

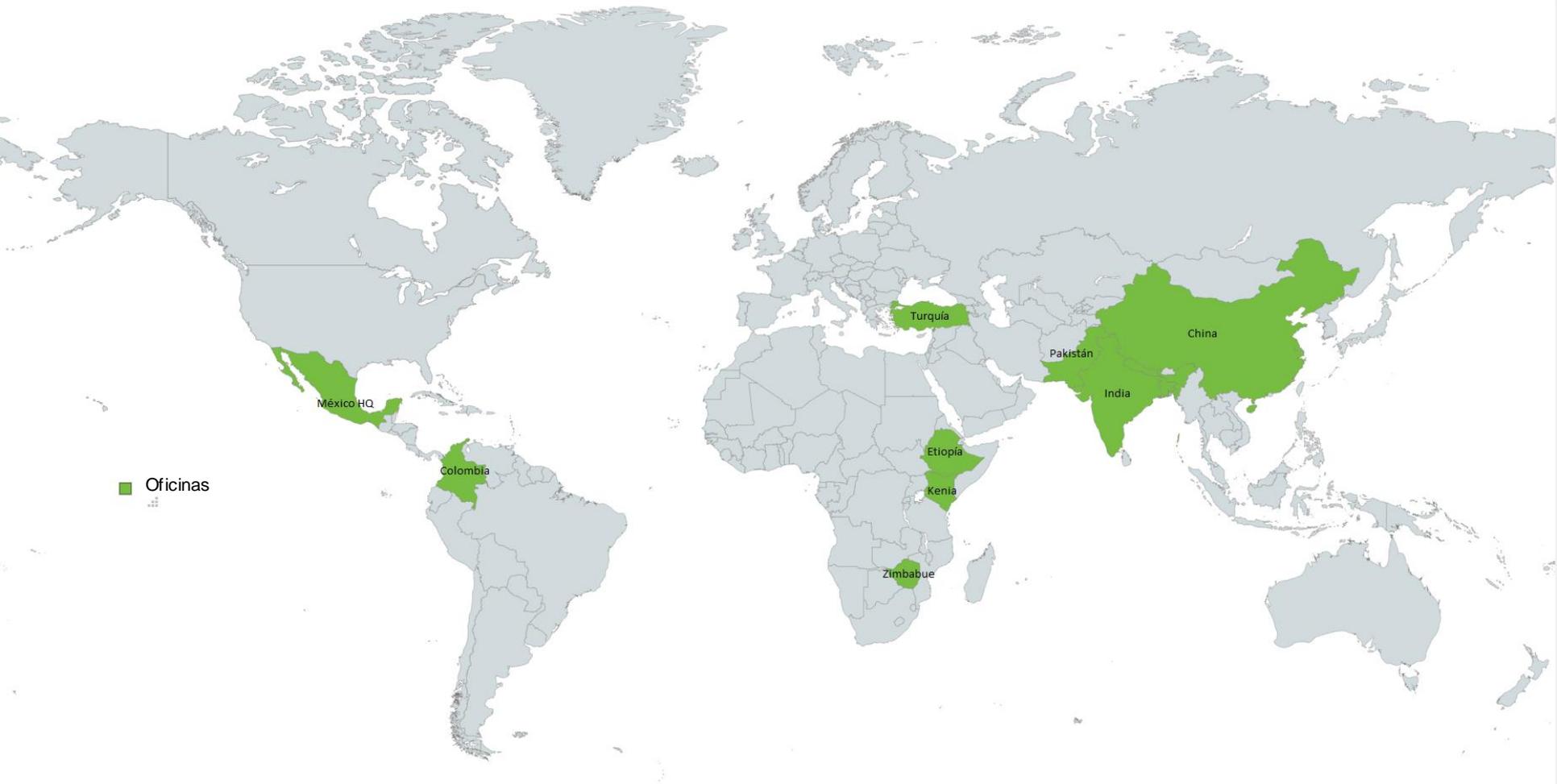
# CIMMYT

## Proyecto Dataverse

Alejandra Tenorio y Jesús Herrera

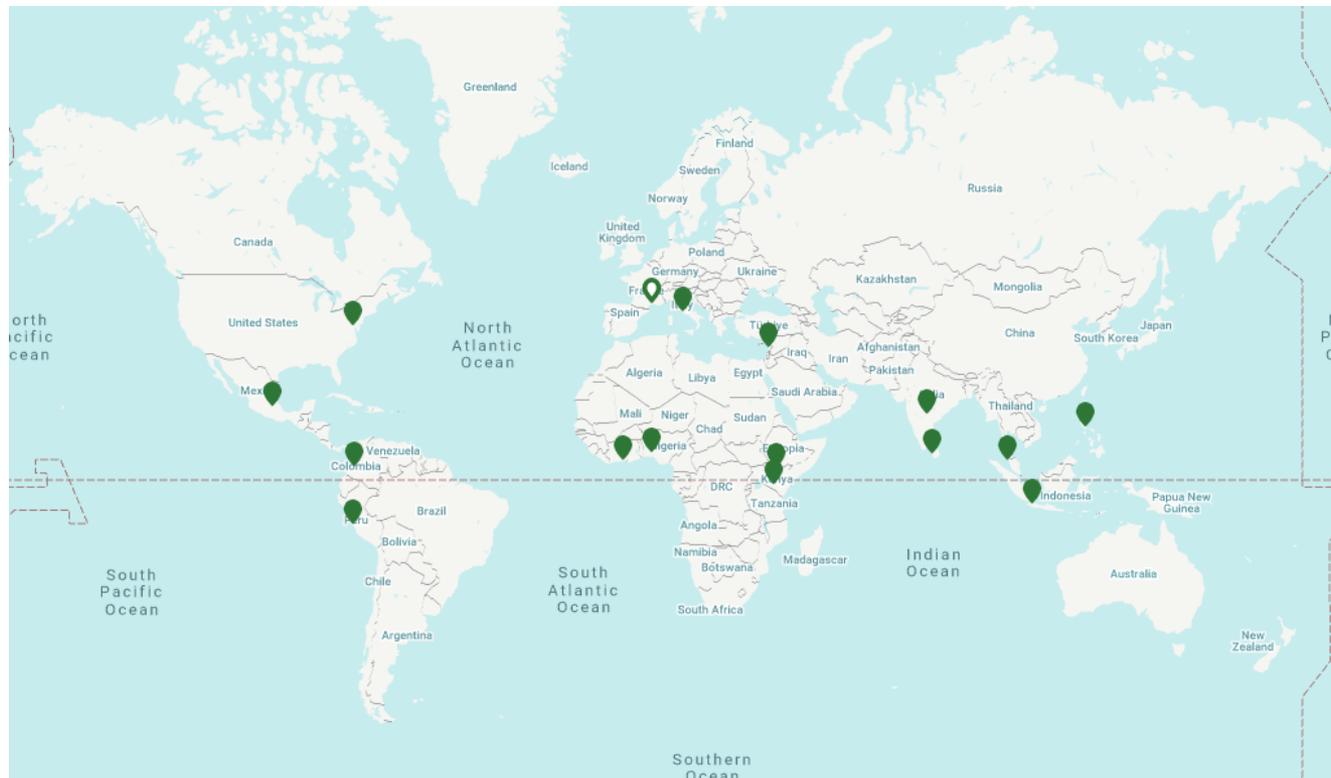
17 de abril de 2023, Buenos Aires, Argentina.

# CIMMYT en el mundo



# Centros de investigación del CGIAR

CIMMYT es uno de los miembros del CGIAR en 65 países



# CIMMYT Research Data & Software Repository Network

- Construir una plataforma de datos abiertos



- Inició en 2014

- Versión 3.0

- 3 subrepositorios

## CIMMYT Dataverse Network

CIMMYT institutional network of scientific datasets and software repositories.

