G11B-0642

Updates to the IGS Data Center Infrastructure

ABSTRACT: The global data centers are key components of the IGS infrastructure, providing access to a full suite of GNSS data, products, and information. These archives are essential tools for a diverse international scientific user community. Since the start of the IGS in 1992, the number of stations, types of station data and derived products has increased substantially, requiring modifications to data center operations. Updates to the archives occur due to new data and products, re-publishing of previously archived data and products, transmission of historic data, and the resulting re-distribution of these data and products from data center to data center. In addition, statistics on the usage of the data and products are important to the data centers as well as data and product providers. This poster will review the status of the IGS data center infrastructure and present ideas, both near-term and long term, for data center equalization and validation of data holdings, modernization of compression schemes, and reporting of usage statistics.

Carey Noll *(Carey.Noll@nasa.gov)*
NASA Goddard Space Flight Center
Code 690, Greenbelt, MD 20771, USA

Patrick Michael *(Patrick.Michael@nasa.gov)*
Catholic University of America
NASA Goddard Space Flight Center
Code 690, Greenbelt, MD 20771, USA

Michael Schmidt *(MSchmidt@nrcan.gc.ca)*
Geological Survey of Canada, Natural Resources Canada
P.O. Box 6000, Sidney BC, V8L 4B2, CANADA

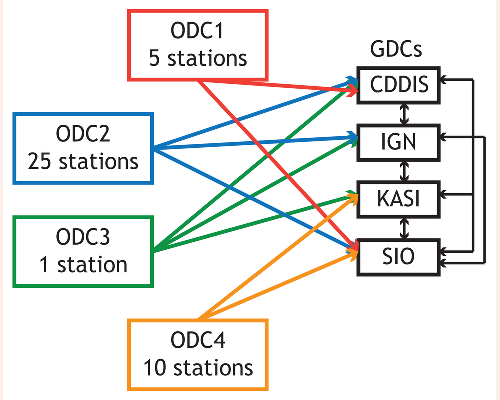
Yuan Lu *(yuan.lu@nrcan.gc.ca)*
Geological Survey of Canada, Natural Resources Canada
P.O. Box 6000, Sidney BC, V8L 4B2, CANADA

Natural Resources Canada   Ressources naturelles Canada

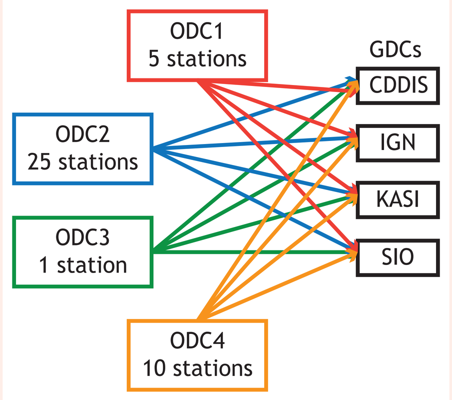
Improvements to IGS GDC Archive Contents and Procedures

Recommendations for Improvements to IGS Data Dissemination:

The main goal of the IGS Global Data Centers (GDCs) is to ensure reliable, consistent, ready user access to the full suite of GNSS data and products. In order to satisfy this goal, the IGS GDCs need to review and update their procedures for populating and maintaining their archives.



Current situation:
- Individual Station Operators (SOs) and Operational Data Centers (ODCs) provide data to one or more "upper level" Regional Data Centers (RDCs) or Global Data Centers (GDCs)
- RDCs/GDCs currently equalize selected data from selected sites (Note: There is some equalization of data but there is no rigorous mirroring of primary data submissions or subsequent re-submissions)
- Result: Inhomogeneous data set (inconsistent files, bad data, and errors in data) in GDCs



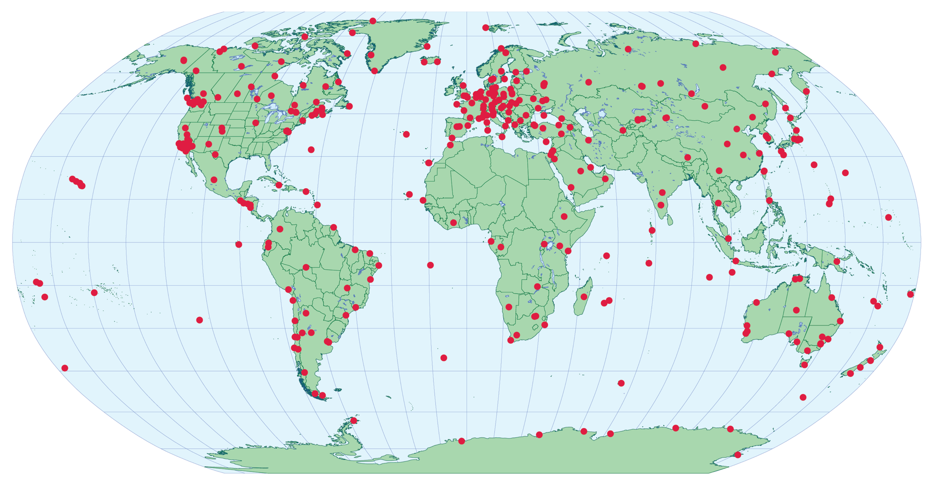
Proposed solution:
- Data should be pushed directly from source (SO/ODC) to archive centers (RDCs/GDCs)
  - Ensures responsibility for data remains with source (SO/ODC)
  - Data distributed directly by the SO/ODC to ALL IGS GDCs
  - Responsibility remains with the SO/ODC to ensure that data are distributed to a GDC after GDC outage, etc.
- One source for data distribution for each IGS site
- Back up source encouraged where possible, but only one active source at any one time

