Data Provenance and Software Genology of Genomic Surveillance Infrastructure

Nathanael Sheehan & Sabina Leonelli

2022-11-30

Abstract

Introduction

Data provenance (often referenced as data-lineage) is defined as the: "record trail that accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place" (Gupta 2009). Data provenance is thereby concerned with tracking the inputs and outputs of computational processes (programs or scripts) that manipulate data across space and time. In academia, data provenance has been formally studied throughout the database community since the early 2000s and at present marks itself an important component to both Open Science and Open Government research and technological policy formulation (Leonelli 2019,Leonelli 2020). The TRUST Principles for digital repositories directly associates data provenance with data responsibility, arguing that it "greatly enhances the discoverability and usefulness of the data" (Lin et al. 2020). In the context of SARS-CoV-2 genome sharing, data provenance ensures that the data adheres to not only quality standards but also ethical and legal obligations. This includes ensuring that the genomes are correctly attributed, linked to metadata records (i.e., a description of how and when the genome was sequenced), and can be tracked through time for further analysis should new findings emerge from future studies.

While there is presently no international standard for data provenance; however, there are various perspectives on its definition as well as numerous examples of proposed or implemented systems that track and/or record provenance information throughout a computational process (Yogesh et al. 2005). Following these perspectives, we argue that data provenance in SARS-CoV-2 genomics encompasses three core components:

- (1) accurate attribution denoting who produced the dataset;
- (2) associated metadata describing when, where, under what circumstances it was collected;
- (3) an audit trail detailing all transformations applied to the dataset during processing from sampling until sharing.

In the rest of this paper we will explore how these three elements have been treated within existing practices of genomic sequence sharing in order to discuss both their benefits as well as limitations with respect to comprehensive treatment of SARS-CoV-2 genomes being shared within open repositories such as GISAID (https://www.gisaid.org/) and the Covid-19 Data Portal (https://www.covid19dataportal.org/).

Methods

Two types of data are necessary in order to map the high level data provenance pipeline for each data portal: (1) software and security information regarding the portals infrastructure and (2) records of complete

metadata for a given sequence. In order to obtain the relevant software and security information regarding each data portal, a penetration test (pen test) was conducted using the open source `whatWeb` command line tool. A penetration test is commonly used to find and exploit vulnerabilities in a computational system. While the aim of this is not to find the vulnerabilities of the data portals, the results from pen test map out many of the infrastructures being used to host the data.

For each data portal the following command was used:

```
./whatweb -v https://www.dataportal-url.org -a 1
```

Where `-v` gives a verbose output of the results and `-a 1` represents a soft level of pen test. The results for each data portal can be found in the following code chunks.

The Covid-19 Data Portal

```
WhatWeb report for https://www.covid19dataportal.org/
Status    : 200 OK
Title     : COVID-19 Data Portal - accelerating scientific research through data
IP        : 193.62.193.83
Country   : UNITED KINGDOM, GB

Summary   : HTML5, UncommonHeaders[x-content-type-options,x-download-options,x-permitted-cross-domain-p

Detected Plugins:
[ Apache ]
    The Apache HTTP Server Project is an effort to develop and
    maintain an open-source HTTP server for modern operating
    systems including UNIX and Windows NT. The goal of this
    project is to provide a secure, efficient and extensible
    server that provides HTTP services in sync with the current
    HTTP standards.

    Google Dorks: (3)
    Website     : http://httpd.apache.org/

[ Google-Analytics ]
    This plugin identifies the Google Analytics account.

    Version     : Universal
    Account     : UA-163982534-1
    Website     : http://www.google.com/analytics/

[ HTML5 ]
    HTML version 5, detected by the doctype declaration


[ HTTPServer ]
    HTTP server header string. This plugin also attempts to
    identify the operating system from the server header.

    String      : Apache (from server string)

[ Script ]
    This plugin detects instances of script HTML elements and
```

```
    returns the script language/type.

    String       : module

[ Strict-Transport-Security ]
    Strict-Transport-Security is an HTTP header that restricts
    a web browser from accessing a website without the security
    of the HTTPS protocol.

    String       : max-age=63072000; includeSubDomains; preload

[ UncommonHeaders ]
    Uncommon HTTP server headers. The blacklist includes all
    the standard headers and many non standard but common ones.
    Interesting but fairly common headers should have their own
    plugins, eg. x-powered-by, server and x-aspnet-version.
    Info about headers can be found at www.http-stats.com

    String       : x-content-type-options,x-download-options,x-permitted-cross-domain-policies,referrer

[ X-Frame-Options ]
    This plugin retrieves the X-Frame-Options value from the
    HTTP header. - More Info:
    http://msdn.microsoft.com/en-us/library/cc288472%28VS.85%29.
    aspx

    String       : SAMEORIGIN

[ X-XSS-Protection ]
    This plugin retrieves the X-XSS-Protection value from the
    HTTP header. - More Info:
    http://msdn.microsoft.com/en-us/library/cc288472%28VS.85%29.
    aspx

    String       : 1; mode=block

HTTP Headers:
    HTTP/1.1 200 OK
    Date: Mon, 31 Oct 2022 17:34:27 GMT
    Server: Apache
    X-Content-Type-Options: nosniff
    X-XSS-Protection: 1; mode=block
    X-Frame-Options: SAMEORIGIN
    X-Download-Options: noopen
    X-Permitted-Cross-Domain-Policies: none
    Strict-Transport-Security: max-age=63072000; includeSubDomains; preload
    Referrer-Policy: same-origin
    Access-Control-Allow-Origin: *
    Last-Modified: Mon, 31 Oct 2022 08:09:31 GMT
    ETag: "263a-5ec502092487f-gzip"
    Accept-Ranges: bytes
    Vary: Accept-Encoding
    Content-Encoding: gzip
    Content-Length: 2714
```

```
    Connection: close
    Content-Type: text/html; charset=UTF-8
```