POWERED BY THE **Dataverse Network** PROJECT V. 3.0

[Search](#) [Info](#) [Help](#) [Create Account](#) [Log In](#)

### Released Dataverses

[Advanced Search Tips](#)

[ALL](#) [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

Dataverses: 3 | Studies: 94 | Files: 297

Name	Affiliation	Released	Activity
CIMMYT Research Software <a href="#">View Info [+]</a>	CIMMYT	3/06/2015	■■■■■
CIMMYT Seeds of Discovery <a href="#">View Info [+]</a>	CIMMYT	19/12/2014	■■■■■
CIMMYT Research Data <a href="#">View Info [+]</a>	CIMMYT	22/10/2014	■■■■■

Copyright © 2014 International Maize and Wheat Improvement Center, CIMMYT



# CIMMYT Research Data & Software Repository Network



Metrics 178,419 Downloads

Contact Share

CIMMYT institutional network of scientific datasets and software repositories.



Search this dataverse...   Advanced Search

- Datasets (937)
- Files (9,354)

**Dataverse Category**  
Organization or Institution (1)  
Researcher (1)

**Publication Year**  
2023 (11)  
2022 (168)  
2021 (63)  
2020 (145)  
2019 (133)

1 to 10 of 945 Results

Sort ▾

The reference genomes of 7 diploid and 1 triploid banana variety

Mar 17, 2023 - Crops to End Hunger (CEH) Data



Shah, Trushar; Uwimana, Brigitte; Amah, Delphine; Brown, Allan; Swennen, Rony. 2023. "The reference genomes of 7 diploid and 1 triploid banana variety". <https://hdl.handle.net/11529/10548875>. CIMMYT Research Data & Software Repository Network. V1

3 wild banana diploid varieties (ITC0249 Calcutta 4, ITC0253 Borneo, ITC0766 Palilama) 1 edible diploid banana variety (ITC0809 Maleb) 3 improved diploid banana parents (SH 3142, SH 3217, TMB2xg128-3) 1 plantain variety

The reference genome of cowpea (*Vigna unguiculata* L. Walp.) variety IT99K-573-1-1

- Versión: 5.10.1



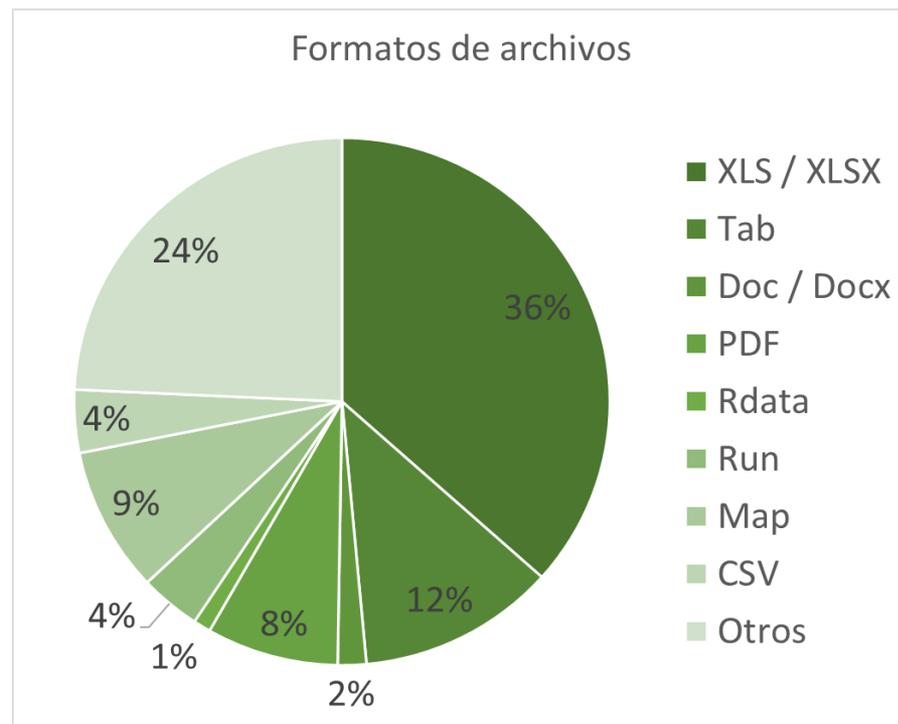
# CIMMYT Research Data & Software Repository Network

