Data Ingest at the IVS Data Centers

Dirk Behrend, Mario Bérubé, John Gipson, Anastasiia Girdiuk, Markus Goltz, Taylor Yates, Pat Michael, Christophe Barache

12th IVS General Meeting

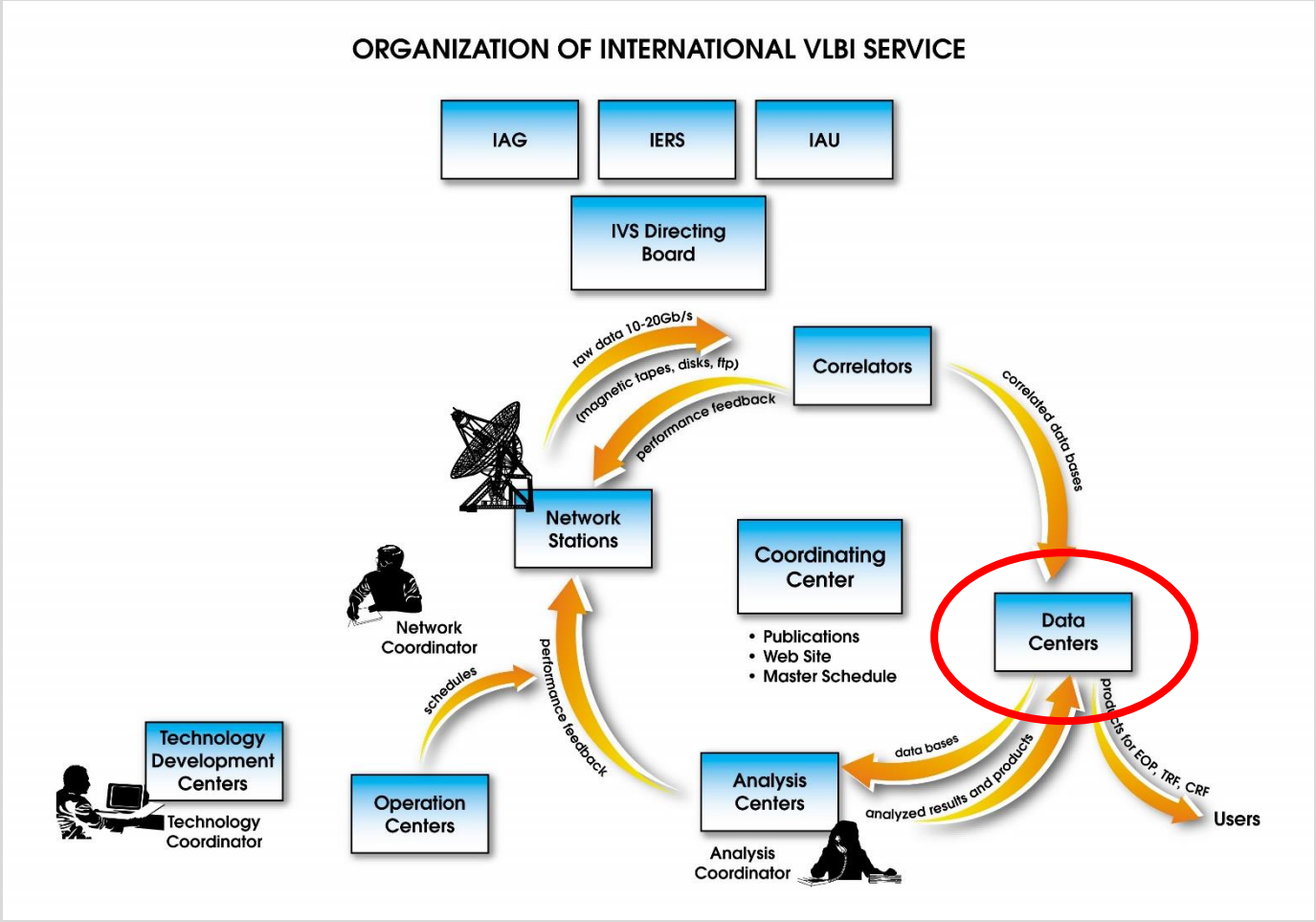
Finnish Cyberspace

March 29, 2022

Background

- Three **Primary IVS Data Centers** hold the IVS products and data files:
  - Crustal Dynamics Data Information System (CDDIS), Goddard, MD, USA
    - ► Poster S2-P09 (Yates et al.) on Tue @ 12:45 UT
  - Bundesamt für Kartographie und Geodäsie (BKG), Frankfurt, Germany
    - ► Poster S2-P08 (Girdiuk et al.) on Tue @ 12:45 UT
  - Observatoire de Paris (OPAR), Paris, France
- Data Centers mirror each other daily (every 4 hours) to ensure common holdings
- Primary Data Centers serve as the main method for disseminating IVS data and products

IVS Structure and Flow Diagram



- Data Centers (DCs) are *one of seven* component types

Data Center Structure

```
vlbi/
    |-- ivscontrol
        |-- ac-codes.txt
        |-- masteryy.txt
        |-- ...
        |-- ns-codes.txt
    |-- ivsdata
        |-- aux
        |   |-- yyyy
        |       |-- <ssssss>
        |           |-- <ssssss>.skd
        |           |-- <ssssss>.txt
        ...
        |-- swin
        |   |-- yyyy
        |       |-- yyyymmdd_<ssssss>_vnnn_swin.tar.bz2
        ...
        |-- vgosdb
        |   |-- yyyy
        |       |--yyMMMddCC.tgz
    |-- ivsdocuments
    |-- ivsformats
    |-- ivsproducts
```

ivsdocuments to be moved under **ivsproducts** and renamed to **soln_descr**

History of “ingest” Software

Author	Software	Data Centers
Frank Gomez	ivsincoming2ivs (ingest v.1)	CDDIS, BKG, OPAR
Nathan Pollack	ingest v.2	CDDIS
Justine Woo, Taylor Yates	**ingest v.3** [CDDIS]	CDDIS
Mario Bérubé, Anastasiia Girdiuk, Dirk Behrend	**ingest v.3** [BKG, OPAR]	BKG, OPAR

- Some features of “ivsincoming2ivs”:
  - monolithic script (10,000+ lines of code)
  - difficult to maintain, evolved over time
  - used for 20 years at all three DCs
- Divergence of data handling with “ingest v.2”

Some Basics of ingest v.3

