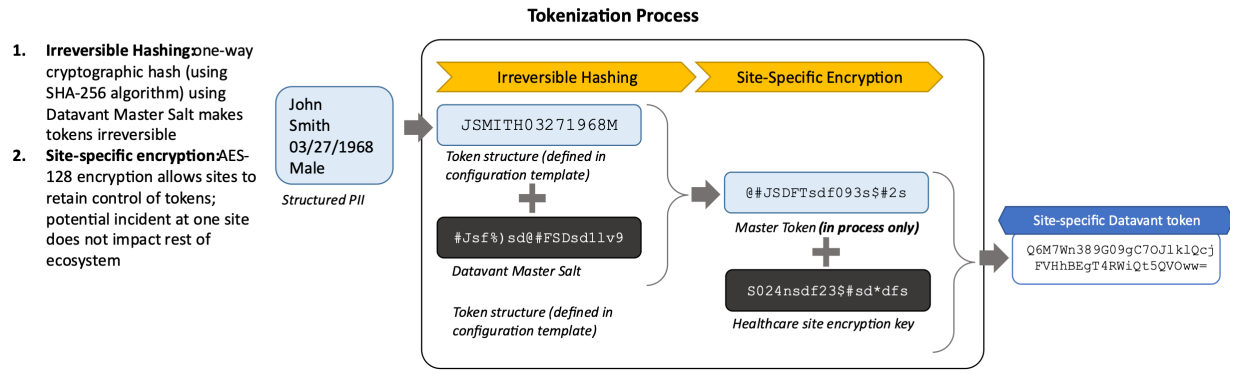
Datavant tokenize will create as many of the tokens defined in the configuration template as possible, provided the input data fields are valid.

The token designs are pre-certified via the HIPAA Expert Determination method, where an independent third-party expert statistician evaluates the statistical re-identification risk of the tokens to certify that the tokens can adequately be considered de-identified per the HIPAA Privacy and Security Rules. Refer to Appendix E for Expert Determination certificate.

Token Generation Process

The token generation process is a seamless process that occurs in a single step through the DeID tool. For illustrative purposes it is broken down into two sub-steps below:

1. **One-way cryptographic hash to create Master Token:** the underlying PII for a token is irreversibly hashed (i.e., you cannot regenerate the PII from the hash value) using SHA-256 and the Datavant Master Salt. This process creates the Master Token (in process only).
2. **Site-specific encryption to create site-specific tokens:** the Master Token is then encrypted using AES-128 with a site-specific encryption key. The same PII will always create the same set of Master Tokens, but the Master Token is never present in any output or log stream from DeID. Only the site-specific tokens are written to the output file.



Step 1.1 Prepare input file

Extract PII

To prepare your input file, first extract PII for your cohort following the instructions here. After extracting PII from your data source into a flat file, perform any necessary cleaning or file preparation.

Prepare your tokenization input file

The following data elements from your site-specific N3C cohort should be extracted to a UTF8 flat file for use as the input for the Datavant application.

Important
Note that the first column, record_id, represents a *pseudo ID* and is a unique identifier for each patient. This identifier must remain consistently associated with the same patient over time and across refreshes. These identifiers must also match the person identifiers included in the Phenotype delivered to the N3C Data Enclave.

Column Number	Column Name	Requirements for Use in Token Creation
1	record_id	N/A - not used to create tokens
2	First Name	Must have 2 or more characters in the set (a-zA-Z) and only contain characters from the set (a-zA-Z '-. and space character)
3	Last Name	Must have 2 or more characters in the set (a-zA-Z) and only contain characters from the set (a-zA-Z '-. and space character)
4	Date of Birth	Must be one of: MM/DD/YYYY, MMDDYYYY, MM-DD-YYYY, YYYY/MM/DD, YYYYMMDD, YYYY-MM-DD. Must be a valid date (for example, date cannot be 2/31/2020).
5	Gender	All versions can take in characters from the set (MmFf). V3.2.1 and later can take in the words Male or /Female, case insensitive. Must not contain any other characters.
6	SSN	Must be a valid US Social Security Number in format FNN-NN-NNNN or FNNNNNNNN, where F is a numeric digit from 0-8 and N is a numeric digit from 0-9. First three digits cannot be 666 and none of the three sections (first three, middle two, or last four digits) can be all zeroes. NNN-NN-NNNN or NNNNNNNNN
7	ZIP	Must be a valid US 5 or 9 digit zip code, with an optional hyphen after the 5th digit in 9 digit zip codes (NNNNN, NNNNN-NNNN, NNNNNNNNN). Additionally, zip codes not issued by the US government will be considered invalid.
8	Email	A valid email must contain exactly one character '@', and the prefix and domain (i.e., the portions to the left and right of the unique address operator '@') must both consist of only letters (a-zA-Z), numbers, underscores, periods, and dashes. An underscore, period, or dash must be followed by one or more letter or number. The domain must contain at least one period and the portion following the last period in the domain name must have at least two alphabetic characters, and no non-alphabetic characters. The prefix and domain must each have at least four distinct characters, and the domain must have at least five characters, not counting the periods. The email prefix (before the "@" symbol) and the domain (after the "@" symbol) must have 4 or more distinct characters. Must contain an @.
9	Cellphone	Field should contain nine or ten numeric digits with an optional separator (space, dash, hyphen or period). Also, can optionally start with a +1 or have parentheses around the area code. Field will be stripped of all non- numeric characters and leading +1 before token creation.

When creating Datavant tokens, note that the above requirements must be met for each individual input field of PII. If the requirement for any field in a token is not met, the token will not be generated for that record. For example, Token1 relies on [LastName + FirstInitial + Gender + DateOfBirth]. If only the Gender is invalid and all other fields are valid, Token1 will not be generated for that record.

Data hygiene best practices

As a result of Datavant's token certification and in accordance with industry best practices, Datavant software ***automatically*** performs some data operations prior to tokenization. These operations ensure consistent tokenization. For example, for the purposes of tokenization, the first names "Mary Jo" and "Mary-Jo" and "MARYJO" are identical. Additionally, we maintain recommendations for your identified data to maximize match rates. Refer to Appendix G for complete best practice documentation.

Step 1.2 Prepare environment to run Datavant software

Datavant is installed and run on-premise at your site. The software is available as both a command line interface (CLI) as well as a graphical user interface, Datavant desktop. Datavant recommends using the CLI because of its flexibility and automation capabilities, and this section assumes the CLI is being used. If you would like to use Datavant desktop, refer to Appendix J.

> **Important**
> The following sections contain details specific to v4 of Datavant software. If you are an existing Datavant organization, we encourage you to use v4 because of its streamlined installation and increased security through use of user-specific credentials. However, v3 instructions are available in Appendix K.

Follow these steps to ensure your environment meets the technical requirements:

- Ensure your operating system is supported:
  - Mac
    - macOS 10.13 (High Sierra) or later
  - Windows
    - Windows 8.1 or later
    - Windows Server 2012 R2 or later
  - Linux (command line executable only)
    - Ubuntu 13.04 or later
    - Red Hat 6.9 or later
    - CentOS 6.10 or later
- Ensure your machine is appropriately sized. For optimal performance, we recommend at least a 3GHz Quad-Core Processor and 8GB RAM.
- Ensure your environment has the appropriate networking configuration. The Datavant application communicates with Datavant's Amazon Web Services (AWS) environment to retrieve configurations and securely retrieve encryption keys. The following endpoints must be accessible *outbound over port 443*:
  - sec.datavant.com
  - auth.datavant.com
  - api.datavant.com

> **Important**
> If you are upgrading from v3 to v4, ensure you have the above endpoints accessible. Depending on which version of v3 you upgraded from, you may have only allowed a previous set of endpoints.

