IGS Data Flow – Today and Proposal for the Future

Angelyn Moore
JPL/Caltech
USA

Carey Noll
NASA GSFC
USA

Carine Bruyninx
ROB
Belgium

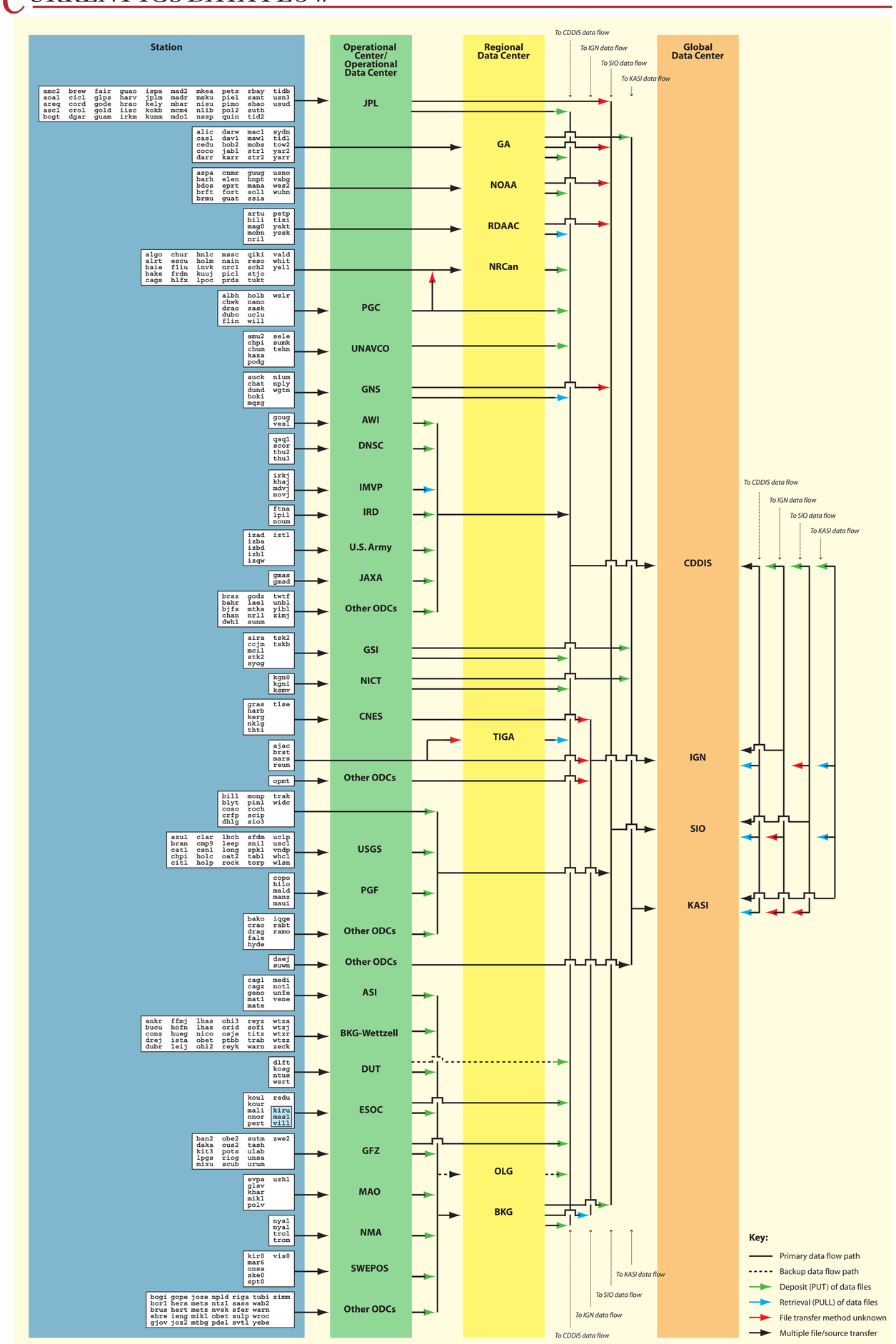
Michael Scharber
SIO/IGPP
USA

ABSTRACT

The IGS analysis centers and user community in general need to be assured that the data centers archive a consistent set of files. Changes to the archives can occur because of the re-publishing of data, the transmission of historic data, and the resulting re-distribution (or lack thereof) of these data from data center to data center. To ensure the quality of the archives, a defined data flow and method of archive population needs to be established. This poster will diagram and review the current IGS data flow, discuss problems that have occurred, and provide recommendations for improvement.

The associated position paper details specifications for defining the IGS data flow in such a way to solve several actual problems that exist in the IGS infrastructure. This plan requires agreement and action from all stations' operational centers, all Data Centers, of the Central Bureau, and quite possibly of Analysis Centers. Relatively simple rules discussed in the paper will address the actual problems in the IGS data flow without undue pain to participants and users.