- Modular design, Python-based
- First at CDDIS (GSFC), then for BKG/OPAR
- CDDIS: main program part of larger suite that supports all geodetic techniques: cannot be disentangled and ported to other DCs
- BKG/OPAR: different main program written that implements CDDIS main program functions
- Two common pieces between both suites:
  - data description files (DDF) and
  - validation scripts (for QC)

Some Statistics of ingest v.3

- Statistics on BKG/OPAR implementation
- Lines of code (incl. comment/blank lines): ~2500
- Main program:
  - Seven modules with a total of ~900 lines
  - Main module has ~500 lines
- Validation routines:
  - Some 30 routines with code of <50…125 lines
  - Average length of module: ~60 lines
- DDFs:
  - Some 70 files
  - Several DDFs call same validation routine

Code Excerpt for product_snx.py


```
def main(filename):
    """Validate a VLBI product file given in SINEX format."""
    try:
        # Initiate list of warnings
        warn_list = []

        # Read the SINEX file
        with open(filename, "r", encoding="latin-1") as reader:
            lines = reader.readlines()
        reader.close()

        # Verify SINEX format of header and trailer lines
        header = lines[0]
        trailer = lines[-1]
        if (header[:5] != "%=SNX") or (trailer[:7] != "%ENDSNX"):
            print ("FATAL: The given file is not in SINEX format!")
            return [False, "FATAL: Wrong format file."]
            exit(1)
        if header[58:59] != "R":
            print ("FATAL: The SINEX solution is not based on VLBI data!")
            return [False, "FATAL: Not a VLBI SINEX file."]
            exit(1)

        # Verify integrity of SINEX file blocks, ensure inclusion of required blocks
        # incorporating SINEX solution methods 6a, 6b, 6c (after reading contained
        # blocks)
        required_blocks = ["FILE/REFERENCE", "NUTATION/DATA", "PRECESSION/DATA",
                           "SOURCE/ID", "SITE/ID", "SITE/ECCENTRICITY",
                           "SOLUTION/EPOCHS", "SOLUTION/APRIORI"]
```