- Download and retrieve the following from the Datavant portal:
    - Datavant CLI executable for your operating system
    - Your user-specific credentials (generate and download from the Datavant portal)

Step 1.3: Run Datavant on the input file to create site-specific tokens

Files Required
This step requires the following files:
- Datavant CLI executable (download from the Datavant portal)
- Your user-specific credentials (generate and download from the Datavant portal)
- Input file generated in Step 1.1

After downloading the Datavant CLI executable and preparing your input file, you are ready to run the software.

- Ensure the executable and input file are both stored in the same directory. Then, navigate to that directory from the command prompt.
- Run Datavant CLI with the `tokenize` sub-command on the input file (see example commands below).

After this step, you will have a de-identified output and Datavant tokens in your site's specific encryption key.

Example commands for Datavant `tokenize`

Windows

```
type credentials.txt | .\Datavant_Win.exe tokenize -s yoursite -c
n3c_tokens -i input.csv -o output_tokenize.csv
```

Mac

```
cat credentials.txt | ./Datavant_Mac tokenize -s yoursite -c
n3c_tokens -i input.csv -o output_tokenize.csv
```

Linux

```
cat credentials.txt | ./Datavant_Linux tokenize -s yoursite -c
n3c_tokens -i input.csv -o output_tokenize.csv
```

where

- `Credentials.txt` is a file containing your user-specific credentials, retrieved from the Datavant portal
- `yoursite` is the name of your site, also available in the Datavant portal
- `input.csv` is the input file created in Step 1.1
- `output_tokenize.csv` is the name of the file where the output will be written

Example input and output with **`tokenize`** sub-command

Input:

```
record_ID|First Name|Last Name|Date of Birth|SSN|Gender|ZIP|Email|Cellphone
6131918848|Alicia|Duran|1990/05/02|389299386|F|10005|aliciaduran@gmail.com|3329348841
9507913578|Ellie|Ortiz|1991/03/22|120174556|F|94108|ego775@live.com|4859332123
8649099455|Lee|Gardner|1976/02/23|893756602|M|72210|gardner21@outlook.com|2319938412
4865748664|Jason|Pacheco|1984/10/11|470799204|M|19082|jpacheco_34@hotmail.com|5659492311
3063354932|Lily|Sutton|1980/01/30||U|02903|lilylouisesutton@gmail.com|8753002918
7239604172|Hanna|Vaughan|1953/02/23|322218840|F|79918|hanv34039@yahoo.com|4858332101
7281305057|Darell|Simmons|1934/12/01|193900587|M|97408|darrellsimmons18@gmail.com|2149093822
2456673715|Marcus|Goodwin|1930/07/17|329894893|M|96823|mg1994@aol.com|3430490032
2005931419|Lucia|Patel|1974/08/29|129488492|F|88435|lepatel_9433@yahoo.com|3455993848
3921770024|Erik|Moon|1995/03/04|399288293|M|48219|erikmoon@gmail.com|5064938421
```

Identified data with record_ID (pseudo ID)

Output:

```
record_ID|token_1|token_2|token_3|token_4|token_5|token_16|token_29|token_30|token_6|token_7|token_8|to
6131918848|PW+CdlbX+1rpRxg5qwQveh5sVQj6y4zbEGtTFkBPbUc=|1Pqd2uxCNg8n2OemAgTXV33n86uh0hYceIR2IpXWjPw=|0C
9507913578|pGzmX23tKsHRE1xqpj6EjYSm7Gp9o5/Afeeaw8z+aBw=|kf9cun4Yu+6v8zweHsqr6dR/mjbqY/jpI5tw/60g0Tk=|EF
8649099455|JnpHkzNBr/HtpyyPfDRPh4NuWYXSD+EdUxNCIIDd1tY=|puvlE8ILxCiRLn4Tao9zE0zq3asrhf4rG0n7+DVpZX0=|FE
4865748664|HXAvTot1cQf7V38g/dnwkk8dDS4G0qA7d8jI3bycg/E=|lbdCbf6Es2Bj57rqY6SL6Qvx+uJ4z54VI9RzBSYjW8c=|cy
3063354932|aT9zgITWDQCsn3YirYWLAehqBZKMLdIg1opXkeHpmpg=|jMj3AvTdEunJGv3+PCvBHxiamIUDoJmxRsSYhWVd0S8=|mb
7239604172|KvgyURVF499bg241OmBRK+ffbgbLHJlUtfnPrMRkdu4=|Ch02eDk5W9vu0UQUjtaWmdsMAEiwf6aRwBoC3M1O4LY=|TG
7281305057|XRPWaUC1FQkWE5sm5iJ1ZF1J3nrtCL0eSzOjYBLJoFo=|HbU1AmHv++llLucMariGurBOxtopPbw3G+dlnZmD7Mk=|oo
2456673715|zi86uqjN35IlqC2ov02y6U4SF9/FW1S33dINCPEymGE=|p/Q4Zm7MB5MkDK/vIeJpSUPAm/pJgY/AbM8ePwZOh8Q=|wN
2005931419|FHAnSiFgTPUWV5iDhmbE6skTU4RfJm/oj0KvgC8vpgg=|xx/JWspslzwERI/7exdLgRUGX4AkSTBOdNFuHR8i6uY=|IH
3921770024|nZZq4R30eSkUGmuU/xF89vHAfJ1T/gs9EFeZMIJ2lYw=|bAnykaCwvT7WZZQ5YX4ygIjr/2paK3DtzR4sQVf+mgs=|OA
```

Set of Datavant site-specific tokens with record_ID (pseudo ID) and demographic data is removed

Step 1.4: Run Datavant to transform site-specific tokens to transit tokens

Files Required

This step requires the following files:

- Datavant CLI executable (download from the Datavant portal)
- Your user-specific credentials (generate and download from the Datavant portal)
- Output file generated in Step 1.3

After running tokenize, your site will have Datavant tokens in your site's encryption key. The use of site-specific encryption protects all N3C participating organizations, and you should never send your Datavant tokens in your site-specific encryption to another site or entity.

To send your data to the LHB, you will run Datavant CLI with the transform-tokens sub-command. This step will transform your site-specific tokens to transit tokens, which are encrypted in a shared encryption specific to the recipient (i.e., in an encryption specific to your site as the sender, and the LHB as the recipient). The input file in this step is the output from running Datavant CLI with the `tokenize` sub-command.

- Ensure the executable and the output file generated in Step 1.3 are both stored in the same directory. Then, navigate to that directory from the command prompt.

- Run Datavant CLI with the `transform-tokens` sub-command, with `output_tokenize.csv` used as the input file (see example commands below).

Example commands for Datavant **`transform-tokens`**

Windows

```
type credentials.txt | .\Datavant_Win.exe transform-tokens --to n3c -s
yoursite -i output_tokenize.csv -o output_transformed.csv
```

Mac

```
cat credentials.txt | ./Datavant_Mac transform-tokens --to n3c -s
yoursite -i output_tokenize.csv -o output_transformed.csv
```

Linux

```
cat credentials.txt | ./Datavant_Linux transform-tokens --to n3c -s
yoursite -i output_tokenize.csv -o output_transformed.csv
```

where

- `Credentials.txt` is a file containing your user-specific credentials, retrieved from the Datavant portal
- `yoursite` is the name of your site, also available in the Datavant portal
- `output_tokenize.csv` is the output from Step 1.3
- `output_transformed.csv` is the name of the file where the output will be written

