CDCB Data Flow and File Exchange

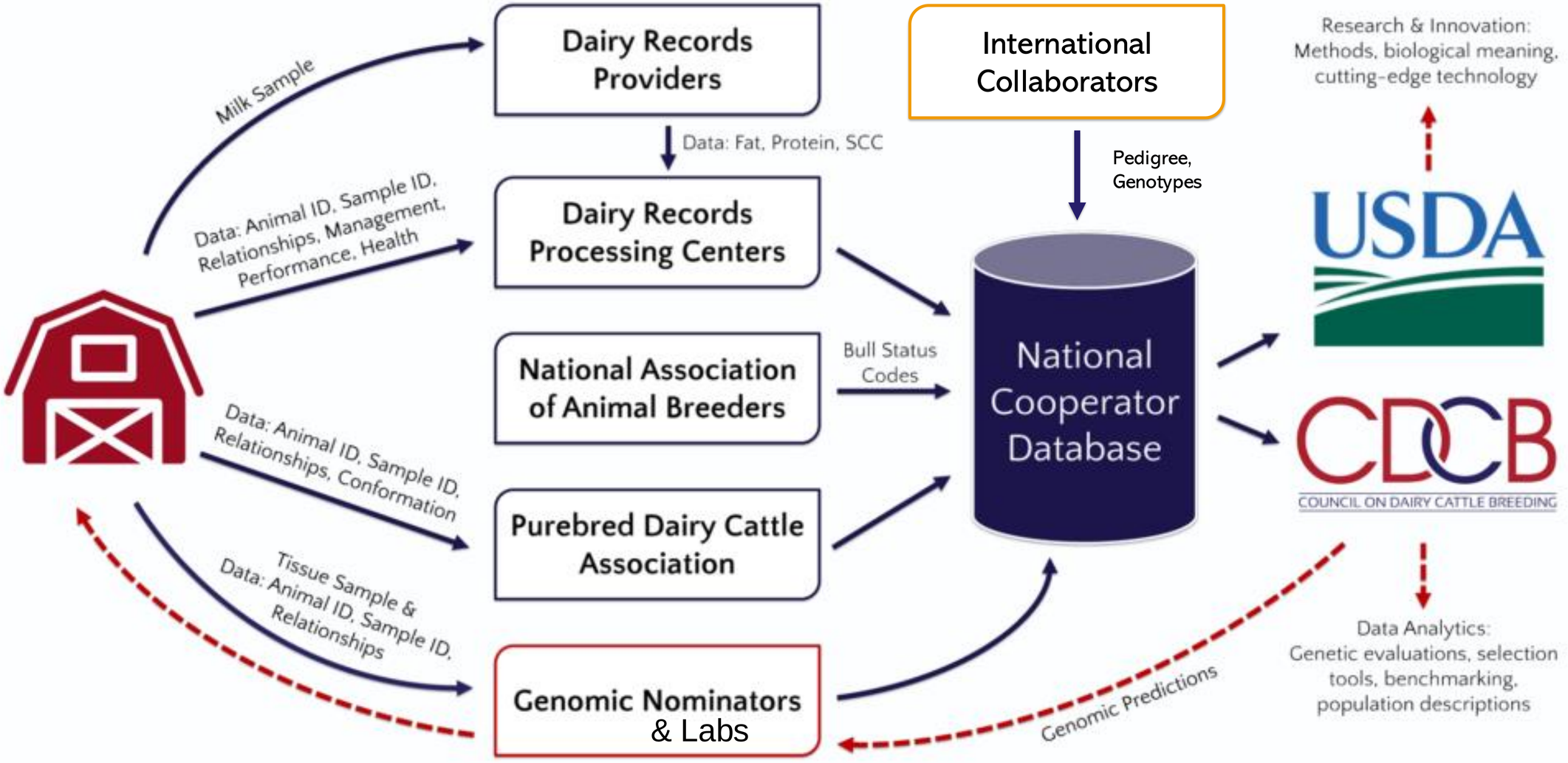
Genomic Data Analyst

Kaori Tokuhisa

7/30/2024

COUNCIL ON DAIRY CATTLE BREEDING

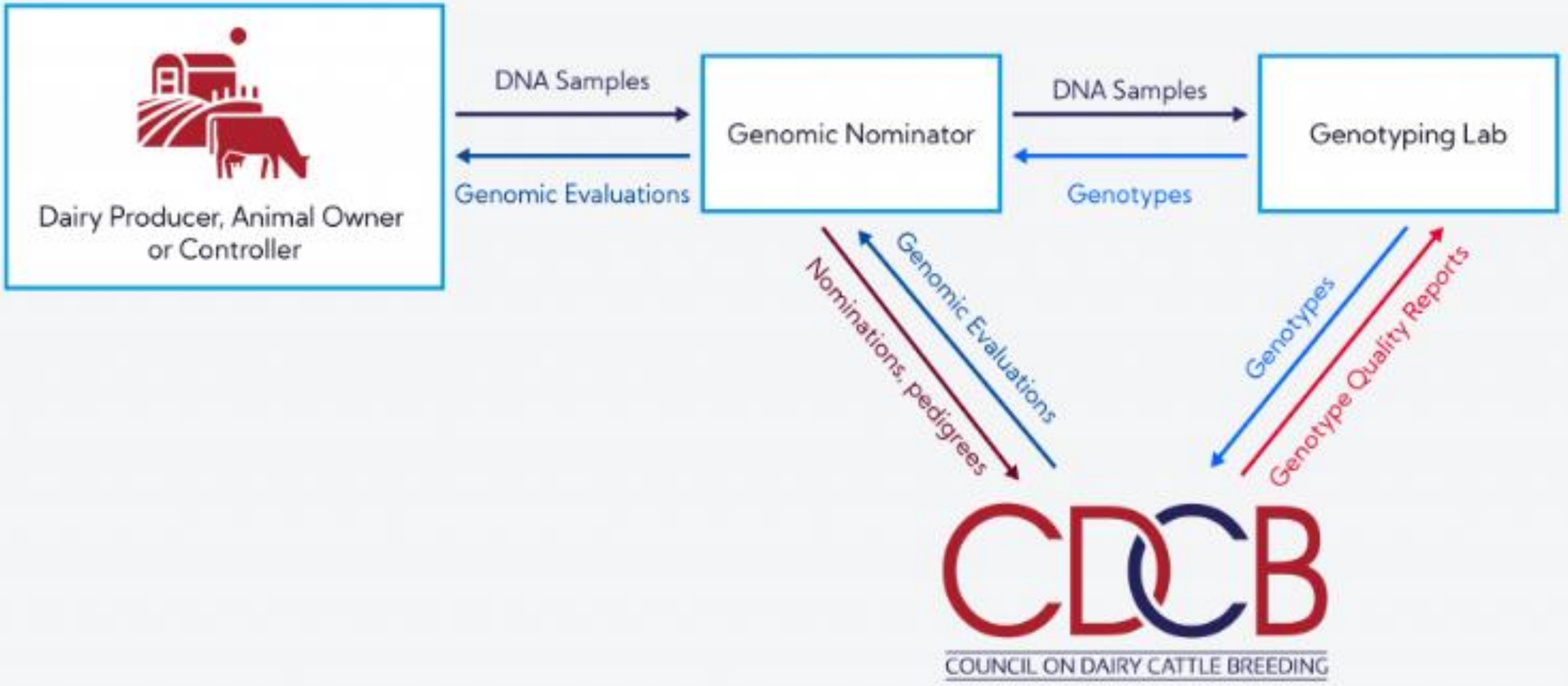
Data Providers



National Cooperator Database

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

How to Request a Genomic Evaluation



How Do Nominators & Labs Enter Data to DB?

- **Nominators**

Nomination + pedigree

-Submission of format1(G,P) through SFTP

-Submission through WebConnect

- **Labs**

Genotype Files (Sample Sheet +Final Report)

-Submission through SFTP

ONLY

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

SFTP

```
drwsrwsr-x 2 mneupane ftprun  8192 May 23 08:54 check
dr-xr-sr-x 2 root     ftprun    17 Jan 24 14:48 dev
drwsrwsr-x 2 mneupane ftprun     6 May 25  2022 FinalReport
drwsrwsr-x 2 mneupane ftprun     6 May 21 07:19 in
dr-srwsr-x 2 mneupane ftprun 12288 May 23 08:54 out
drwsrwsr-x 2 mneupane ftprun     6 May 25  2022 QC_Report
```

- Data exchanges via files happens in SFTP
- “in” folder for incoming files and “out” folder for outgoing files
- Lab gets additional folder called “check”
- Files get picked up by our automation if there is no pending job prior to your submission
- Some files in “out” directory are zipped after 24 hours to save storage (20240730.zip)
- Genotypes can be loaded from SFTPs only

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

SFTP vs WebConenct

Comparison	SFTP	WebConnect
Ease of use	Required to follow format1 format	Graphical User Interface. User friendly.
Number of records+submission	Batch submissions allowed	Limitation in number of submission (Data exchange for batch is available for some data)
Record type flexibility	Most record types are accepted (P,X,C,R etc)	Some record types are not supported (such as C, R,D)
System	Some infrastructure might need to be built, but automation is possible	No system required. Manual work involved

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Genotype Submission

- Labs submit Sample Sheet files (contains sample info), and Final Report files (contains actual genotypes) in zipped files to “Check” directory for QC.
- Once QC is done, many QC report files are placed in “out” directory.
- **Data QC** file contains pass/fail from QC check
  - put a comment (if there is any failed metrics), after investigating the issue and still want to proceed
  - No comments are required for submission that pass QC check
- Placing **Data QC** file in “in” directory triggers the system to proceed with loading.