Advantages:
- Implies simplified data flow
- Ensures responsibility for data remains with SO/ODC
- Ensures data are distributed/mirrored across all GDCs (primary submission and subsequent re-submissions)
- Ensures global data availability should individual GDCs temporarily go off-line
- Allows for timely notification of data re-submissions

Recommendations for IGS GDC Data Holdings and Handling Data Resubmissions:

Current situation:
- GDCs provide access to different sets of IGS sites => Users must "shop" multiple GDCs to retrieve required data
- GDCs not synchronized and therefore not necessarily holding most current data => Inhomogeneous data set in GDCs and RDCs

Proposed solution:
- GDCs archive data from, at a minimum, all sites in IGS network
- GDCs provide on-line resources to summarize data holdings
- GDCs post data files as received without re-formatting or changing data files in any way
- GDCs verify integrity of data holdings by comparing file check sums, flagging any differences and updating archive(s) as required
- SO/ODCs distribute updated data files to all GDC's
- SO/ODC issues advisory immediately upon data resubmissions using a prescribed format
- GDCs extract summary info from advisory email to update resubmission file at each data center, one line per message



The IGS network currently consists of 420 global distributed sites (as of November 2010).

Advantages:
- Ensures that data are consistent among GDCs
- Ensures replacement data are distributed to all GDCs
- Ensures users can easily get data from any GDC for any IGS site
- Provides redundant data availability including resubmitted data files

```
Author: Michael Schmidt

************************************************************************
                          NRCan/PGC


SITE ID:              ALBH
DOMES No.:            40129M003
MONUMENT No.          927000
LOCATION:             Albert Head, British Columbia, CANADA
START DATE:           2010-10-30
END DATE:             2010-10-31
REASON:               Invalid header records
```

Example of a data replacement email.

```
IGS Tracking Network Status for 10-May-10 100510 10130    GPS Week 1583 Day 2         As of date:  May 17 2010 10:27:21

     Dly  No.   No.   Pts.     Avg. Avg. Pos.  No.                                      Ant.                         Marker            RINEX    Dly
Site (H)  Exp.  Obs.  Del.  %  MP1  MP2  Diff Slps V    Receiver Type        Antenna Type     Height     Marker Name        Number   Type  Version  (M)
---- --- ----- ----- ----- --- ---- ---- ---- ---- - -------------------- -------------------- ------ -------------------- --------- --- --------- -----
abmf  10 25780 25640   136  99 0.41 0.51 0.04   23 1 TRIMBLE NETR5        TRM55971.00     NONE 0.0000 ABMF               97103M001  M    2.11      630
abpo   1 26268 25800     0  98 0.35 0.37 0.09   23 1 ASHTECH UZ-12        ASH701945G_M    SCIT 0.0083 ABPO               33302M001  G    2.11       14
ade1  20 25447 25133     0  98 0.43 0.41 0.04    4 1 ASHTECH Z-XII3       ASH700936B_M    SNOW 0.0000 ade1               50109S001  G    2.11     1253
ade2  20 25447 25071     1  98 0.42 0.41 0.04    5 1 ASHTECH Z-XII3       ASH700936B_M    SNOW 0.0000 ade2               50109S001  G    2.11     1253
adis  24 28500 22796   672  79 0.53 0.58 0.03   22 1 JPS LEGACY           TRM29659.00     NONE 0.0010 ADIS               31502M001  M    2        1490
aira   1 24806 22489   608  90 0.35 0.48 0.03   30 1 TRIMBLE 5700         TRM29659.00     DOME 0.0000 AIRA               21742S001  G    2.11      111
ajac   8 23674 23644     0  99 0.19 0.16 0.04    2 1 LEICA GRX1200GGPRO   LEIAT504GG      NONE 0.0000 AJAC               10077M005  M    2         511
albh   1 25563 25235     0  98 0.22 0.26 0.03    3 1 AOA BENCHMARK ACT    AOAD/M_T        SCIS 0.1000 albh  WCDA-ACP   927 40129M003  G    2.11        8
algo   1 24839 24730     0  99 0.22 0.23 0.07    1 1 AOA BENCHMARK ACT    AOAD/M_T        NONE 0.1000 ALGO CACS-ACP   8831 40104M002  G    2.11       11
.
.
.
zamb
zeck   9 24586 24311    32  98 0.49 0.51 0.07    5 1 ASHTECH Z-XII3       ASH700936D_M    SNOW 0.0450 ZECK               12351M001  G    2.10      570
zim2   1 24423 24342     2  99 0.29 0.34 0.06    4 1 TRIMBLE NETR5        TRM59800.00     NONE 0.0000 ZIM2               14001M008  M    2.11       12
zimj   9 24432 24069    48  98 0.51 0.55 0.03    3 1 JPS LEGACY           JPSREGANT_SD_E  NONE 0.0770 ZIMJ               14001M006  M    2         541
zimm   1 24423 24177   197  98 0.38 0.34 0.07   47 1 TRIMBLE NETRS        TRM29659.00     NONE 0.0000 ZIMM               14001M004  G    2.1         8
zwe2   9 25076 24785     0  98 0.28 0.28 0.07   39 1 SEPT POLARX2         TPSCR3_GGD      NONE 0.1300 ZWE2               12330M003  M    2.11      541

Program:  QC 2009Mar23 by UNAVCO run with elevation angle cutoff of 10 degrees
    Field       Size   Type                          Explanation
-------------   ----  ------  -------------------------------------------------------
Site              4   char    Site name
Dly (H)           3   number  Delivery delay in hours
No. Exp.          5   number  Total number of observations expected
No. Obs.          5   number  Total number of observations in file
Pts. Del.         5   number  Total number of points deleted
%                 3   number  Data collection percentage
Avg. MP1          4   number  Average L1 multipath (rounded to two decimal places)
Avg. MP2          4   number  Average L2 multipath (rounded to two decimal places)
Pos. Diff         4   number  RINEX vs QC point position difference (Km)
No. Slps          4   number  Number of detected slips
V                 1   number  Version of the data file (set to 1 for initial delivery)
Receiver Type    20   char    Type of GPS receiver from RINEX header
Antenna Type     20   char    Type of GPS antenna from RINEX header
Ant. Height       6   number  Height of antenna from RINEX header
Marker Name      20   char    Marker name from RINEX header
Marker Number    10   char    Marker DOMES number from RINEX header
Type              1   char    GPS type
RINEX Version     7   char    RINEX Version number
Dly (M)           3   number  Delivery delay in minutes
```