Example input and output with **`transform-tokens`** sub-command

Input:

```
record_ID|token_1|token_2|token_3|token_4|token_5|token_16|token_29|token_30|token_6|token_7|token_8|token_12|token_18|token
6131918848|PW+CdlbX+1rpRxg5qwQveh5sVQj6y4zbEGtTFkBPbUc=|1Pqd2uxCNg8n2OemAgTXV33n86uh0hYceIR2IpXWjPw=|0ClJ6xQ24h4S/1UhfJaI+82
9507913578|pGzmX23tKsHRE1xqpj6EjYSm7Gp9o5/Afeeaw8z+aBw=|kf9cun4Yu+6v8zweHsqr6dR/mjbqY/jpI5tw/60g0Tk=|EFjDX7F8eSVDvXiWnLPhcUQ
8649099455|JnpHkzNBr/HtpyyPfDRPh4NuWYXSD+EdUxNCIIDd1tY=|puvlE8ILxCiRLn4Tao9zE0zq3asrhf4rG0n7+DVpZX0=|FEds+85XrJG69h1YJ12LWx+
4865748664|HXAvTot1cQf7V38g/dnwkk8dDS4G0qA7d8jI3bycg/E=|lbdCbf6Es2Bj57rqY6SL6Qvx+uJ4z54VI9RzBSYjW8c=|cy0u787Ads5gCIGyWV6HHT/
3063354932|aT9zgITWDQCsn3YirYWLAehqBZKMLdIg1opXkeHpmpg=|jMj3AvTdEunJGv3+PCvBHxiamIUDoJmxRsSYhWVd0S8=|mbu9h4fmcudWBl87I1UHiz7
7239604172|KvgyURVF499bg241OmBRK+ffbgbLHJlUtfnPrMRkdu4=|Ch02eDk5W9vu0UQUjtaWmdsMAEiwf6aRwBoC3M1O4LY=|TGDKsaEoVPwi7iHjPGRwS2L
7281305057|XRPWaUC1FQkWE5sm5iJ1ZF1J3nrtCL0eSzOjYBLJoFo=|HbU1AmHv++llLucMariGurBOxtopPbw3G+dlnZmD7Mk=|oohIjsCUkSipaK1D24lKY2a
2456673715|zi86uqjN35IlqC2ov02y6U4SF9/FW1S33dINCPEymGE=|p/Q4Zm7MB5MkDK/vIeJpSUPAm/pJgY/AbM8ePwZOh8Q=|wNABct9kL5QBAqVT5MUaE++
2005931419|FHAnSiFgTPUWV5iDhmbE6skTU4RfJm/oj0KvgC8vpgg=|xx/JWspslzwERI/7exdLgRUGX4AkSTBOdNFuHR8i6uY=|IHWt1UB6yw7bGRbZZ+VVAS1
3921770024|nZZq4R30eSkUGmuU/xF89vHAfJ1T/gs9EFeZMIJ2lYw=|bAnykaCwvT7WZZQ5YX4ygIjr/2paK3DtzR4sQVf+mgs=|OAa+IlErzWz1yzs+NCjor0y
```

Input to `transform-tokens` *is the output from* `tokenize`

Output:

```
record_ID|token_1|token_2|token_3|token_4|token_5|token_16|token_29|token_30|token_6|token_7|token_8|token_12|token_18|token_
6131918848|4Yajudv1JF9BJzMRRDF3w1hMIaBCkTrDHiA712dMgGI=|GvV4VvOjIiLaYS24mj0HD+M2+eshlXb1BbqJJF6qxRE=|8dkK5kt9/4acaiKcWDZbXO2
9507913578|UeL4b1XvSu57JHT3SgE3kGU5GcRNDktYWt83iDCxM9c=|OXbleoujQDmGVRd+Qo5A1u4F5M1pgQ806A6BP+irFs8=|LWXbj9KAabnSnuL/q4uzIFL
8649099455|Uxv+Qr3EoPmCMjUZy4W+A/QFkQ+3AsMhK0rGl483uq8=|LX+fqLQcpIquhNt+HF+xLnAEoSTpl7EvA/gvVU8DvjE=|1Y4SWVovpjPqix9TEqQypn/
4865748664|Cy0X9CI3kD6lLwEQGneBzq7D9j9gxHm64OK+o2I5wB4=|iKJc18Q8HbZucW6FZN1BzbFo1Olotll0rUx4SplvgSo=|fz6YAIDvlLlyrhiuQJ1wh3S
3063354932|eamv3yxT0pPzn8IegIlEzDK10OgAWIaVqgARk3SsMlw=|mzln5XsI/kdQHw/U+rFpBo0+3USTvtjHmcdIqCQOpGo=|U7euaHyzaXI8VbkhLvW3L74
7239604172|1yruMTxy/jQW+JQpJav3ei5B4W8V7frCi2YZwNeOSaY=|XXHupKnLpnbnwluHeSPKZt5V/PzlQoCCkxJVrn5QTbU=|z5TVenE0wrDI3mrU/hOTIsc
7281305057|90yRPsSAowDg0i1w+oOd0U7FXI4cq1Ks2mJ0CVcLCVM=|W3Y1CCyPraOFldyOqMBBfMS7AYgXOOMi2bRlD+0FTPs=|/azk92u/PlvcoMd9wGpRwFL
2456673715|yYi3x8hHXT2UyzYnUdlIowv9LNSI2/fKuxHcKV1nwgI=|WXp/oJ2iso+RdvuTPTYQp8qumRvm/6An6ty55egXevE=|mIas1QnvKpfOj4TgxLms6yn
2005931419|CCP3GNylvZBMrKQ8wDRQNdnHquncFXQqCJZk+9dHh0Q=|mmCwlaVslq6x7+o4rEl1RvvqRQ7vudaochUBl8OtyDA=|RC0mU2gBTohvyBErFVpIj/d
3921770024|Xyzc+Gxleeq9efC8/Dp/Ue0BhKWY0pD+cCrbzsy/Fz8=|DbISIqIOD9k2w6CCX0+tIpu1CjjJF1loP+RTJWLt20c=|Cv0lS3EMhZ74s3aoZ/O1pAI
```

Tokens transformed into transit tokens, record_ID is unchanged

Step 1.5: Create Manifest file and data package to send to the LHB

After running transform-tokens, you are ready to create the data package, which contains both the transit tokens created in Step 1.4, as well as a Manifest file with metadata about the submission.

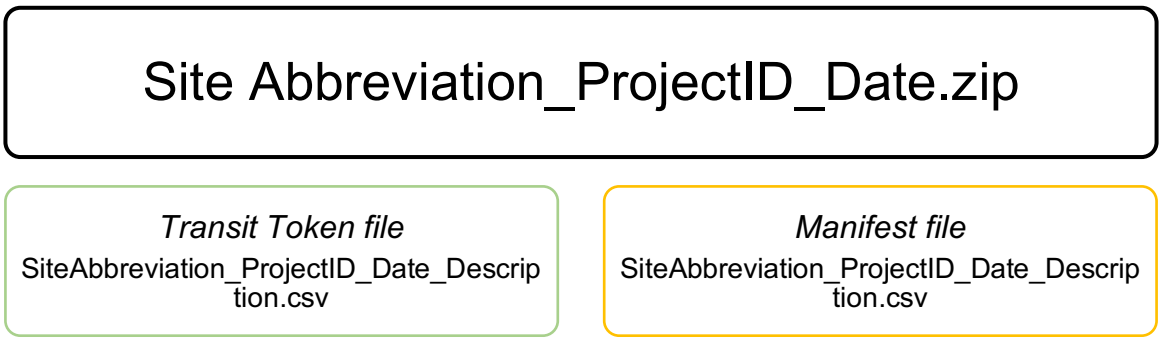
The Manifest file should contain the following data elements:

Header	Definition	Example
site_abbreviation	(same as used for the N3C manifest if participating with NCATS / N3C)	UNC
site_name	Your institution name - Text version	University of North Carolina
project_id	The identifier for this token set (ie. N3C, All of us, etc.)	N3C
number_of_token_sets	The number of patients who have token sets in the file	2
contact_name	Best contact name for Token creation	Jane Doe
contact_email	Best contact email for Token creation	Jane_doe@ohdsi.org
run_date	Date the Tokens are created	2021-01-01
run_time	Time the Tokens are created	10:23
update_date	Date the demographic data was last extracted from data source	2021-01-01
next_submission_date	Anticipated Date when the next Token creation will occur	2021-01-08
datavant_tool_config_name	Name of the configuration used for the Datavant Token tools *Note: should be the same in all Manifest files for a given site*	n3c_tokens
datavant_tool_site_name	Name of your site used for the Datavant Token tools *Note: should be the same in all Manifest files for a given site*	unc_medicine
datavant_tool_send_to	Organization name used when creating the Transit Tokens with Datavant Link tool *Note: should be the same in all Manifest files for a given site*	n3c