CURRENT IGS DATA FLOW



CURRENT IGS DATA CENTERS

The following data centers are referenced in the current data flow diagram:

Operational Center/Operational Data Centers:

ASI	Italian Space Agency
AWI	Alfred Wegener Institute for Polar and Marine Research, Germany
BKG-Wettzell	Bundesamt für Kartographie und Geodäsie, Germany
CNES	Centre National d'Etudes Spatiales, France
DNSC	Danish National Space Center
DUT	Delft University of Technology, The Netherlands
ESOC	European Space Agency (ESA) Space Operations Center, Germany
GFZ	GeoForschungsZentrum, Germany
GNS	Institute of Geological and Nuclear Sciences, New Zealand
GSI	Geographical Survey Institute, Japan
IMVP	Institute of Metrology for Time and Space, Russia
IRD	Institut de Recherche pour le Développement, New Caledonia
OLG	Observatory Lustbuehel Graz, Austrian Academy of Sciences
JAXA	Japan Aerospace Exploration Agency
MAO	Main Astronomical Observatory of the National Academy of Sciences of Ukraine
NICT	National Institute of Information and Communications Technology, Japan
NMA	Norwegian Mapping Authority
PGC	Pacific Geoscience Centre, NRCan, Canada
PGF	Pacific GPS Facility, USA
SWEPOS	Swedish National Network/Lantmäteriet
UNAVCO	University NAVSTAR Consortium, USA
U.S. Army	U.S. Army, Iraq
USGS	United States Geological Survey

Regional Data Centers:

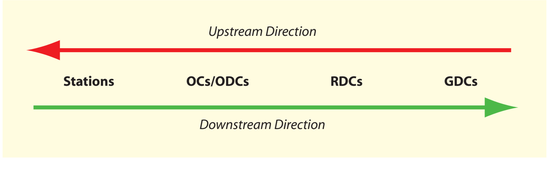
BKG	Bundesamt für Kartographie und Geodäsie, Germany
GA	Geoscience Australia
NOAA	National Oceanic and Atmospheric Administration, USA
NRCan	Natural Resources of Canada
RDAAC	Regional GPS Data Acquisition and Analysis Center on Northern Eurasia, Russia
TIGA	Tide Gauge Benchmark Monitoring

Global Data Centers:

CDDIS	Crustal Dynamics Data Information System, NASA GSFC, USA
IGN	Institut Géographique National, France
KASI	Korean Astronomy and Space Science Institute
SIO	Scripps Institution of Oceanography, USA

DEFINITIONS

- Data Centers (DCs): make data available to the public and include Global Data Centers (GDCs), Regional Data Centers (RDCs), and Operational Data Centers (ODCs).
- Operational Center (OC): The agency that is responsible for making a station's data available to the IGS community. An OC may or may not be an ODC, depending on whether it makes data publicly available. A station may be its own OC. An OC pushes data to DCs.
- Operational Data Center (ODC): an OC that offers public access to data from a set of stations, usually those managed by itself or a partner agency. The details of getting the data from the stations may vary.
- Regional Data Center (RDC): collect and offer data from many stations and agencies in a region. The RDC must allow data to be pushed to it from OCs/ODCs. RDCs will push data to GDCs.
- Global Data Center (GDC): collect and offer data from all IGS Reference Frame stations, as well as offering products (see DC charter). GDCs must allow data to be pushed to them from OCs/ODCs, RDCs, and other GDCs. GDCs will push data to other GDCs to equalize data holdings of IGS Reference Frame stations (at a minimum).
- Upstream: the direction of flow of data (from station to final destination).



CEN 04/26/2006

MOTIVATION FOR NEW IGS DATA FLOW SPECIFICATION

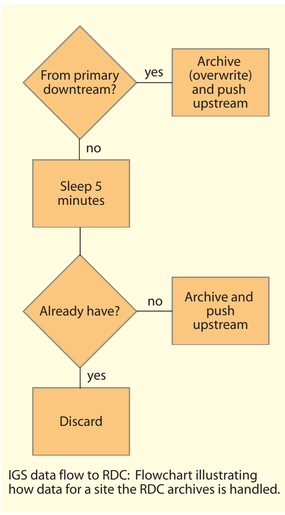
- Users expect consistent, timely data flow despite data center outages
- IGS data archiving must be fully automated
- Data flow in failover mode when outages occur must be fully automated
- Data files must be consistent between data centers
- Data revisions must populate to all archiving data centers
- Users must be aware of data revisions