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

General Files Distributed for Labs

- Zip file generated from **check** process results. YYYYMMDDXX_ZZZ.zip

- Zip file generated from **write mode** process final.results. YYYYMMDDXX_ZZZ.zip

YYYY-Year (2024)

MM- Month (07)

DD- Day (30)

XX –Any number or character

ZZZ- chip type (ex 50K for Illumina BovineSNP50 BeadChip. V1)

https://redmine.uscdcb.com/projects/cdcb-customer-service/wiki/Files_Generated_During_Check_Process

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

- DataQC_lab10-byte-date_*.csv
- Possible_Switch.csv
- chip_sum_bar.csv
- Conflicts_by_plate.csv
- Reassigned_genotypes.txt
- key_not_found.txt
- No_nomination.txt
- No_Match_Sample_ID_10-byte-date.txt
- Sample_ID10-byte-date.txt
- count.gt
- Lab_Conflicts10-byte-date.htm
- HWE10-byte-date.html
- Parent_Progeny_Conflicts10-byte-date.htm
- PPC_bySNP10-byte-date.htm
- LowCallSNP10-byte-date.htm
- Genomic_conflicts10-byte-date.htm (Check_Errors_20200911.csv)
- GE1k_progeny.csv
- Existing_genotype_negative_key.csv
- Excessive_Homozygous_LABYYYYMMDD.csv (upon loading)

General files distributed for Nominators

- Files are generated upon briefly 5 occasions. Weekly, monthly, Tri-annual, upon data submission, and daily processing (-o job)
- Files include information like: Evaluation, notifications, QC, and updates
- Notify files might be the file that you check most frequently

Frequency	Event	File Name
Weekly	Evaluation	BRD_NOM_yyyymmdd.csv (xml)
		BRD_NOM_yyyymmdd_haplo_data.csv
		BRD_NOM.BBRdata.yyyymmdd.csv
		XXXX_Add_Sire_YYYYMMDD.csv
		XXXX_Add_Cnst_Dam_YYYYMMDD.csv
		XXXX_Sugg_Dam_YYYYMMDD.csv
Monthly	Evaluation	NOM_YYMM.zip/NOM_YYMM.csv
		NOM_YYMM.zip/BRD_NOM_YYMM_BBRdata.csv
		NOM_YYMM.zip/NOM_YYMM_haplo_data.csv
		NOM_Check_Fee_Code_YYMM.csv
		NOM_Report_Card.csv
		NOM_Supplemental.txt
		BRD_NOM_MGS_unlikely_YYMM.csv
		XXXX_Add_Sire_YYMM.csv
		XXXX_Add_Cnst_Dam_YYMM.csv
		XXXX_Sugg_Dam_YYMM.csv
	ID-based breed code conflict	NOM_BRD_BBR90_BreedConflict_YYMM.csv
Triannual	Evaluation	Evaluation results available online
Upon data submission	Quality Control	notify.yyyymmdd.1
		LAB.YYYYMMDDXX.NOM.zip/NOM_LABCHIPYYYYMMDDXX_No_Nomination.csv
		LAB.YYYYMMDDXX.NOM.zip/NOM_Genomic_conflicts.htm
		LAB.YYYYMMDDXX.NOM.zip/NOM_Genotype_Conflicts.csv
		LAB.YYYYMMDDXX.NOM.zip/NOM_Parentage.csv
		LAB.YYYYMMDDXX.NOM.zip/NOM_Nominator_Report.csv
		LAB.YYYYMMDDXX.NOM.zip/NOM_PGS_unlikely.csv
		LAB.YYYYMMDDXX.NOM.zip/NOM_Excessive_Homozygous_LABYYYYMMDDXX.csv
Daily (5:00am and noon)	Update job (-o)	NOM_Parentage_yyyymmdd_TIMESTAMP.csv