The data package should have the following files:

- Output token file from Step 1.4, saved as a .csv file
    - Naming convention for file: SiteAbbreviation_ProjectID_Date_Description.csv
    - Description: TOKENS
    - Date Format: YYYYMMDD
    - Example file name: UNC_N3C_20210401_TOKENS.csv
- Manifest file containing metadata about submission, saved as a .csv file

- Naming convention for file: SiteAbbreviation_ProjectID_Date_Description.csv
- Description: MANIFEST
- Date Format: YYYYMMDD
- Example file name: UNC_N3C_20210401_MANIFEST.csv



Save the data package contain the token file and the Manifest file as a .zip file

- Naming convention for file: SiteAbbreviation_ProjectID_Date.zip
- Date Format: YYYYMMDD
- Example file name: UNC_N3C_20210401.zip

REGENSTRIEF X DATAVANT

Part 2. Submit Data Package to Linkage Honest Broker

Introduction

The Regenstrief Institute functions as the Linkage Honest Broker, receiving data packages from participating sites and performing privacy-preserving record linkage. This process creates a linkage "map" of Datavant tokens that correspond to the same patient. No files sent to the LHB are shared with the Data Enclave and the LHB only receives tokens from sites.

Prerequisites

Before you can submit your data package to the LHB, you must complete the Linkage Honest Broker Agreement.

Step 2.1: Complete LHB registration

Important
Do not start LHB registration until your organization has completed the Linkage Honest Broker Agreement.

Important
You must create your private and public SSH key before filling out the Individual User Access Form by going to https://www.regenstrief.org/n3c-lhb/ and follow the instructions.

Register your site and team members with the LHB using this link: https://redcap.link/N3C

1. Complete the Site Registration Form which will provision the proper firewalls. In the form, list the site personnel who require LHB SFTP account(s).
    1. Formal site name (full name of your institution)
    2. Project (i.e., N3C, All of Us, Mortality, Viral Variant, etc. Check all that apply.)
    3. Principal investigator's name
    4. Public Static IP or CIDR Block
    5. Names (first and last) and email addresses for LHB SFTP users
    6. Primary Technical Contact Name and email address for your site
2. After completing the Site Registration Form, an e-mail to the Individual User Access Form will be sent to the individuals listed in the Site Registration Form. The email will be sent from RILHB@Regenstrief.org. The Individual User Access Form requires the following information:
    1. First name, last name, and middle initial (optional)
    2. Project (N3C/De-duplication, Imaging, Mortality, etc.)
    3. Email address
    4. Phone number
    5. SSH Public Key (upload)
        i. Go to https://www.regenstrief.org/n3c-lhb/ and follow the LHB Site Configuration Instructions or open the embedded attachment.

        LHB+Site+Data+Inbox+Configuration.pdf

        ii. This ***MUST*** be completed prior to submitting the Individual User Access Form

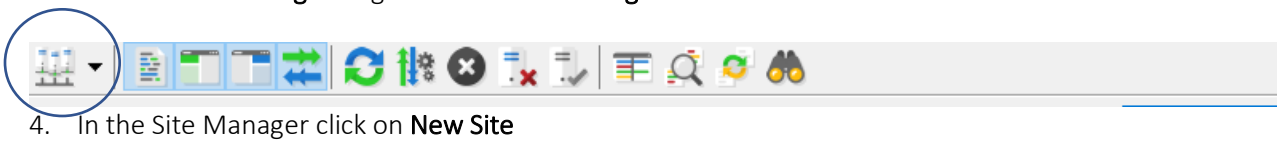
3. After the Individual User Access Form is submitted, users will receive an e-mail with your LHB SFTP username.
4. After receipt of your LHB username, set up the LHB SFTP (see Step 2.2 or go to https://www.regenstrief.org/n3c-lhb for detailed instructions)

Important
The LHB SFTP is separate from the NCATS SFTP, which is used to submit data to the N3C Data Enclave. Before submitting data, please confirm you are using the correct SFTP.

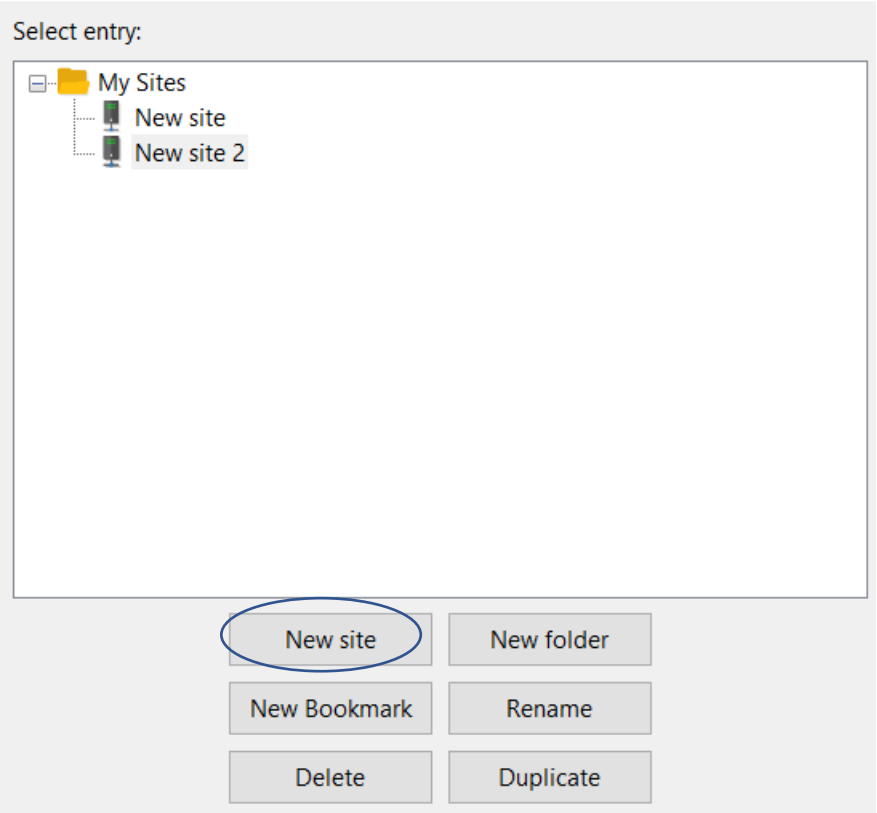
Step 2.2: Submit the data package to the LHB Secure Landing Zone

To submit your data package, you first need to set up your SFTP account. You only need to complete this once. Follow the below instructions

1. Download FileZilla at https://filezilla-project.org/download.php?platform=win64
2. Open **FileZilla** program
3. Click on **Site Manager** or go to **File → Site Manager**