SETUP

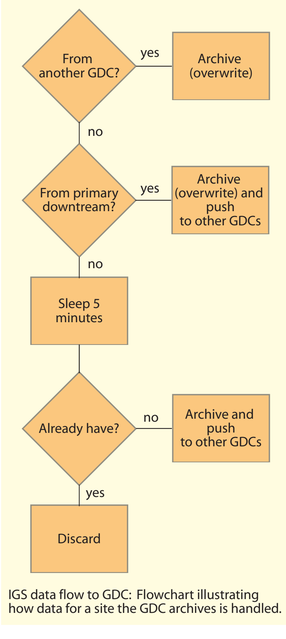
- OCs/ODCs will define two upstream DCs of any level to flow data to: one primary and one secondary. The most upstream DC will be the primary. These DCs will be listed in section 13 of each station's IGS site log.
- RDCs will define two GDCs to flow data to: one primary and secondary.
- The CB will collect, combine, and publish a table giving the primary DC for each station/OC, ODC, and RDC.
- Each DC must be able to automatically determine whether it is a primary recipient for a given data file submitted to it, either by consulting the official table at the IGS CB or by setting up, with downstream providers, data drop-off rules that identify the pushed data are on a primary vs. secondary path. GDCs must also be able to distinguish data submitted from other GDCs vs. data submitted from downstream sources.

RULES

- The official IGS data flow will be a **PUSH-ONLY method** (currently DCs use both PUSH and PULL to populate their archives).
- **OCs** will always push station data to their two DCs simultaneously or in immediate succession. If a data file is later corrected, the station will transmit the corrected file to both DCs.
- **All DCs** must sleep (wait) 5 minutes before acting on a file pushed to it on a secondary path. This action is presumed in all the rules below.
- **ODCs** and **RDCs** will always push all IGS data pushed to them from downstream to their two upstream DCs. If a file already received is received again from a primary route downstream, the ODC/RDC will replace the file with the version most recently received through the primary route, and push it to its two upstream DCs. If a file already received is received again from a secondary route downstream, the ODC/RDC will discard the newly received version.
- **GDCs** will push all IGS data received from a primary downstream source to all other GDCs. A GDC will also push data received from a secondary source to other GDCs after a 5 minute wait, if a copy has not already been received from a primary source. If a file already received is received again from a downstream station/DC for which the GDC is primary, the file will be replaced with the new copy and pushed to the other GDCs. If a file already received is received again from a secondary downstream route, the new copy will be discarded and not pushed to other GDCs. GDCs will overwrite data in the archive with data newly provided from another GDC. GDCs may discard data from stations they do not archive, recalling that GDCs must minimally archive all IGS Reference Frame stations.
- Any file PULLed by a DC becomes a private copy and must not be offered on public IGS DC areas or pushed to upstream DCs. The typical scenario for this activity is a DC gathering data for the convenience of an AC at the same institution. Data obtained by a pull mechanism becomes outside the IGS data flow.
- DCs must regard the pushed data as **READ-ONLY**. The only exception to this rule is that a DC may agree to accept data in alternate compression formats (e.g., .gz) and recompress the files in the IGS standard compression mode (.Z, Unix compression) prior to publishing. The underlying RINEX file must still be regarded as READ-ONLY.



IGS data flow to RDC: Flowchart illustrating how data for a site the RDC archives is handled.



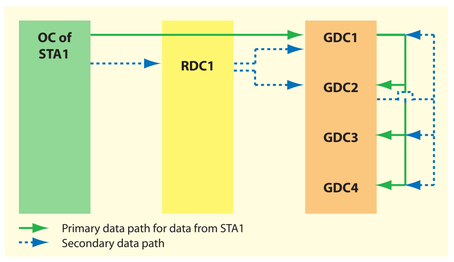
IGS data flow to GDC: Flowchart illustrating how data for a site the GDC archives is handled.

EXAMPLES

Given the following data flow scenario:

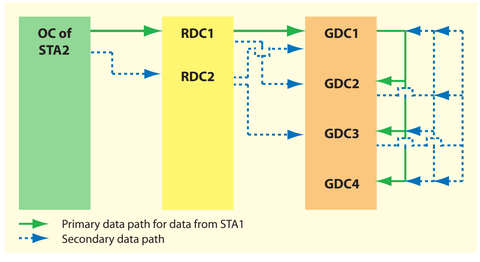
OC	Primary DC	Secondary DC
STA1	GDC1	RDC1
STA2	RDC1	RDC2
STA3	GDC1	GDC2

DC	Primary DC	Secondary DC
RDC1	GDC1	GDC2
RDC2	GDC1	GDC3



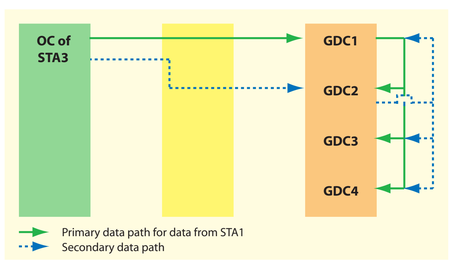
Upload to a GDC (primary) and RDC (secondary):