## Tipos de datos principales:

- Datos procesados
- Datos limpios
- Subconjuntos
- Encuestas
- Software

## Almacenamiento:

- Sistema de archivos local



# Nuevos desafíos – Big Data

- Cada dataset podría llegar a ~2 TB
  - Datos de las secuencias
  - Datos crudos o borrosos
- Herramientas tradicionales como,
  - Dropbox
  - globus
  - SFTP
  - Discos Duros



# Big Data

Integrar Dataverse con infraestructura tecnológica con las siguientes características:



- Escalable (almacenamiento)
- Alta disponibilidad
- Duradera
- Seguridad



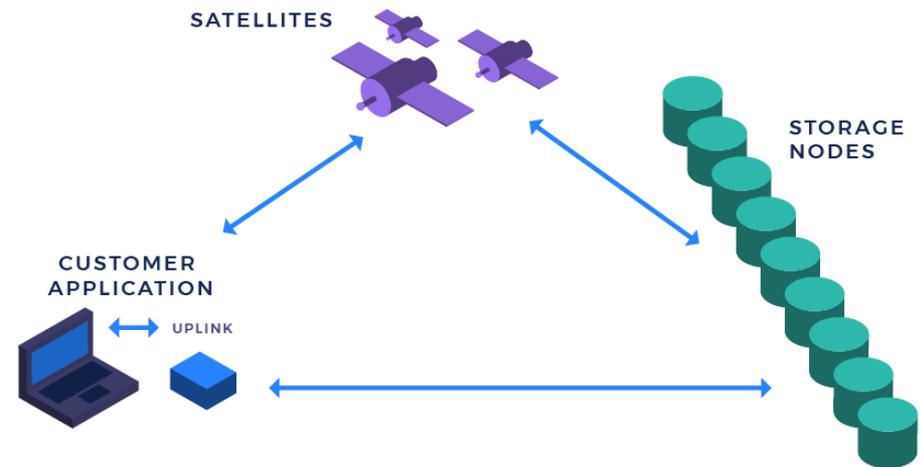
# StorJ

- Plataforma de almacenamiento en la nube descentralizada
- Alta disponibilidad: codificación Reed Solomon
- Fragmentación de ficheros en 80 piezas
- Alta redundancia
- Alta durabilidad de los datos
- Compatible con el servicio de almacenamiento de objetos S3
- Ahorrar al menos 80% - 90% que con AWS S3



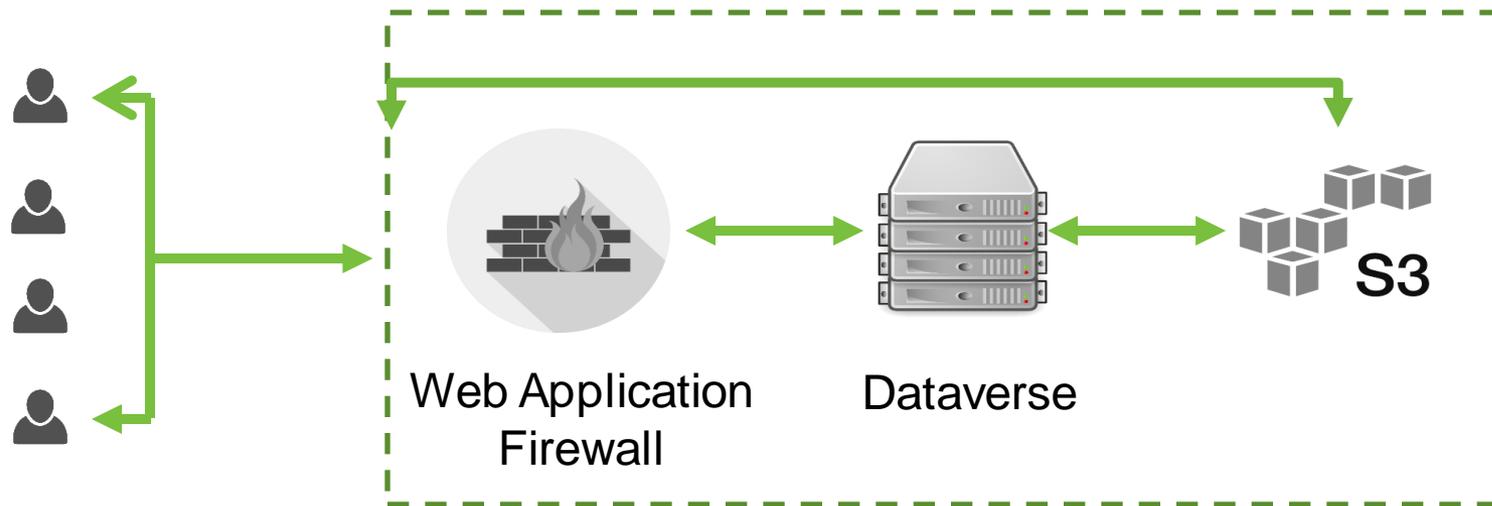
# Integración de Dataverse con un almacenamiento S3 o compatible