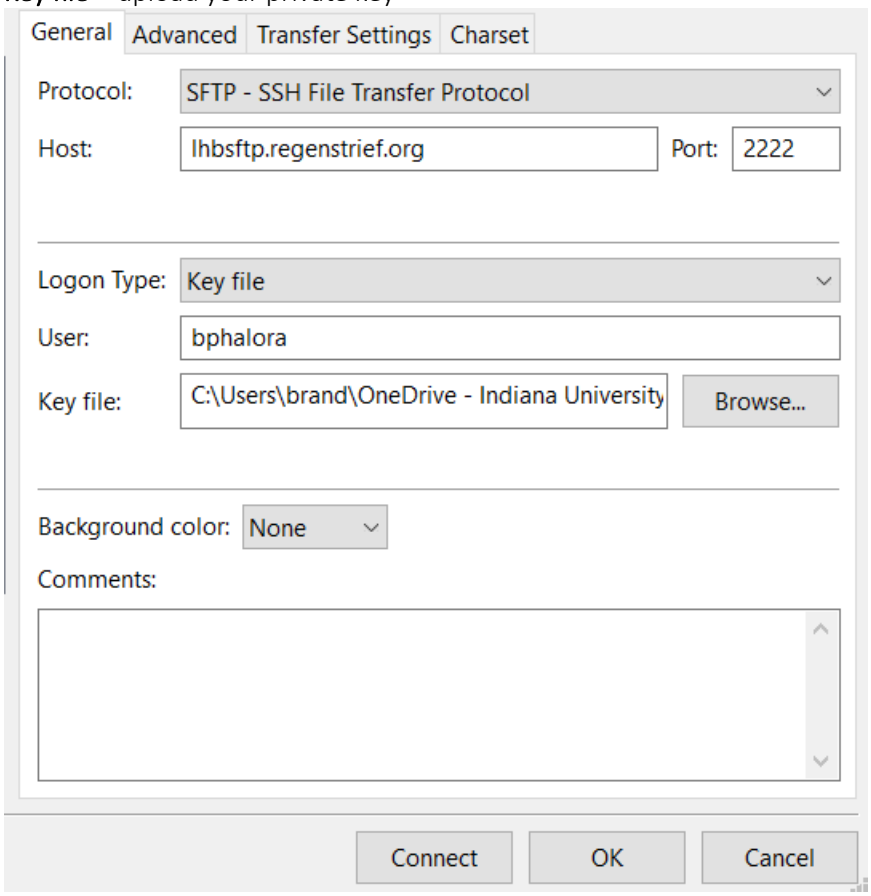


4. In the Site Manager click on **New Site**



5. Enter the following:
   a. **Protocol** (right hand side of program) – SFTP SSH File Transfer Protocol
   b. **Host** – lhbsftp.regenstrief.org
   c. **Port** – 2222
   d. **Logon Type** – key file

e. **User** – your LHB provided username
f. **Key file** – upload your private key



6. Click **Connect**. You will know connection to the SFTP is successful by looking at the status box at the top

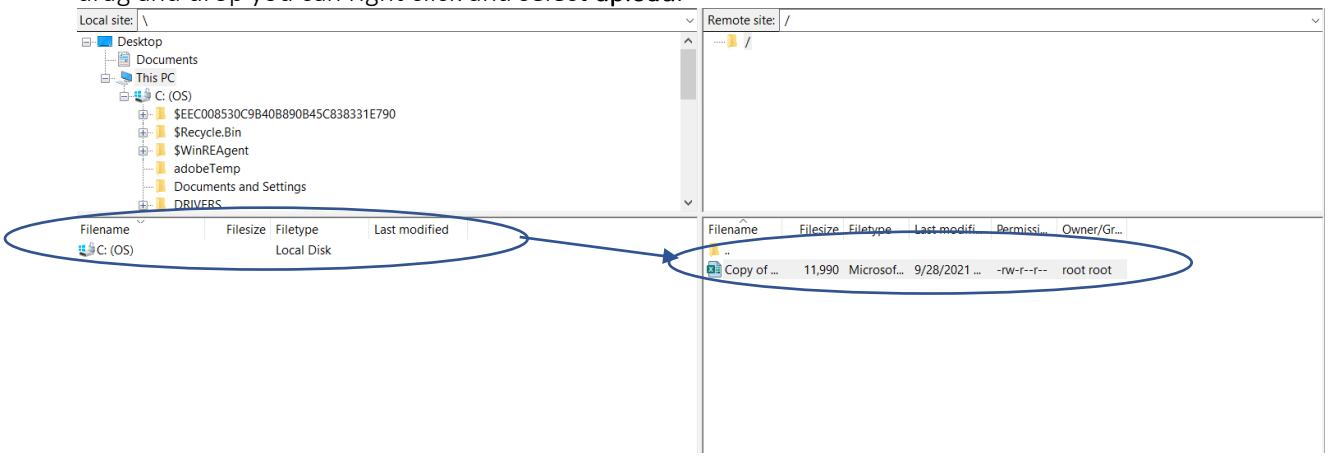
```
Status:    Connected to 34.133.160.229
Status:    Retrieving directory listing...
Status:    Listing directory /
Status:    Directory listing of "/" successful
```

7. For each file submission you will need to log in to FileZilla and connect to the LHB SFTP. You do not need to re-enter the information each time. Simply click connect.

File Submission to SFTP:

1. Log in to the LHB SFTP using FileZilla.
2. On the left-hand side navigate to the file you wish to upload to the LHB SFTP. The SFTP will show files on your C:\. Select on the .ZIP file you wish to upload to the LHB and simply drag it to the right-hand side of the screen under the Remote Site column (filename). If you do not wish to

drag and drop you can right click and select **upload**.



3. Confirm receipt of successful upload. Check the **status box** for successful upload

REGENSTRIEF X DATAVANT

Frequently Asked Questions

How often should I submit my token files to the Linkage Honest Broker?

You should aim to submit a data package to the Linkage Honest Broker at the same cadence as you submit your data to the N3C Data Enclave.

I am having issues with the Datavant software.

Submit an N3C Service Desk ticket at https://n3c-help.atlassian.net/servicedesk/customer/portal/2:

- Select N3C Record Linkage from the main menu.
- Under PPRL Question Category, select "Datavant Software Questions."

I am having issues uploading the data package to the Linkage Honest Broker SFTP.

Submit an N3C Service Desk ticket at https://n3c-help.atlassian.net/servicedesk/customer/portal/2:

- Select N3C Record Linkage from the main menu.
- Under PPRL Question Category, select "Linkage Honest Broker Questions."

I cannot remember my LHB login credentials.

Submit an N3C Service Desk ticket at https://n3c-help.atlassian.net/servicedesk/customer/portal/2:

- Select N3C Record Linkage from the main menu.
- Under PPRL Question Category, select "Linkage Honest Broker Questions."

I accidentally submitted the wrong file.

Submit an N3C Service Desk ticket **immediately** at https://n3c-help.atlassian.net/servicedesk/customer/portal/2:

- Select N3C Record Linkage from the main menu.
- Under PPRL Question Category, select "Linkage Honest Broker Questions."

In the ticket, include:

- Name of the file sent to the LHB SFTP Secure Landing Zone
- Your project name, site abbreviation, and date/time of submission

I need to remove the transit token file or Manifest file sent in my data package submission.

Submit an N3C Service Desk ticket **immediately** at https://n3c-help.atlassian.net/servicedesk/customer/portal/2:

- Select N3C Record Linkage from the main menu.
- Under PPRL Question Category, select "Linkage Honest Broker Questions."

In the ticket, include:

- Name of the file sent to the LHB SFTP Secure Landing Zone
- Your project name, site abbreviation, and date/time of submission

The file I sent contained patient identifiers. How do I remove the file?

Submit an N3C Service Desk ticket **immediately** at https://n3c-help.atlassian.net/servicedesk/customer/portal/2:

- Select N3C Record Linkage from the main menu.
- Under PPRL Question Category, select “Linkage Honest Broker Questions.”

In the ticket, include:

- Name of the file sent to the LHB SFTP Secure Landing Zone
- Your project name, site abbreviation, and date/time of submission

Appendix A. Datavant Security General Overview

Details regarding certain policies that are often brought up or discussed and not mentioned in other documents are listed below. If you have further questions on a specific policy or one not mentioned here or elsewhere, feel free to contact Datavant at n3csupport@datavant.com.

SOC 2 Type 2 Report

- **SOC 2 Type 1 Report; SOC 2 Type 2 Report:** These reports demonstrate Datavant's compliance with SOC 2, which certifies that Datavant's systems and operations meet trust principles of security, availability, processing integrity, confidentiality, and privacy. Our auditors require that our SOC 2 Type 2 Report be provided only upon request with Confidentiality agreements in place and an expiration date.

