Curation and Quality Control of Data in EpiFlu™ Database

Anne Pohlmann
FLI Riems

Singapore
2nd BLE-GISAID Symposium
24/25 October 2014

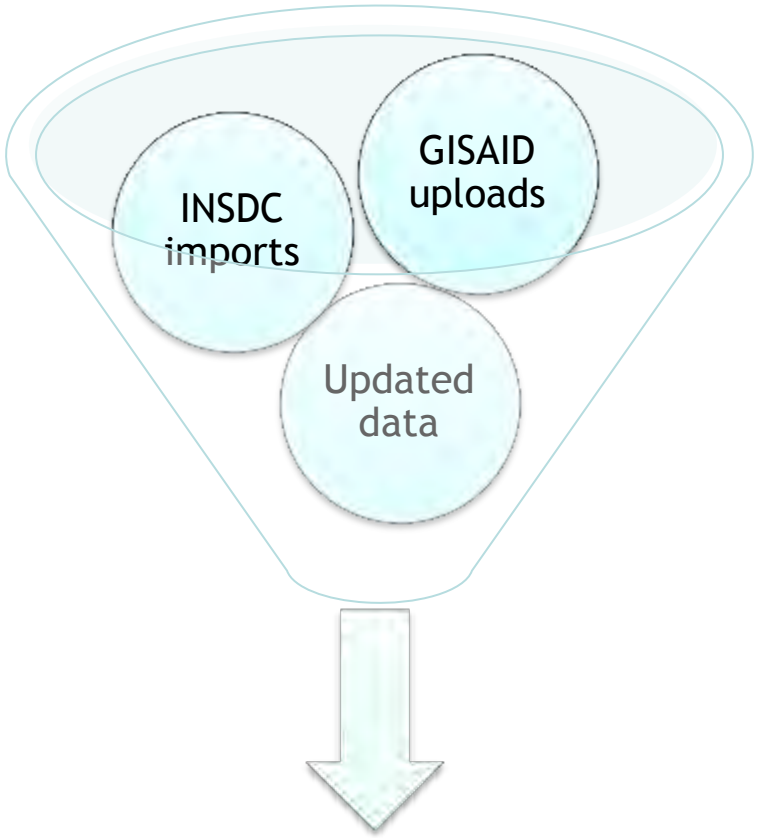
The host of GISAID database

As of 2011 the Federal Republic of Germany represented by the Federal Ministry of Agriculture and Food (BMEL) is official host of the GISAID database

- **Friedrich-Loeffler-Institute (FLI), Federal Research Institute for Animal Health**
  - Content-based data curation
  - Data import and correction
  - Contact point for questions concerning data curation

- **Federal Office for Agriculture and Food (BLE)**
  - Technical operation and maintenance of the database and web server
  - Monitoring and surveillance of worldwide availability
  - Data security and integrity

What data need to be curated?

