CZ ID (formerly IDseq) Security White Paper

Last Updated: February 28, 2022

This white paper is intended for technical readers, such as Data Protection Officers, Chief Information Officers, Chief Technology Officers, Laboratory Directors, Technical Supervisors, Institutional Research Directors, and Principal Investigators. It is intended to provide additional clarity and specifics about how we act on our commitments made in our Privacy Notice, Terms of Use, and FAQs. Protection and proper handling of human genomic data is a top priority for us. We understand that the files that Users upload to the platform may contain data that is sensitive, and we implement security measures designed to safeguard it. We seek to implement security best practices like encrypting data, hosting it on leading cloud providers with robust physical security, regular security assessments, data loss prevention systems, and working to ensure that only authorized staff have access to the data. The security of this sensitive data is a shared responsibility between Users, CZ ID, and the providers of infrastructure that we use (our Providers). This white paper outlines how these responsibilities align and complement each other to provide these safeguards.

Should you have security or privacy questions, please reach out to our team at security@czid.org or privacy@czid.org respectively.

1. CZ ID Overview & Security Principles

About CZ ID

CZ ID (formerly IDseq) is a hypothesis-free, open source, global software platform that helps scientists identify pathogens in metagenomic sequencing data. CZ ID accepts sequencing data from labs and quickly processes the results to provide actionable information on the state of pathogens in the given set of samples. This allows researchers to make data-driven decisions such as when to deploy antibiotics, where to prioritize immunization campaigns, and how to shape vector-borne disease surveillance and control efforts.

CZ ID Data Principles

When using CZ ID, Users submit Upload Data. This data may contain human and non-human genomic data (Raw Sample Data), as well as information about those sequences (Sample Metadata), such as the date the sample was collected and the species it was collected from. Upon upload, Raw Sample Data is processed through our data pipeline that both filters out human genomic data, and prepares the data for CZ ID reports and visualizations. Details about this pipeline are discussed in a paper published in GigaScience[1]. The Raw Sample Data remains accessible to the user who uploaded it, however it is not accessible to other users or to anyone working on CZ ID unless specifically requested by a User, such as to debug an issue. Those working on CZ ID are granted access only on an as-needed basis, and this access is logged and individually attributable. Users providing Raw Sample Data are responsible for ensuring that they only upload data that they are authorized to process via a tool like CZ ID, and that no Personally Identifiable Information (PII) Is contained within the Sample Metadata of the files they upload. When feasible, Users are advised to limit the data uploaded to that necessary for their analysis.

[1] K Katrina L., et al.: IDseq—an open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. GigaScience 9.10 (2020): giaa111, https://doi.org/10.1093/gigascience/giaa111

Infrastructure Partnership Overview

CZ ID is provided by a partnership between the Chan Zuckerberg Biohub Inc (CZ Biohub), an independent nonprofit research center, and the Chan Zuckerberg Initiative Foundation (CZIF), a 501(c)(3) nonprofit private foundation. This partnership is responsible for ensuring the security of the CZ ID software and platform. Privacy and security are top of mind for this partnership; data are only used for the purposes set out in our Privacy Notice, and will not be sold.

Where third party service providers are used, CZI is responsible for evaluating, monitoring, and maintaining the policies, controls, and agreements with regard to these third party service providers in order to keep the platform and its contents secure. For example, the third party service providers are contractually obligated to secure data from unauthorized access and use, and to limit their use of data only to the extent permitted in our agreements.

Our information security program implements and maintains controls that align to the Center for Internet Security Critical Security Controls. We regularly evaluate our policies and practices to improve security and to keep up with the latest practices of the security industry.

You can learn more about our relationship with CZI and other service providers in our Privacy Notice, Terms of Use, and FAQs.

2. Infrastructure Security

Encryption at Rest and In Transit

Access to the CZ ID service occurs via encrypted connections (HTTP over TLS, also known as HTTPS) which encrypt all data before it leaves CZ ID's servers and protect that data as it transits over the internet. We use HTTP Strict Transport Security to ensure that pages are loaded over HTTPS connections and our TLS configuration receives an A+ from Qualys SSL Labs.

All personally identifiable User data, such as account information and site analytics, is encrypted at rest using modern encryption algorithms such as AES-256 or stronger, and is retained no longer than necessary to provide the services for which it was collected. Raw Sample Data files uploaded by Users may include human and non-human genomic data and are similarly encrypted.