Internal Network Security

- Datavant uses a WeWork space. Because all our services are cloud hosted, we do not own our own IT infrastructure. We use a zero-trust model. Every user authenticates against our cloud services individually over SSH, HTTPS, SFTP, etc. Our company network grants no access privileges beyond being a whitelisted IP address that allows for SSH connection attempts.
- As such, we do not utilize VPNs or other services for out of office work. Every user authenticates directly against the service they are signing into from their location when necessary.
- Our company network is provided by WeWork.

Facilities Security

- Datavant's offices are located in a WeWork space.
    - WeWork administers security for our office, which is in an office tower shared with non-WeWork tenants as well. There is an on-duty security guard in the building lobby, and keycards are necessary to take the elevator to our building level. Beyond that, users must use keycards to access the general WeWork space, and again to access our offices. Visitors are controlled and logged by WeWork at a front desk for the WeWork space.
- Datavant's IT infrastructure is handled by AWS. We have no data center facilities to secure.

Software Updates

- We follow an agile development model with continuous integration. We are frequently making minor changes to our software. Running the most recent version is not necessary for security.
- We do not have a set release schedule for updates.
- We will notify if a major security issue causes a need for an update through our implementations contact.

Business Continuity and Disaster Recovery

- Datavant has a comprehensive Business Continuity and Disaster Recovery Policy (BCDR). It covers an overview of preparations and recovery actions in case of disruptions to our information assets. In support of this, we have the following policies:
    - Data Management Policy
    - Disaster Recovery Plan

Onboarding

- Training: All employees undergo a set of security and privacy related new hire orientations and trainings.

  - Privacy Orientation: General overview on privacy regulations and issues, HIPAA, other regulatory concerns, as well as how our company keeps these in the forefront.
  - Security Orientation: General rules and good practice on security, along with an overview of our security policies. Employees are assisted with setting up and configuring multi factor authentication, password managers, etc.
  - SDLC Orientation (Engineering Only): Familiarization with our SDLC and Change Management Procedure for making changes to our software, along with other associated procedures.
  - HIPAA Training: Third-party HIPAA training covering HIPAA privacy rules and regulations, satisfying the HIPAA training requirement. Engineers, product, and implementations take further training in HIPAA security, with a more in depth look at the technical aspects of HIPAA rules and regulations.
- Background checks: All employees are background checked before onboarding. Checks are done by a third-party vendor and include SSN Trace, Criminal Background, OFAC, FACIS Level 1, OIG/SAM, Education Verification, Employment Verification

Appendix B. Security Policies List

1. Datavant has a number of security policies. All policies are reviewed for changes quarterly and changed if necessary.
2. Below is a list of our security policies and the most recent revision date. As a general rule, we do not disclose our full company policies, however sanitized versions redacted to include only version, outline, contents, summaries may be made available upon request.
    - Acceptable Use Policy
    - Access Control Policy
    - Auditing and Logging Procedure
    - Background Check Procedure
    - Business Continuity and Disaster Recovery Policy
    - Change Management Procedure
    - Customer Password Reset Procedure
    - Data Management Policy
    - Disaster Recovery Plan
    - Discipline Procedure
    - Exceptions Procedure
    - Facilities and Physical Security Policy
    - Incident Response Policy
    - Information Security Management Program Overview
    - Information Security Policy
    - Investigation Policy
    - IT Hardware Security Policy
    - Malicious Code Procedure
    - Password Management Procedure
    - Risk Management Procedure
    - Software Development Lifecycle Policy
    - Standard Secure Configuration of Services Procedure
    - Training and Awareness Procedure
    - User Provisioning and Deprovisioning Procedure
    - Vendor Management Procedure
    - Vulnerability Management Procedure

Appendix C. Software Development Lifecycle Overview

1. **Software Development Lifecycle (SDLC) Policies:** Datavant uses an SDLC Policy along with a Change Management Procedure to effect changes effectively and securely to our software.
    - SDLC Policy: Outlines the general guidelines and requirements for developing and implementing new software and make changes to existing software while ensuring that development work results in secure, compliant software.
    - Change Management Procedure: Outlines the exact steps to be taken to make a change in our codebase, and the necessary approval steps.
2. **Development model summary:** Datavant follows an agile development model, with technical testing and deployment controls enforced by GitHub.
    - Development and Production environments are kept separate.
    - Software development happens in 3 phases:
        - Concept: defining product, security and privacy requirements, technical specs.
        - Development: code development and testing.
        - Release/Maintenance: Deployment, support, updates.
3. **Change procedure overview**
    - Configuration Changes: Changes to configurations or settings, which follow the SDLC process but are subject to less approvals.
    - Regular Changes: Normal changes to software that follow the regular SDLC process, can be major or minor.
    - Change Approval Process:
        - Automated testing must fully pass before changes can be made; testing is done on every change.
        - Second developer must approve all changes, for major changes, a change authority must approve the change.

Appendix D. Penetration & Code Security Testing Overview

1. Penetration Testing
    - Datavant contracts with third party auditing firms to perform penetration tests on our on-prem command line applications, Desktop applications, and web application.
    - We do not penetration test internally.

   Our penetration tests are conducted annually, and attestations of these tests are available with a Mutual Non-Disclosure Agreement in place (site's or Datavant's), or the Site License Agreement in place.

   The Datavant Mutual Non-Disclosure Agreement can be reviewed and executed through DocuSign, if your institution accepts electronic signatures.
2. Static Analysis
    - We use the static analysis tool Bandit to check our code.
    - Static analysis is heavily integrated into our SDLC. We run static analysis on our entire codebase for every change made to our code. We have our system configured to prevent the change unless all tests are passed, one of which is the static analysis tests.
    - Any issues from our static analysis tool counts as a failure and prevents the change. This is enforced by GitHub, our Continuous Integration system.

Appendix E. HIPAA Expert Determination Certification

Datavant's Certification that the N3C CDM Tokens generated by Datavant's De-Identification Engine are de-identified data per the HIPAA Expert Determination method is provided by Scheuren-Ruffner Consultants, respected statistical re-identification experts. Datavant is always working on new token designs for/with its customers and updates the Certification frequently.

Summary of the current certification, Version 4, June 2021, is available with either the N3C Confidentiality and Disclosure Agreement (CDA), Mutual Non-Disclosure Agreement (site's or Datavant's), or the Site License Agreement in place.

The Datavant Mutual Non-Disclosure Agreement can be reviewed and executed through DocuSign, if your institution accepts electronic signatures.

Appendix F. Trusted Third Party Certification

Datavant's Certification for our privacy framework, which is the technical and procedural framework used for its privacy engineering framework, which includes protection and distribution of the cryptographic secrets and encryption keys is provided by provided by Scheuren-Ruffner Consultants, our certifier for the de-identified data status of the tokens per the HIPAA Expert Determination method.

The current certification, April 2021, is available upon request with either the N3C Confidentiality and Disclosure Agreement (CDA), Mutual Non-Disclosure Agreement (site's or Datavant's), or the Site License Agreement in place.

The Datavant Mutual Non-Disclosure Agreement can be reviewed and executed through DocuSign, if your institution accepts electronic signatures.

Appendix G. Tokenization File Preparation and Data Hygiene

The underlying format of your data is integral to the tokenization and ultimately patient linking process. As a result, standardized data hygiene across the Datavant ecosystem is crucial. Please refer to the Data Hygiene Best Practices for both recommendations and automatically applied operations for identified data.