Tasks Done by “ingest”

- Main program: ► filename check
  - Build proper name from applicable control files (i.e., Master files, *ac-codes.txt, ns-codes.txt*)
  - Check proper name vs. filename, compression
  - Reject file if no match or wrong compression
- Validation routine: ► QC step
  - Check integrity of content (e.g., header/trailer lines and block structure in SINEX files)
  - Extract metadata (e.g., start and stop times of session related files)
  - Reject file if prior steps fail

Impact on Submissions

- Enhanced quality control (QC):
  - Strict enforcement of filename conventions
  - Stricter quality checks of file content, i.e., verify that standard formats are followed (SKD, VEX, SINEX, EOP format, etc.)
  - ► Some files that used to pass are now rejected!
- Need for improved notification system:
  - E.g., at CDDIS “successful upload” indicates “file received” but not successful pass of QC
  - Possible options (feedback requested):
    - *Email notification of success/failure* or
    - *Webpage listing of last ~200 submissions*

Status of Rollout

- August 2, 2021: all three data centers switched to new ingest
- Information on “Conventions for Submitting Data and Product Files to the IVS Data Centers”: https://ivscc.gsfc.nasa.gov/products-data/DataCenter_File_Conventions.pdf
- Cleanup of repository
  - Reprocessing of existing data holding
  - Removal of erroneous files
  - Renaming of misnamed files
  - Tentative date: April 30, 2022

Excerpt of Conventions Doc

File type	Name convention	Compression	Example
schedule file	<ssssss>.skd		r11002.skd
session notes	<ssssss>.txt		r11002.txt
log files	<ssssss>nn.log		r11002ht.log
full log files	<ssssss>nn_full.log	.bz2	r11002k2_full.log.bz2
...			
SWIN files	yyyymmdd_<ssssss>_vnnn_swin.tar	.bz2	20210607_r11002_v001_swin.tar.bz2
vgosDB	yyMMMddCC	.tgz	21JUN07XA.tgz
...			
CRF	aaaccccc.crf	.gz	opa2021a.crf.gz
	aaaccccc.stats.crf	.gz	opa2021a.stats.crf.gz
EOPS	aaaccccc.eops	.gz	gsf2020a.eops.gz
	aaaccccc.stats.eops	.gz	gsf2020a.stats.eops.gz
Daily SINEX	yyMMMddCC_aaaccccc.snx	.gz	21JUN07XA_bkg2020a.snx.gz
...			
DOCS	aaaccccc.crf.txt		opa2021a.crf.txt
	aaaccccc.eops.txt		gsf2020a.eops.txt
	aaaccccc.dsnx.txt		bkg2020a.dsnx.txt

How to Add New Data Type

- Community:
  - Discuss and define format description
  - Determine storage needs
- DC group:
  - Write DDF, including location in directory tree
  - Write validation routine
  - Extend storage capacity, if needed
  - Test DDF, routine in shadow ingest system
- Community and DCs:
  - Submit files of new data type
  - Correct any kinks

Summary and Outlook

- As of August 2, 2021, a new ingest software is running at the IVS DCs of CDDIS, BKG, OPAR
- BKG and OPAR use same suite; CDDIS uses its own flavor of the main ingest program
- DDFs and validation scripts are the same for all
- DCs will reprocess existing holding (cleanup); tentatively scheduled for April 30, 2022
- Following cleanup at all three DCs, the data holdings will be synchronized
- Then mirroring should ensure identical holdings going forward
- Contact the DCs: ► ivs-datcen@lists.nasa.gov

Thank you.