Network Security

The CZ ID platform is provided as a cloud-based application. CZ ID uses Amazon Web Services (AWS), a leading cloud provider, to host the infrastructure, and we have evaluated its security controls, processes, and practices to ensure alignment with our policies. AWS undergoes strict ongoing security assessments from external audit firms to ensure compliance with security standards including ISO 27001, SOC 2, PCI DSS Level 1, and FISMA. See AWS' Compliance Programs for more details.

Network access to the CZ ID infrastructure is highly restricted. AWS hosted infrastructure resides in a dedicated Virtual Private Cloud (VPC) which is designed to ensure that only authorized traffic over approved ports is allowed. Development infrastructure resides in a separate VPC from that used for the production environment that Users access. Both production and development environments use the same

security controls, though differ in access control rules based on how these environments are used and the data that they contain.

We leverage built-in AWS services, such as AWS GuardDuty, to monitor for suspicious activity. GuardDuty is a continuous monitoring service that analyzes and processes event, flow, and DNS logs to identify unexpected and potentially unauthorized or malicious activity.

Access Management

Access to the CZ ID infrastructure is highly restricted, particularly with respect to Raw Sample Data (that could include human genomic data) and Sample Metadata both of which are hosted by CZ ID. Account data related to Users is restricted by the same controls. We limit network access to the AWS-hosted Virtual Private Cloud (VPC) to individuals who need access to do their jobs such as engineers, data scientists, product managers, and support personnel. We use a codified infrastructure tool to minimize human error in configurations and provide consistent testing, staging, and production environments. These configurations include management of controlled access through automated onboarding and offboarding based on the projects, roles, and positions these individuals have. Individuals who no longer work on CZ ID or for the organization are automatically removed from access lists. All access to the VPC and our infrastructure is logged. All access to our infrastructure requires the use of strong passwords and multifactor authentication.

In addition, CZ ID's cloud infrastructure host, AWS, provides a comprehensive set of technical, operational, and contractual measures to ensure strong data protection that we have confirmed to be consistent with our policies. See AWS Data Protection descriptions for more details.

Patching

Infrastructure patching is managed by the providers providing each component of the infrastructure platform. CZ ID reviews provider processes to ensure that they meet the stringent security protocols required of the application.

Our third party providers are responsible for providing tools, information, and often automation to ensure that security of these infrastructure components are up-to-date. Specifically,

- **AWS Managed Services (e.g. Relational Database Service)**: AWS proactively notifies our engineering team when updates are available and these updates are automatically applied.
- **AWS EC2**: All EC2 instances are configured to automatically apply operating system and kernel patches. This includes automatic restarts as needed.
- **Auth0 (User Access Management)**: Auth0 regularly monitors their services for vulnerabilities and software bugs and releases new versions of their software with updates. We monitor and apply these updates.

Backups

We have an automated data backup and recovery capability that is designed to provide a timely restoration of CZ ID, with minimal data loss, in the case of catastrophic failure. These backups are encrypted.

3. Physical Security

CZ ID uses several methods to ensure the physical security of the equipment that houses our infrastructure and application cybersecurity. The responsibility of physical security is shared across CZ ID, our Partners, and the Users of the platform.

When choosing Partners, CZ ID evaluates and monitors their security policies, processes, and controls to ensure that they meet or exceed the security of our own. CZ ID is currently hosted in Amazon Web Services (AWS), which employs industry-leading physical security measures to protect their data centers such as a full 24/7 onsite security team, video surveillance, and perimeter intrusion detection systems. These security features are regularly audited by third-party auditors. See the documentation about AWS' physical security controls.

The servers, networking equipment, and storage used to run CZ ID are housed in secure AWS buildings. In addition, laptops and other equipment that may be used remotely by CZ ID team members are encrypted and have password policies that automatically lock after too many missed attempts. Those working on CZ ID are required to use multifactor authentication for core tools and are trained on security protocols.

Similarly, Users of CZ ID are advised to be vigilant with the security of the computers that they use to access CZ ID, particularly if they contain stored passwords for accessing Raw Sample Data that they have uploaded to their CZ ID account.

4. Application Security

Secure Software Development Lifecycle

In addition to designing our systems with privacy and security in mind, we employ a combination of manual and automated processes to identify potential vulnerabilities. This includes mandatory code review, automated source code scanning, software quality regression testing, automated dependency scanning, as well as comprehensive black box penetration testing of CZ ID by external security experts. Access to software code repositories is governed by enterprise single sign on and automated access control.