Appendix H. Configuration Layout

Input Layout:

Column Number	Column Name
1	record_id
2	First Name
3	Last Name
4	Date of Birth
5	Gender
6	SSN
7	ZIP
8	Email
9	Cellphone

Output Layout:

Column Number	Column Name	Token Elements/Note
1	record_id	N/A, not a token; passed through from input (no operation performed) *Note: this is the Pseudo_ID*
2	token_1	Last Name + First Initial + Gender +DOB
3	token_2	Last Name Soundex + First Name Soundex + Gender + DOB
4	token_3	Last Name + First Name + DOB + ZIP3
5	token_4	Last Name + First Name + Gender + DOB
6	token_5	SSN + Gender + DOB
7	token_16	SSN + First Name
8	token_29	First Name + Email
9	token_30	First Name + Cell Phone Number (US)
10	token_6	Last Name + First Name (first 3 characters) + Gender + DOB + ZIP3
11	token_7	Last Name + First Name (first 3 characters) + Gender + DOB
12	token_8	Last Name + First Name (first 3 characters) + Gender + ZIP5
13	token_12	Last Name + First Name + Gender + ZIP5
14	token_18	Last Name + First Name + Gender + ZIP5 + Birth Year + Birth Month
15	token_23	Last Name + First Initial + DOB + ZIP3
16	token_24	Last Name Soundex + First Name Soundex + DOB + ZIP3
17	token_26	Last Name + First Name + DOB + ZIP5
18	token_38	Last Name + First Name + DOB
19	token_39	SSN + DOB
20	token_encryption_key	N/A, not a token; a descriptor used for the site running DeID

Appendix I. Glossary

- **Data Enclave**: a data repository for storing Phenotype data for patients from multiple sites.
- **Datavant tokens:** non-reversible strings created from a patient's PII, allowing a patient's records to be matched across different de-identified health data sets without exposure of the original PII. Each site has a unique encryption key, so the same PII from two different sites will produce different tokens.
- **Datavant transit tokens:** tokens that have been encrypted using an encryption key that is shared between a data source and recipient. Using transit tokens provides additional encryption when sending data between parties and ensures that a site's tokens and their site-specific encryption key never leave their environment.
- **Linkage Honest Broker (LHB):** a system that will collect encrypted patient Identifiers from participating sites and provide a mechanism for securely "matching" patients across multiple sites and multiple cohorts.
- **Pseudo ID**: a unique ID for all patients from a source healthcare site. These identifiers are sent from sites to both the LHB and the Data Enclave.

Appendix J. Use Datavant Desktop

Introduction

In addition to a command line interface, the Datavant application is also available as a graphical user interface, Datavant desktop. Datavant recommends using the command line interface because of its advanced capabilities, including automation.

Step 1. Prepare environment to run Datavant desktop

Refer to the Datavant technical prerequisites, noting the following differences:

- Datavant desktop is only supported on Windows and Mac. You can download the desktop application from the Datavant portal.
- You must manually enter your credentials; they cannot be passed through automatically.

Step 2. De-identify and tokenize using Datavant desktop

After opening desktop and authenticating with your email, set up the inputs as follows:

- **I am…** tokenizing identified data and sending it to a partner
- **My Site Name**: select your site's name
- **Partner's Site Name**: n3c
- **Configuration Name**: n3c_tokens
- **My Data to Tokenize**: select the input file created in Step 1.1

Appendix K. Use Datavant v3

Introduction

You may use a v3 executable of the Datavant command line interface (DeID and Link) to create tokens and prepare them for sending to the Linkage Honest Broker. The input and output files created after each step are the same in both v3 and v4. The two major differences between versions are:

- v3 has two separate executables, DeID and Link, while v4 has a combined executable, Datavant CLI. Datavant CLI executes the same functions as DeID and Link.
- v3 used a site-specific authentication file and site software password to authenticate users when running the software. In v4, user-specific credentials are used instead.

Step 1. Ensure you are using a compatible v3 version

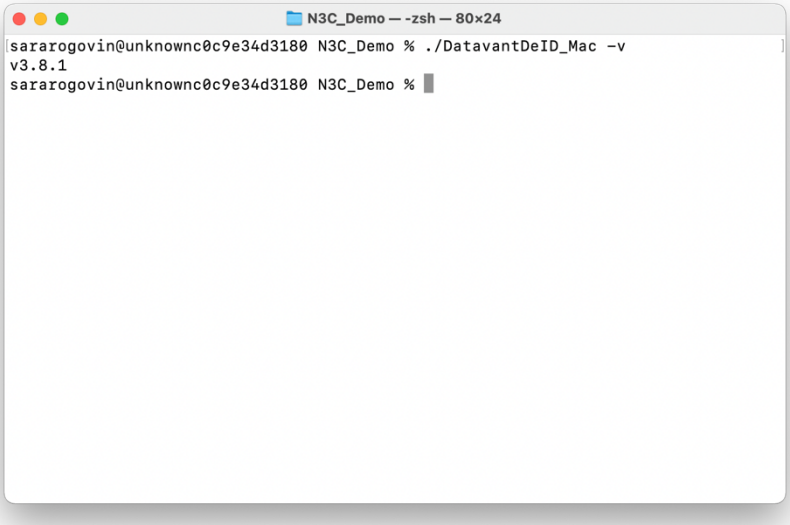
Important

To access the `n3c_tokens` software configuration and generate all the tokens in the configuration, **you must use a Datavant version at least 3.5.0 or higher**. If you are using a version prior to v3.5.0, refer to technical prerequisites to download the latest version of Datavant software. Note that this *may* require updates to your networking configuration so that the following endpoints are allowed outbound over port 443:

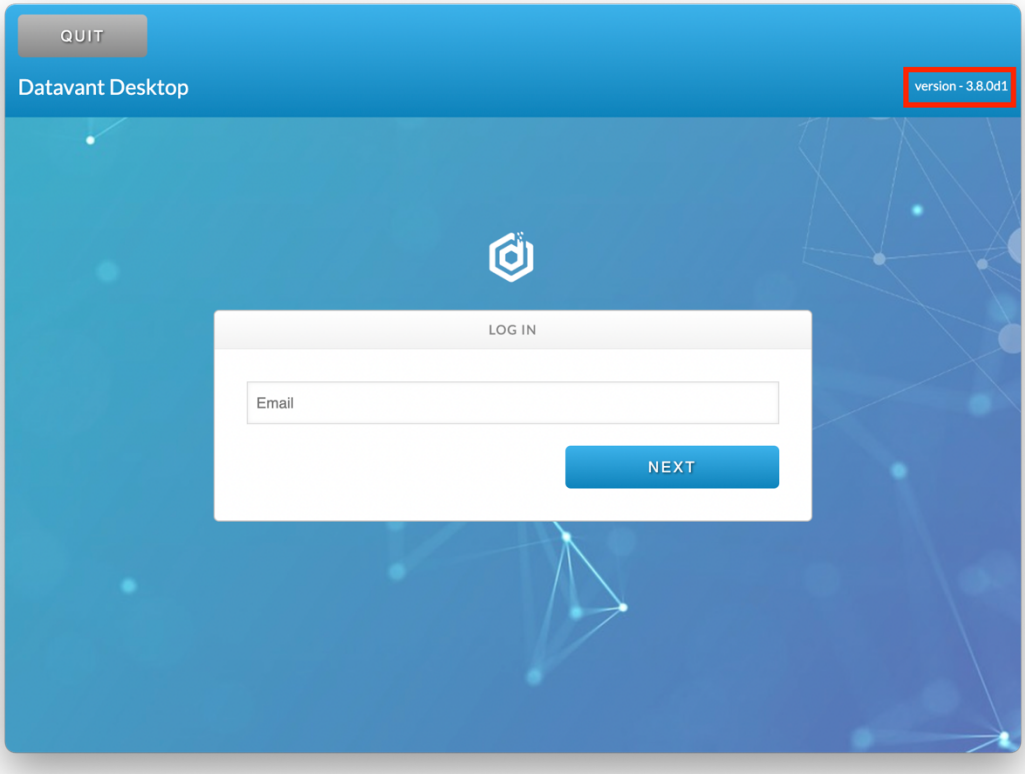
- sec.datavant.com
- auth.datavant.com
- api.datavant.com

If you are unable to make the networking configuration changes in a timely manner, reach out to n3csupport@datavant.com for a download link for v3.8 for your operating system.

If you are using the command line DeID or Link, you can determine the version using the `-v` command line argument.



If you are using Desktop, the version is listed in the top right corner of the application on startup.