- Al ser compatible con S3 se activó como almacén remoto para el repositorio de CIMMYT
- Se migró todo el almacenamiento del sistema de archivos local a StorJ



# Integración de Dataverse con StorJ

- Repositorio tiene la siguiente arquitectura



# Herramientas para cargar archivos

- Desde la interfaz de usuario en el repositorio
- Directa desde DropBox
- DVUploader

Ficheros

Si necesita más información sobre formatos de ficheros soportados, puede dirigirse a [Guía de usuario](#).

Subir desde HTTP usando su navegador 

Seleccione los ficheros o arrástrelos al widget de subida. Maximum of 1.000 files per upload. El tamaño máximo de cada fichero son 200.0 GB bytes. **Tabular file ingest** is limited to 200 B. Ingest is limited to the following file sizes based on their format: Rdata: 200 B.

 **Seleccione los ficheros que quiera añadir**

Arrastre y suelte aquí los ficheros.

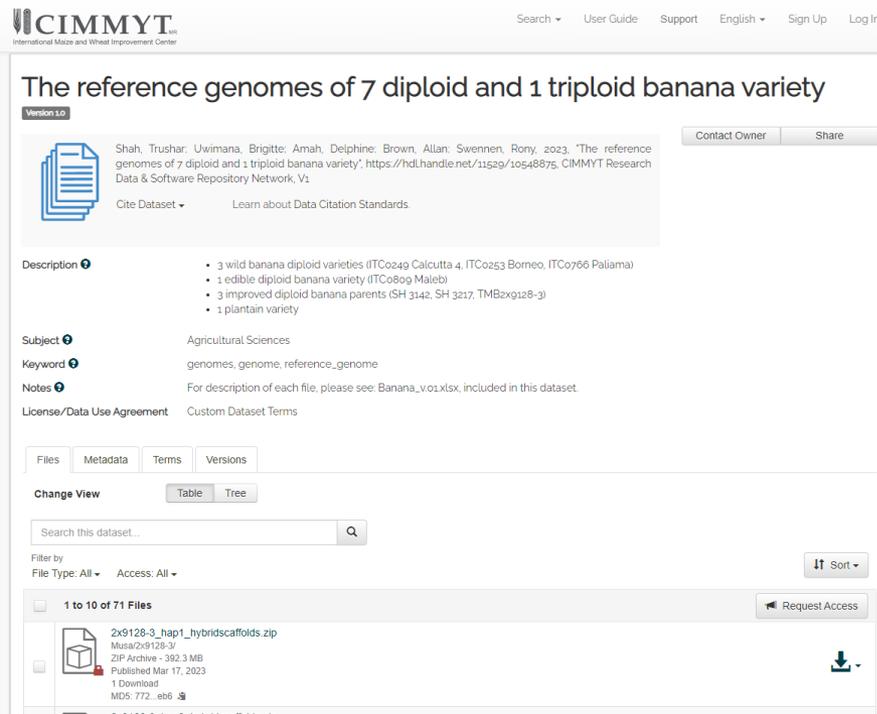
Seleccionar ficheros desde Dropbox.