GISAID

```
WhatWeb report for https://epicov.org/
Status    : 301 Moved Permanently
Title     : 301 Moved Permanently
IP        : 104.22.50.160
Country   : UNITED STATES, US

Summary   : UncommonHeaders[cf-cache-status,cf-ray], HTTPServer[cloudflare], RedirectLocation[https://w

Detected Plugins:
[ HTTPServer ]
    HTTP server header string. This plugin also attempts to
    identify the operating system from the server header.

    String       : cloudflare (from server string)

[ RedirectLocation ]
    HTTP Server string location. used with http-status 301 and
    302

    String       : https://www.gisaid.org/ (from location)

[ UncommonHeaders ]
    Uncommon HTTP server headers. The blacklist includes all
    the standard headers and many non standard but common ones.
    Interesting but fairly common headers should have their own
    plugins, eg. x-powered-by, server and x-aspnet-version.
    Info about headers can be found at www.http-stats.com

    String       : cf-cache-status,cf-ray (from headers)

HTTP Headers:
    HTTP/1.1 301 Moved Permanently
    Date: Mon, 31 Oct 2022 17:39:48 GMT
    Content-Type: text/html; charset=iso-8859-1
    Transfer-Encoding: chunked
    Connection: close
    Location: https://www.gisaid.org/
    CF-Cache-Status: DYNAMIC
    Server: cloudflare
    CF-RAY: 762e2c16af9e188f-MAN

WhatWeb report for https://www.gisaid.org/
Status    : 403 Forbidden
Title     : 403 Forbidden
IP        : 78.46.3.243
Country   : GERMANY, DE

Summary   : Apache, HTTPServer[Apache]
```

```
Detected Plugins:
[ Apache ]
    The Apache HTTP Server Project is an effort to develop and
    maintain an open-source HTTP server for modern operating
    systems including UNIX and Windows NT. The goal of this
    project is to provide a secure, efficient and extensible
    server that provides HTTP services in sync with the current
    HTTP standards.

    Google Dorks: (3)
    Website     : http://httpd.apache.org/

[ HTTPServer ]
    HTTP server header string. This plugin also attempts to
    identify the operating system from the server header.

    String       : Apache (from server string)

HTTP Headers:
    HTTP/1.1 403 Forbidden
    Date: Mon, 31 Oct 2022 17:39:50 GMT
    Server: Apache
    Content-Length: 264
    Connection: close
    Content-Type: text/html; charset=iso-8859-1
```

In order to access data and metadata records, a manual collection of data and metadata was carried out for each platform on October 30th 2022. To standardise this process, each manual collection was led by the accession of the same sequence for each portal, the number of "hops" were then recorded to show the infrastructural lineage of the data, as well as documenting supplementary metadata associated with a given sequence. Data accessed through GISAID was conducted following their terms of use policy.

Results and Discussion

Conclusion

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101001145). This paper reflects only the author's view and that the Commission / Agency is not responsible for any use that may be made of the information it contains.

Biblography

Gupta, A. (2009). Data Provenance. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_1305

Leonelli, S. (2019). Data Governance is Key to Interpretation: Reconceptualizing Data in Data Science. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.17405bb6