Step 2. Prepare environment to run DeID and Link

Refer to the Datavant technical prerequisites, noting the following differences:

- Instead of retrieving your user-specific credentials, retrieve your authentication file and site software password. Both can be retrieved from the Datavant portal in My Settings > Datavant Application Credentials > v3 Authentication File and Password Management

Step 3. Run DeID to create site-specific tokens

Files Required

This step requires the following files:

- DeID executable
- Your site's authentication file (see Step 2 above for retrieval instructions)
- Your site's software password (see Step 2 above for retrieval instructions)
- Input file generated in Step 1.1

- Ensure the DeID executable and input file are both stored in the same directory. Then, navigate to that directory from the command prompt.
- Run DeID on the input file (see example commands below).

After this step, you will have a de-identified output and Datavant tokens in your site's specific encryption key.

Example commands for DeID

Windows

```
echo password | .\DatavantDeID_Win.exe -a yoursiteauthfile.yaml -s
yoursite -c n3c_tokens -i input.csv -o output_tokenize.csv -p
```

Mac

```
echo password | ./DatavantDeID_Mac -a yoursiteauthfile.yaml -s
yoursite -c n3c_tokens -i input.csv -o output_tokenize.csv -p
```

Linux

```
cat credentials | ./DatavantDeID_Linux -a yoursiteauthfile.yaml -s
yoursite -c n3c_tokens -i input.csv -o output_tokenize.csv -p
```

where

- `password` is your site's software password, retrieved from the Datavant portal
- `yoursiteauthfile.yaml` is your site's authentication file, retrieved from the Datavant portal
- `yoursite` is the name of your site, also available in the Datavant portal
- `input.csv` is the input file created in Step 1.1
- `output_tokenize.csv` is the name of the file where the output will be written

Step 4. Run Link to create transit tokens

Files Required

This step requires the following files:

- Link executable
- Your site's authentication file (see Step 1 above for retrieval instructions)
- Your site's software password (see Step 1 above for retrieval instructions)
- Output file generated in Step 3

- Ensure the Link executable and the output file generated in Step 2 are both stored in the same directory. Then, navigate to that directory from the command prompt.
- Run Link with `output_tokenize.csv` used as the input file (see example commands below).

Example commands for Datavant **`transform-tokens`**

Windows

```
echo password | .\DatavantLink_Win.exe -a yoursiteauthfile.yaml --
send-to n3c -s yoursite -i output_tokenize.csv -o
output_transformed.csv -p
```

Mac

```
echo password | ./DatavantLink_Mac -a yoursiteauthfile.yaml --send-to
n3c -s yoursite -i output_tokenize.csv -o output_transformed.csv -p
```

Linux

```
echo password | ./DatavantLink_Linux -a yoursiteauthfile.yaml --send-
to n3c -s yoursite -i output_tokenize.csv -o output_transformed.csv -p
```

where

- password are your user-specific credentials, retrieved from the Datavant portal
- yoursiteauthfile.yaml is your site's authentication file, retrieved from the Datavant portal
- yoursite is the name of your site, also available in the Datavant portal
- output_tokenize.csv is the output from Step 1.3
- output_transformed.csv is the name of the file where the output will be written

Appendix L. Send Viral Variant Data to N3C

Introduction

To augment EHR data within N3C for greater analytic context, your organization may submit viral variant summary data to the N3C Data Enclave, in addition to the regular EHR data you send today. There are three main steps in this process:

- **Viral variant summary data sent to N3C Data Enclave:** your organization or your sequencing lab will send viral variant summary data (e.g., "Delta variant") in the appropriate table(s) to the N3C Data Enclave, either in your existing payload, or as a separate, summary-only payload. You will send the summary data along with the appropriate N3C Pseudo ID.
- **Viral variant sequence data sent to NCBI Repository:** your sequencing lab will send viral variant sequence data to the NCBI Repository, along with your site's local ID (preferred ID is N3C Pseudo ID).
- **Crosswalk between N3C Pseudo ID and Accession ID sent to Linkage Honest Broker:** NCBI will send a crosswalk between a site's N3C Pseudo ID and the Accession ID (generated by NCBI) to the Linkage Honest Broker. This crosswalk will be used to link EHR data within the N3C Data Enclave to sequence data in the NCBI Repository.

Step 1. Evaluate your existing sequencing process to determine the workflow your site will use

Depending on your sequencing process, you will use a different process to implement viral variant data with N3C. Answer the following questions:

1. **For which patients are you sending viral variant data?**
    a. We are sending viral variant data for all tested patients, including ones that aren't a part of our N3C COVID cohort. *Choose this option if your sequencing lab performs sequencing for patients that aren't seen at your organization (e.g., for your entire state or region).*
    b. We are sending viral variant data only for patients who are in our N3C COVID cohort.
2. **Who will send viral variant summary data to the N3C Data Enclave?**
    a. Someone at our organization will send the summary data. *Choose this option if your sequencing lab is internal/in-house and you receive back the summary data from the sequencing lab.*
    b. The sequencing lab will send the summary data. *Choose this option if your organization doesn't receive back summary data from the sequencing lab.*
3. **How will you send viral variant summary data to the N3C Data Enclave?**
    a. We will always send it as its own, separate transaction, separate from the rest of our EHR data payload. *This should be the choice for all organizations whose external sequencing lab is sending summary data, or for organizations sending viral variant data for all tested patients.*
    b. We will always send it with the rest of our EHR data payload.
    c. We will send a backfill of summary data as its own separate transaction, and ongoing submissions will be sent with the rest of our EHR data payload.

Step 2. Generate tokens for tested patients and send to the LHB

If your organization is sending viral variant data only for patients in your N3C COVID cohort (Question 1, Answer B), you should have already generated tokens and sent them to the LHB as part of your N3C PPRL participation.

- Follow the steps in Part 1 of this packet to generate Datavant tokens for tested patients who are not in the N3C COVID cohort. Assign an N3C Pseudo ID to these patients as normal.
- Follow the steps in Part 2 of this packet to send the tokens to the Linkage Honest Broker.

Step 3. Send biospecimens to lab for sequencing

All organizations must complete this step.

Your sequencing lab should produce both a summary data output (to be sent to the N3C Data Enclave) and a complete sequencing data output (to be sent to the NCBI Repository).

- Create a crosswalk between the specimen ID for each sample and the patient's N3C Pseudo ID.
- Send the biospecimen to the lab for sequencing, along with the crosswalk between specimen ID and the N3C Pseudo ID.

Step 4. Send sequence data to the NCBI Repository

All organizations must complete this step.

After the specimen is sequenced, the sequencing lab should send the sequence data to the NCBI Repository along with the N3C Pseudo ID. The N3C Pseudo ID is your "local ID" for each sequence.

Step 5. Send viral variant summary data to the N3C Data Enclave

All organizations must complete this step. The party that sends the summary data will depend on your answer to Question 2 above. The format of the transaction will depend on your answer to Question 3 above.

- Send viral variant summary data to N3C, along with the patient's assigned N3C Pseudo ID, to the N3C Data Enclave.
- Sending party:
    - If your organization is sending the summary data to N3C *(Question 2, Answer A)*, use the crosswalk created in Step 2 to associate the specimen ID with the correct N3C Pseudo ID.
    - If your sequencing lab is sending the summary data to N3C *(Question 2, Answer B)*, the lab should use the crosswalk you sent in Step 2 along with the specimen to associate the summary data with the correct N3C Pseudo ID.
- Transaction format:
    - Send the summary data as its own separate transaction. You may choose this option if you are submitting a backfill of data *(Question 3, Answer C)* or if you are submitting data for patients who are not in the N3C COVID cohort *(Question 1, Answer A)*.
    - Send the summary data as part of your regular N3C EHR payload. You may choose this option for ongoing data submissions *(Question 3, Answer C)*, or if your tested patients will always be in the N3C COVID cohort *(Question 1, Answer B)*.