 Subir desde Dropbox



# Big Data

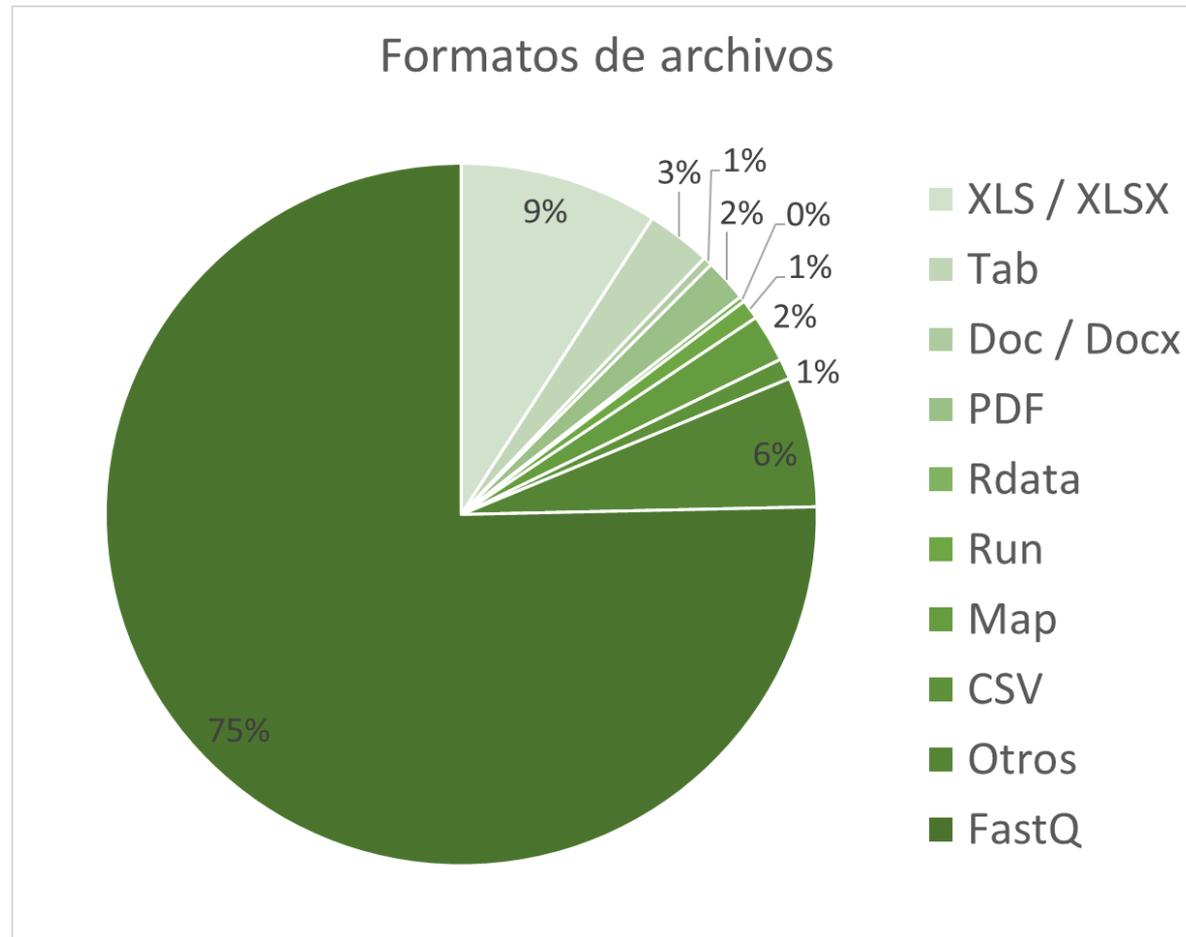
- Repositorio soporta aumentos en el volumen de datos y diversidad de formatos