CZ ID also monitors and applies patching necessary to secure the software. These patches may include proactive updates to correct vulnerabilities in software libraries, developed software, or interfaces to the platform infrastructure as described above.

In addition to our proactive methods, we run a Vulnerability Disclosure Program through our partnership with Bugcrowd, which allows security researchers who identify vulnerabilities to responsibly disclose them to us. If you suspect or know of a security vulnerability in the CZ ID product, please contact us at security@czid.org.

Browser Security

We use an up-to-date Content Security Policy (CSP) to prevent unauthorized JavaScript from running in the context of CZ ID and we use standard countermeasures to protect against Cross-Site Request Forgery (CSRF). Users are responsible for ensuring that the browsers they use to access CZ ID have been updated to

the most recent version. These actions ensure that a User's interactions and data are not subject to known browser vulnerabilities.

Authentication and Access Control

Access to CZ ID is granted for the sole purpose of scientific research. When users create accounts for CZ ID, they are asked to provide certain information ("Account Information") to create an individual user account (an "Account"). As disclosed in our Terms of Use, it is the responsibility of users to ensure that their Account Information is kept up-to-date. CZ ID users may not share their accounts on CZ ID, and the Terms of Use prohibit the disclosure of the Database, or any works serviced from the Database, to any third parties.

CZ ID uses Auth0 for all application authentication and access. Auth0, is a leading cloud-based access management provider, to secure application access to the CZ ID platform. CZ ID has evaluated Auth0's security controls, processes, and practices to ensure alignment with our policies. It complies with key security and privacy frameworks including ISO27001, SOC 2 Type II, ISO27018 and HIPAA BAA, among others, and complies with the General Data Protection Regulation (GDPR). Auth0 supports data sovereignty, allowing those using their services to control where related data is stored. CZ ID stores access data in its Virtual Private Cloud (VPC) for a limited amount of time to operate and secure the platform. For more information about Auth0's data privacy and compliance controls, please review the related Auth0 article.

For user access, Auth0 provides programmatic access control policies that are tested for compliance and monitored for misuse. These controls mean that your data is only accessible to Users who have the proper permissions. The controls include advanced detection of unauthorized access, using tools such as bot detection, suspicious iP throttling, brute-force protection, and breached password detection. More information about these monitoring tools can be found on the Auth0 website.

Our policies set out the types of data which can be uploaded to CZ ID, and our expectations that Users do not re-identify or attempt to re-identify any Database records or content for any purpose other than responding to a request from the individual whom the record within the Upload Data is about.

From time to time, CZ ID team members may need privileged application access to stored data to fix issues or maintain the software. Access is allowed only on an as-needed basis, and is logged and individually attributable.

5. Security Governance & Policies

Incident Response

CZ ID shares the responsibility of Incident Response with our third-party providers. We have configured our Provider tools to ensure effective and responsive detection and mitigation of potential data and application security challenges, and established processes for detecting suspicious or abnormal activity on CZ ID that might have a security implication. Using this shared model, our key infrastructure provider, AWS, has the responsibility of managing the security *of* the cloud (the infrastructure), while CZ ID manages the security *in* the cloud (the application and its functions).

In order to safeguard the personal data which we hold, we have an established process that is followed whenever we detect suspicious or abnormal activity on CZ ID that might have a security implication. In order to support this process and our efforts to ensure CZ ID is available, our engineering and security teams have on-call rotations to provide a designated point person available to respond to any suspicious or abnormal activity.

As part of our incident response process, we perform post-mortem reviews of major incidents including both security and non-security related (such as site outages). These post-mortem reviews are designed to ensure that we learn from past incidents and, if needed, improve CZ ID to prevent them from occurring again in the future.

GDPR Considerations

In the case of an incident that may constitute a personal data breach under the GDPR, CZI, CZIF and/or CZ Biohub will act promptly to investigate and determine whether this breach is a personal data breach required to be reported to a competent supervisory authority in the European Union, and whether the impacted individual users should be notified as well.

In addition to the breach notification obligations under the GDPR, CZI, CZIF and CZ Biohub will assess any other breach notification obligations which may arise pursuant to other legal frameworks (ex. California state law).

Data Retention

Data pertaining to Users is deleted within 60 days of account deletion as per our Privacy Notice.