- The primary path for data for station STA1 is STA1→GDC1 and from GDC1→GDC2,3,4.
- If this path is completed, GDC1 and GDC2 will discard copy received from RDC1 via STA1⇢RDC1⇢GDC2 (RDC1 is not primary source of STA1 data).
- If GDC2 receives STA1 via STA1⇢RDC1 and by 5 minutes later does not receive data via STA1→GDC1, RDC1 will archive STA1 and push data to GDCs (STA1⇢RDC1⇢GDC2,3,4).
- GDC2 will discard copy from RDC1 if it already has a copy from GDC1 (RDC1 is not primary for STA1).
- GDC2 will replace copy received from RDC1 with copy received via STA1→GDC1→GDC2 (GDCs overwrite with a copy provided from another GDC).
- If RDC1 is disabled, data continues from STA1→GDC1→GDC2,3,4 (primary data flow).
- If GDC1 is disabled, GDC2 will receive STA1 data via STA1⇢RDC1⇢GDC2⇢GDC3,4. When primary copy from STA1 is eventually pushed via STA1→GDC1→GDC2,3,4, GDCs will overwrite previous copy with one from GDC1 because GDC1 is primary for STA1.



Upload to two RDCs (primary and secondary):

- The primary path for data for station STA2 is STA2→RDC1→GDC1→GDC2,3,4.
- If this path is completed, GDC1 will discard second copy sent from RDC2 (RDC1 is primary source of STA2 data).
- GDC2 will discard copy from RDC1 (primary path is from GDC1).
- If GDC1 is disabled, GDC2,3,4 will receive STA2 data via STA2→RDC1⇢GDC2⇢GDC3,4 or STA2⇢RDC2⇢GDC3⇢GDC2,4, whichever is completed faster (most likely RDC1, since RDC1 will not sleep for 5 minutes whereas RDC2 will).
- When GDC1 is restored, data will be overwritten with copy from RDC1→GDC1→GDC2,3,4 (GDCs overwrite with a copy provided from another GDC).



Upload to two GDCs (primary and secondary):

- The primary path for data for station STA3 is STA3→GDC1→GDC2,3,4.
- GDC2 knows it is not primary path and waits 5 minutes; GDC2 checks if file was received via GDC1→GDC2,3,4; if yes, discards copy received via STA3⇢GDC2.

HANDLING DATA FILE REVISIONS

Revisions to released data should be documented and forwarded to the user community. Some ideas for user notification:

- When a file is revised by an OC, an email message with a standardized subject (e.g. "STA1 2006 140 revised due to missing data.") is generated and sent out to a mailing list; ACs can use either humans or scripts to read the messages and act accordingly.
  Problem: The OC fails to generate the message.
- When a file is revised by an OC, a COMMENT is placed in the RINEX header with the reason for replacement, and the DCs must extract and log the COMMENT. The log is sent out to a mailing list when appended to.
  Problem: (1) The OC fails to generate the COMMENT. (2) Some additional load on DCs to uncompress and read files. **Note:** This concept certainly should be required on general principle to track revisions to released data.
- When a file is revised by an OC, a separate file with a standardized line explaining the replacement is transmitted with the data. DCs will log the replacement files and send the log to a mailing list when appended to.
  Problem: The OC fails to generate the file.
- Whenever a DC executes an "overwrite" action, it will log the replacement. This action might be combined with the second and third options above to utilize the reason if supplied by the OC, or "Unknown" if not. The log could be emailed to subscribed users and archived at the DCs when modified.

THOUGHTS ON IMPLEMENTATION

More detailed steps will need to be discussed with the IGS DCs in order to implement the concepts detailed in this position paper and poster. These tasks include:

- Gather information on present station/OC data flow
- Define the two upstream DCs for each station/OC and update IGS site logs with this information
- Prepare procedures at all OCs/ODCs for upload to two DCs at higher level
- Prepare procedures at all RDCs and GDCs to receive data from two upload paths and distinguish primary and secondary paths
- Prepare procedures for GDC data equalization
- Test data flow procedures thus verifying operational and backup data flow
- Document data flow and provide information at IGS CB

RECOMMENDATIONS

1. Flow of IGS data from station/OC to a higher-level data center is performed using PUSH only
2. Stations/OCs, ODCs, and RDCs will define primary and secondary data centers to PUSH their data to and will update IGS site logs with this information; IGS CB will create and maintain supplemental material summarizing this data flow
3. Stations/OCs should document replacement of data files and notify the IGS through automated procedures