The screenshot displays the CIMMYT Data Repository interface. At the top, the CIMMYT logo and navigation links (Search, User Guide, Support, English, Sign Up, Log In) are visible. The main heading is "The reference genomes of 7 diploid and 1 triploid banana variety" (Version 1.0). Below the heading, there is a "Cite Dataset" section with a citation: "Shah, Trushar; Uximama, Brigitte; Amah, Delphine; Brown, Allan; Swennen, Rony, 2023. 'The reference genomes of 7 diploid and 1 triploid banana variety'. https://hdl.handle.net/11529/10548875. CIMMYT Research Data & Software Repository Network, V1." and a "Learn about Data Citation Standards" link. The "Description" section lists: "3 wild banana diploid varieties (ITC0249 Calcutta 4, ITC0253 Borneo, ITC0766 Palaima), 1 edible diploid banana variety (ITC0809 Maieb), 3 improved diploid banana parents (SH 3142, SH 3217, TMB2x9128-3), and 1 plantain variety." The "Subject" is "Agricultural Sciences", and the "Keyword" is "genomes, genome, reference\_genome". The "Notes" state: "For description of each file, please see: Banana\_v01.xlsx, included in this dataset." The "License/Data Use Agreement" is "Custom Dataset Terms". Below this, there are tabs for "Files", "Metadata", "Terms", and "Versions". The "Change View" section has "Table" and "Tree" options. A search bar is present with the text "Search this dataset...". The "Filter by" section shows "File Type: All" and "Access: All". The "Files" section shows "1 to 10 of 71 Files" and a "Request Access" button. The first file listed is "2x9128-3\_hap1\_hybridscaffolds.zip" (Musa2x9128-3) with a size of 392.3 MB, published on Mar 17, 2023, and has 1 download. The second file is "2x9128-3\_hap2\_hybridscaffolds.zip".



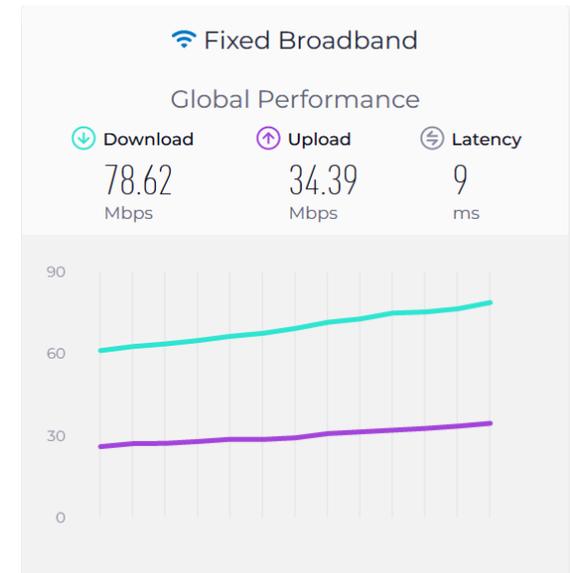
# Tipos de archivos actuales



# Próximos pasos

# Velocidad de Internet

- De acuerdo a los datos de Speedtest, países emergentes tienen una velocidad de Internet menor al promedio global
- Velocidad de descarga (Banda ancha fija hasta marzo de 2023)
  - Estados Unidos: 198.17 Mbps
  - **Promedio mundial: 78.62 Mbps**
  - Argentina: 54.05 Mbps
  - México: 50.15 Mbps
  - Etiopía: 6.13 Mbps



# Ámbitos de interés de CIMMYT

- Países emergentes
- Disminución de latencia
  - Red descentralizada de almacenamiento en la nube
  - Arquitectura distribuida
  - Enrutamiento inteligente



# Adiciones a Dataverse

- Herramientas para mejorar el soporte a BigData
  - DVWebloader: investigadores tengan la facilidad subir archivos desde una carpeta local y subcarpetas
  - Rsync
  - Futuras herramientas que se adicionen



# Contacto

- Alejandra Tenorio Robles  
[a.tenorio@cgiar.org](mailto:a.tenorio@cgiar.org)
- Jesús Herrera de la Cruz  
[j.herrera@cgiar.org](mailto:j.herrera@cgiar.org)





Gracias