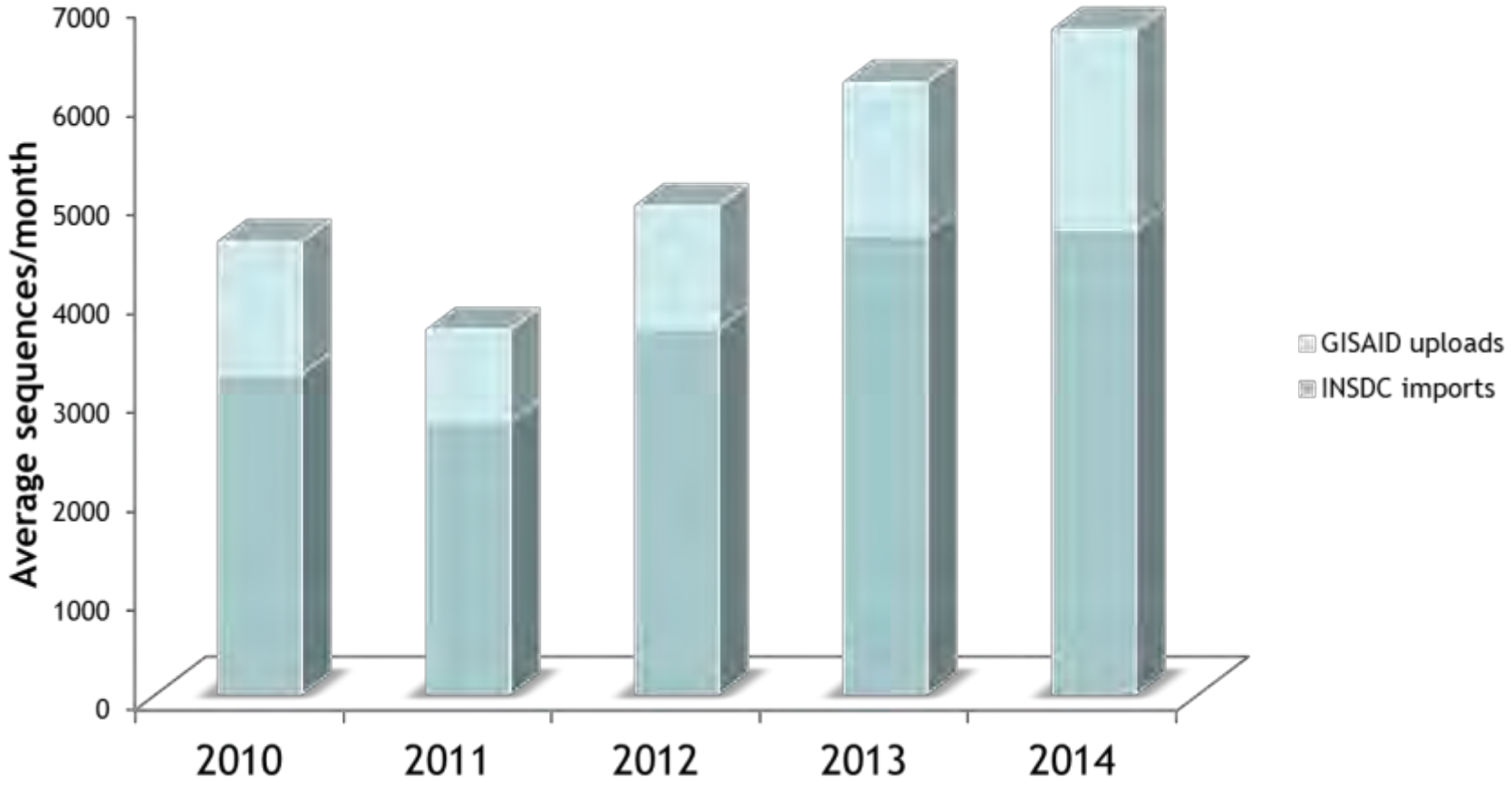


Curated Data

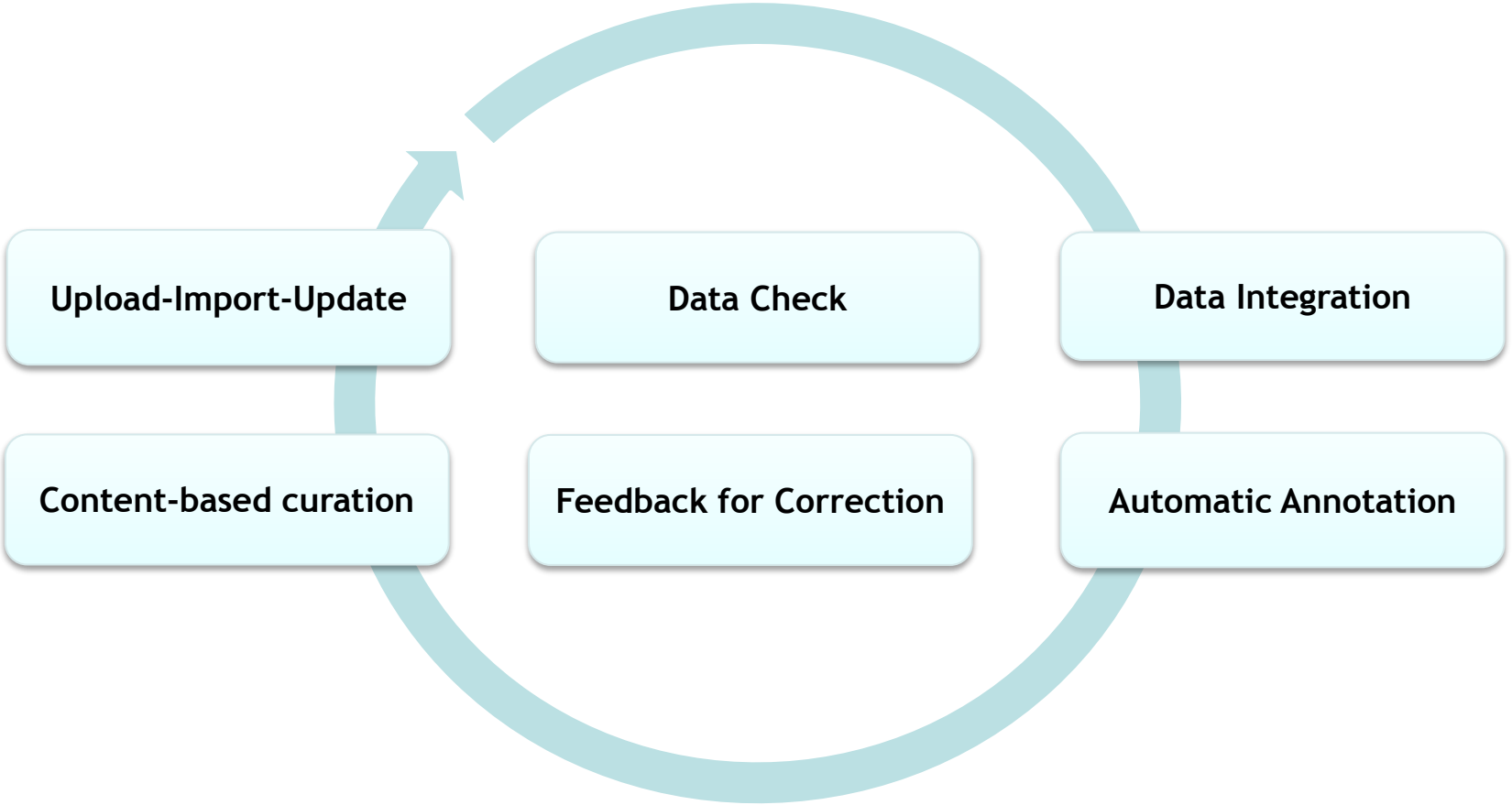
Nucleotide sequences
- Length, code
- Coding sequences
- Protein sequences
- …

Virus information
- Type, subtype, lineage
- Dates
- Host information
- Geographic information
- Resistance information
- Antigenic information
- …

How many data needs to be curated?



Data curation is an interactive and iterative process



Upload tools assists submitter to ensure correct and complete data

Data upload:
- Single Upload
- Batch upload

Submitter

GISAID

GISAID published: 4.269 viruses with 8.130 Sequences Total count: 104

Browse | Upload | Workset Management | Administration

Virus

Virus Detail

Virus Name * A/New York/359/2005
Example: A/Wisconsin/2145/2001 or A/chicken/Rostov/864/2007

Passage Category: Cell

Passage History * cell
Example: c1/c2

Harvest Date

Virus Type * A Subtype H: – Choose Subtype N: 1 Lineage: None

Host Group * Human

Host Group options: None, Animal, Environment, Human, Laboratory derived, Unknown

Subtype H options: – Choose –, None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17

Interactive process to improve data quality

Data upload:
- Single Upload
- Batch upload

Upload Check:
- Mandatory fields
- Drop-down lists

Submitter

Data Integration:
Integrates data to GISAID data format and fields

Automatic Annotation:
Segment type, subtype lineages, CDS, protein

Automatic annotation in GISAID 2.0

CLC Qiagen annotation server:

- Automatic annotation of segment type, influenza type and lineages
- Identification of virus subtype from HA and NA segments
- Determination of coding sequences and integration of corresponding proteins

CDC Atlanta H5N1 Clade annotation:

- Designation of clades for H5N1 sequences

Interactive process to improve data quality of GISAID uploads

Data upload:
- Single Upload
- Batch upload

Upload Check:
- Mandatory fields
- Drop-down lists

Data Integration:
Integrates data to GISAID data format and fields

Submitter

Feedback to the Submitter:
Incomplete data, frameshifts, differences in subtype of virus, …
Correction of entries!