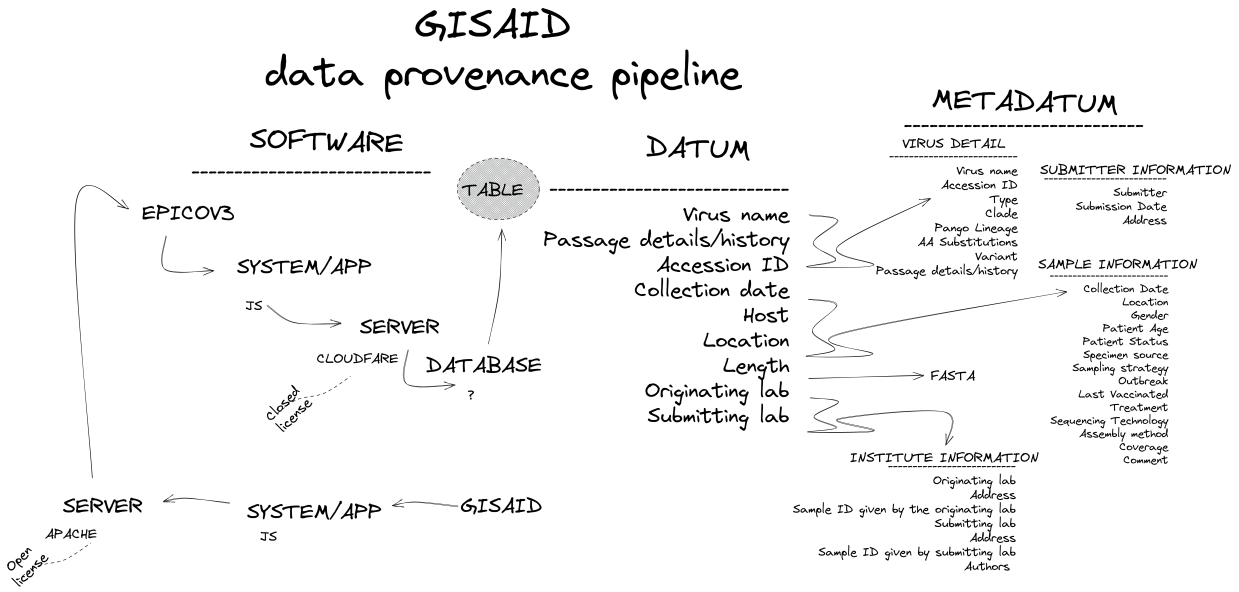

Figure 1: GISAID Data Provenance Pipeline: A data user will most likely start their journey using the GISAID platform, hosted on highly secure and free/OS software in Germany. From here a data user will need to sign up for the EpiCov3 database which is the primary way to access GISAID data. The EpiCov3 database is hosted on highly secure, closed software in the United States. Aggregated data is summarised in nine columns which reference the three sub metadata categories. One important thing to notice here is the direct mapping between the data and metadata. In most cases, a column in the aggregated data will directly reference and match to a metadata sub category that will contain additional and often more private information. It should be noted that there also exists one isolated metadata category, which contains highly personal information regarding the submitter of the sequence

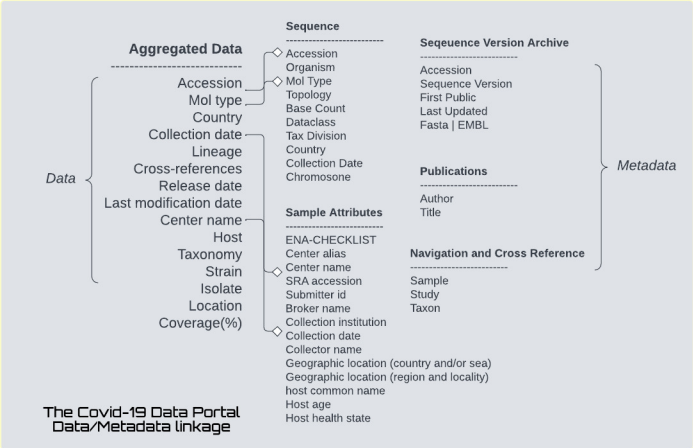

Figure 2: The Covid-19 Data Portal Data Provenance Pipeline: A data user will most likely start their journey using the Covid-19 Data portal, hosted on a fairly secure LAMP stack in the United Kingdom. From here a data user is able to use the data portal to openly query sequence data. However, when they do this request does not come from the Covid-19 Data Portals infrastructure itself, rather it queries the European Nucleotide Archive database and returns back aggregated data. In order to explore the metadata associated with a sequence, the data user will then be redirected to the ENA's app. An important thing to note here, is the dissonance between the aggregated data names and the associated metadata. While some common labels stay the same e.g. accession ID, many take new names after being pooled to the Covid-19 Data portal e.g. center name (data) -> collection institution (metadata). This linkage strategy blurs the direct matches and references between the data and the five metadata sub categories. This being said, two of the metadata categories hosted by the ENA contain linked if a sequence has been linked to a particular study, sample and taxon.

Leonelli, Sabina (2020). "Scientific research and big data." https://plato.stanford.edu/entries/science-big-data/?ref=hackernoon.com

Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. 2005. A survey of data provenance in e-science. SIGMOD Rec. 34, 3 (September 2005), 31–36. https://doi.org/10.1145/1084805.1084812