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

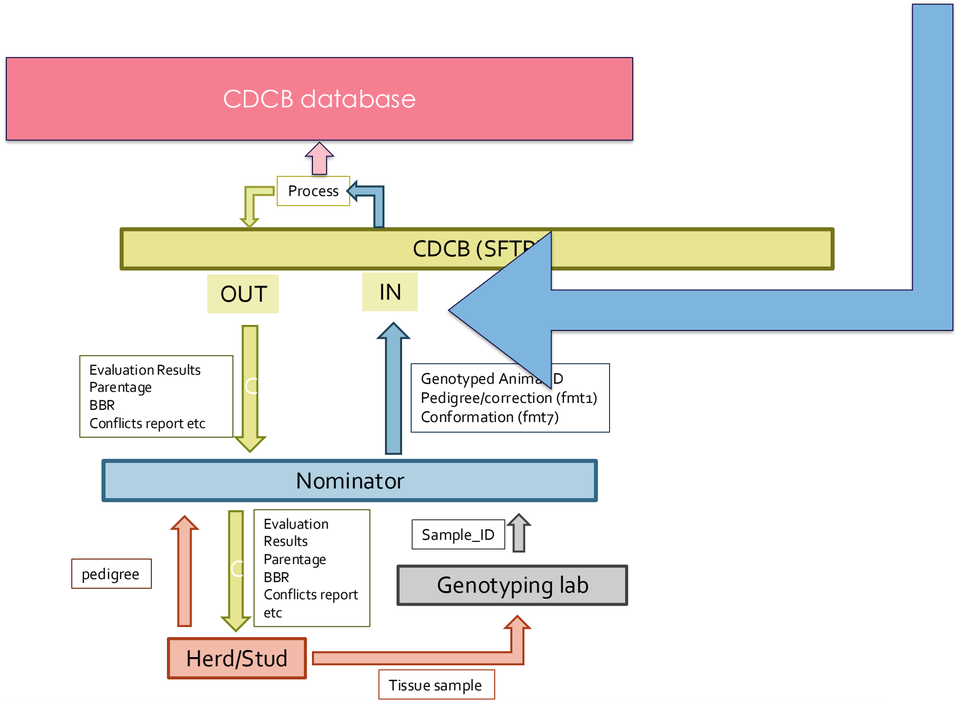
https://redmine.uscdcb.com/projects/cdcb-customer-service/wiki/CDCB_general_files_distributed_to_nominators

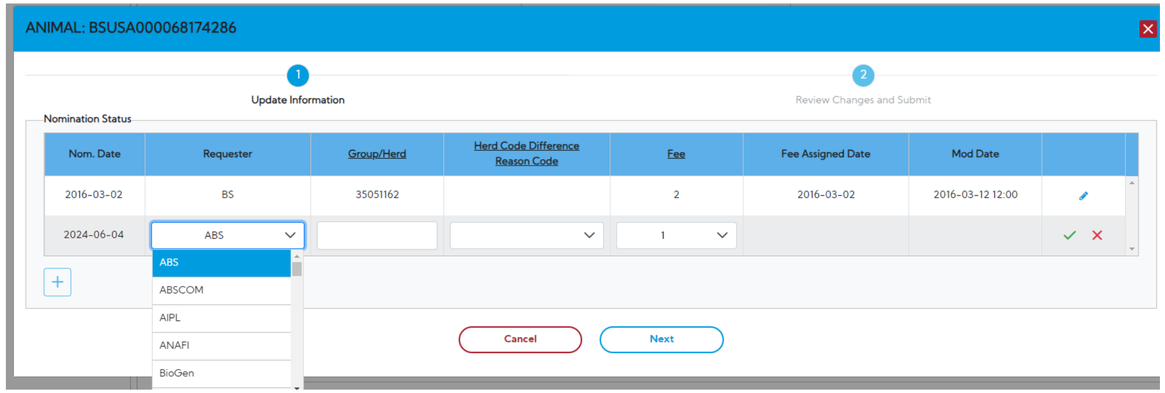
Nomination

YYYYMMDD.1GX

0FHOUSA000011111111HOUSA000000222222HOUSA000033333333A1B2C3D4E5F6 20160219B20160312G013HR000000NAMEOFANIMAL ET 35051162L2

- Nomination is a process of providing pedigree, assigning a fee code, and indicating what service you would like to receive.
- Nomination should be done before the genotype submission
- Notify file will report any nomination related errors
- I will talk more about how to nominate animals using WebConnect in my next presentation





https://redmine.uscdcb.com/projects/cdcb-customer-service/wiki/Create_a_format1G

https://redmine.uscdcb.com/projects/cdcb-customer-service/wiki/Format_1

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

You Are **NOT** Done Yet!!!

Please make sure if the nomination/pedigree was processed successfully by looking at notify files in your **out** directory!!!

- **notify_VAL.YYYYMMDD.1GX**

Notify whether the animal was nominated successfully or not regardless of existence of errors.

- **notify_ERR.YYYYMMDD.1[G]X**

Notify only if error/change needs to be informed.

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Summary

- Due to the variety and complexity of the data we exchange, CDCB outputs many kinds of files with different information (confirmation, error/conflicts, results, notification etc)
- Utilizing CDCB documentation, as well as your own internal documentation, should help with understanding these files
- Nomination using WebConnect will be discussed in my next presentation – Stay tuned!

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Thank you

CDCB

COUNCIL ON DAIRY CATTLE BREEDING