Automatic Annotation:
Segment type, subtype lineages, CDS, protein

GISAID database

Data Release:
Submitter releases data to the public

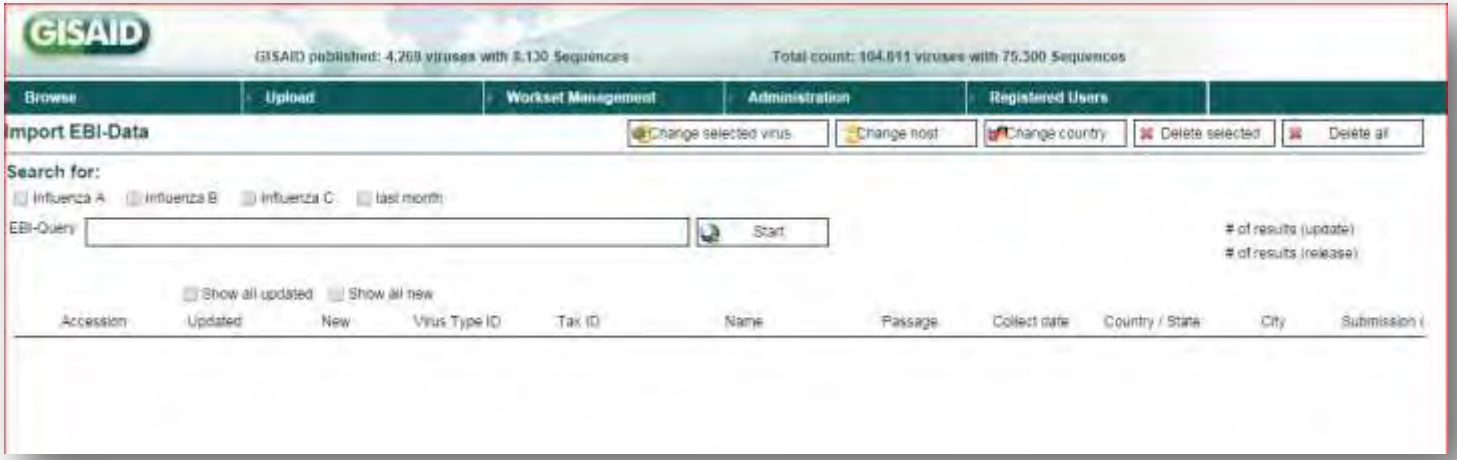
Import and quality control of INSDC data

Data import:
- New Influenza data
- Updated influenza data

Import Check:
- Parse data
- Check data

Curator

Correction by Curator:
Complete data, identification of passages, check for duplication, insert detailed location, use scientific host name (Avian), annotation, …



Import and quality control of INSDC data

Data import:
- New Influenza data
- Updated influenza data

Import Check:
- Parse data
- Check data

Data Integration:
Integrates data to GISAID data format and fields

Curator

Correction by Curator:
Complete data, identification of passages, check for duplication, insert detailed location, use scientific host name (Avian), annotation, …

Automatic Annotation:
Segment type, subtype lineages, CDS, protein

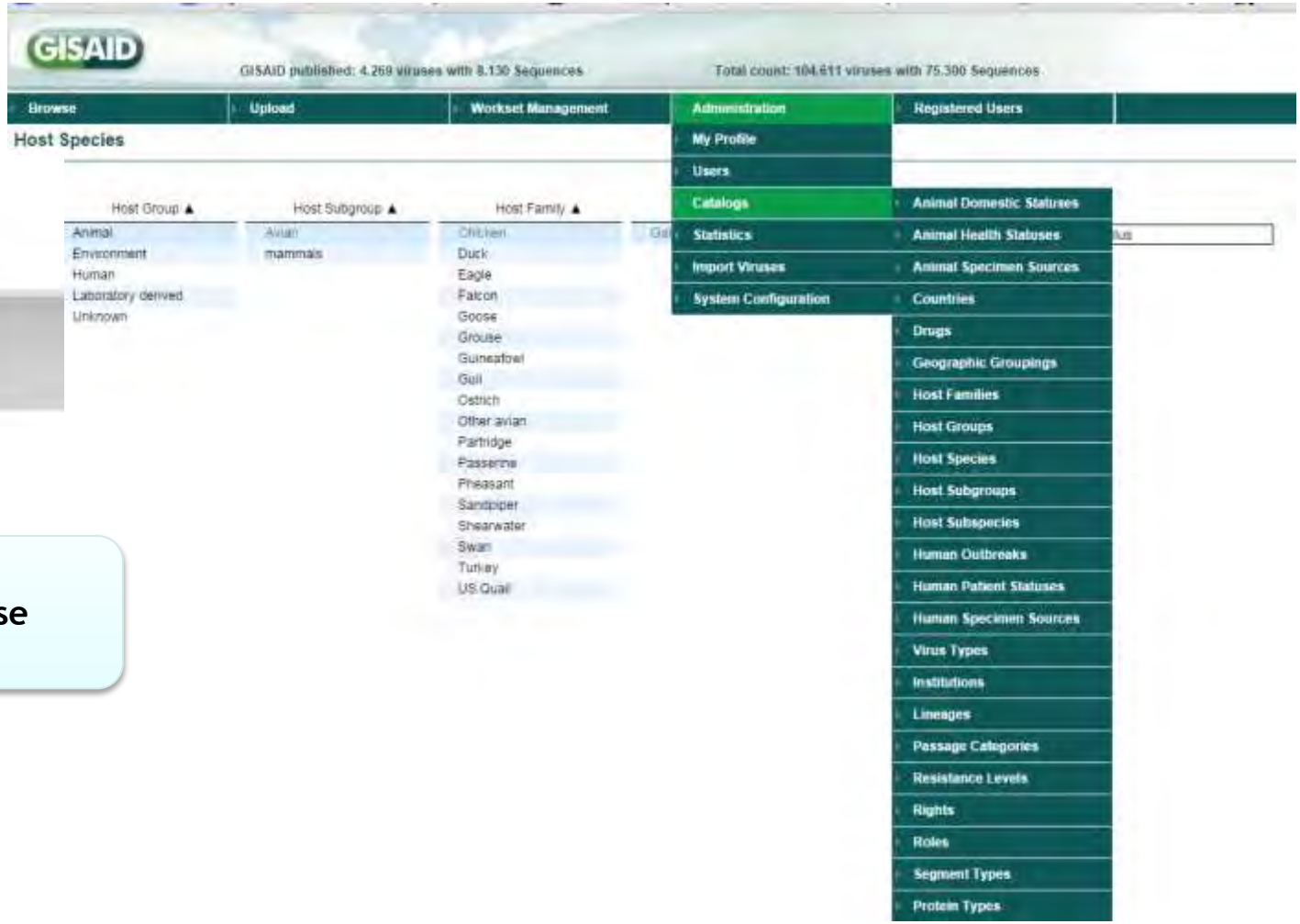
GISAID database

Data Release:
Curator releases data to the public

Curation of Catalogue Content

Admin

GISAID database



Content-based data curation

Curator

GISAID database

Curation

Dedicated software, flexible and dynamic

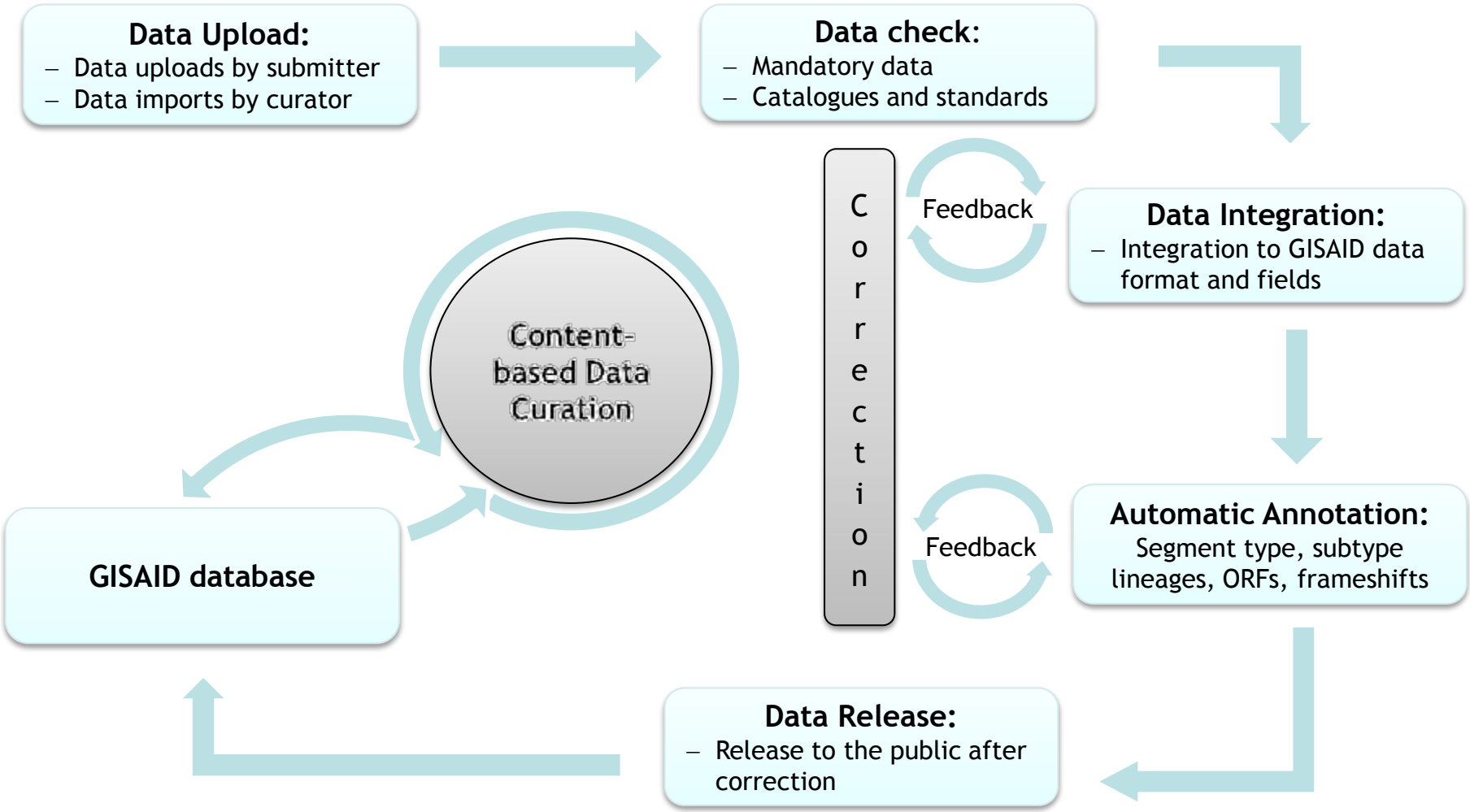
- Find inconsistencies
- Correct entries
- Add information

Isolate_Id	Isolate_Name	Subtype	Lineage	Location	Host	Collection_Date	Update_Date
EPI_ISL_90291	A/Turkey/Germany/R617/2007	A / H6N2		Europe / Germany	Turkey	2011-05-26	2011-05-26
EPI_ISL_90292	A/Turkey/Ontario/6118/1968	A / H8N4		North America / Canada	Turkey	2011-05-26	2011-05-26
EPI_ISL_94864	B/MACAU/200392/2011				Human	2011	
EPI_ISL_94865	B/MACAU/601485/2010	B	Victoria		Human	2010-08-26	2011-08-15

Lack of location

Inconsistent collection date

The GISAID 2.0 Curation Concept



Thanks

FLI | Martin Beer | Dirk Hoeper | Thomas Mettenleiter

BLE | Jan Mark Pohlmann | Holger Heuser | Hanns-Christoph Eiden

GISAID | Alan Hay | Peter Bogner

CDC Atlanta | Catherine Smith | Rebecca Garten | Nancy Cox

WHO CC Melbourne | Naomi Komadina

Scopeland | Claudia Peißert | Robert Lembcke | Marko Bacetic

CLC Qiagen | Ben Turner

BII | Sebastian Maurer-Stroh | Raphael Lee Tze-Chuen

Thank you for your attention.

service@gisaid.org

curator@gisaid.org