An example of a file maintained by the CDDIS for summarizing daily GNSS data holdings. This archive status file of CDDIS GNSS data holdings reflects the timeliness of the data delivered as well as statistics on number of data points, cycle slips, and multipath. The user community can obtain a snapshot of data availability and quality by viewing the contents of such a summary file as shown in the figure. The CDDIS staff has recently enhanced the format of these files as requested by the Infrastructure Committee. Software for generating the files is available to other data centers.

Update to IGS Data Center Compression Scheme

Recommendations for Updates to IGS Compression Scheme:

Current situation:
- IGS infrastructure (e.g., stations, DCs, ACs, etc.) currently uses a two stage compression scheme for RINEX observation data: Hatanaka compacting followed by UNIX compress (Z)
- Z compress is not as efficient as other compression schemes
- DCs compress once (or twice); users decompress often; therefore, efficient size compression is more important from a DC perspective

Proposed solution:
- Change compression tool used by the IGS infrastructure to a more effective, commonly available compression tool
- Three options are considered here: gzip, bzip2, and xz
- All are widely distributed across multiple platforms and operating systems

Advantages:
- All proposed compression software provide an improved factor of compression over Z; some require more memory and CPU resources
- All provide checksum capture
- A DC recompression process will allow DCs to QC their archives, capture checksum information for ensuring intact file transfers (however, a cryptographic checksum is preferred for performing DC content intercomparisons and data rationalization)
- Compression of high-rate files (1Hz and higher) further improves by using one of these new compression tools

Implications:
- A new compression scheme should be implemented across the IGS infrastructure including SOs/ODCs, RDCs, GDCs, ACs, etc.)
- Implementation must be coordinated across the IGS infrastructure as well as with manufacturers, users, etc.; an overlap period will be required

Recommendations:
- Propose new compression scheme and survey IGS infrastructure
- Coordinate changes within IGS but also with general user community, receiver manufacturers, etc.
- Develop timeline for implementation
- Coordinate changes to allow DCs to utilize more efficient compression on historic archives
- Begin use of new compression scheme as part of overall data center archive content evaluation in preparation for IGS reanalysis 2

Analysis of Compression Testing

Tests of commonly available compression schemes (Z, gzip, bzip2, and xz) were performed on a CDDIS server running the Linux OS; the results are shown in Charts 1-5.

Tested Compression Schemes/Levels

Compression Scheme	Available Levels	Default Level
Unix (Z)	1	1
gzip	1-9	6
bzip2	1-9	9
xz	1-9	6

Chart 1 compares all tested compression schemes on Hatanaka-compacted and uncompacted RINEX observation data. Each file was then compressed (using default levels for each of the four tested schemes) and the average percent of compression was graphed. It is clear that using Hatanaka compression combined with any one of the proposed compression tools is more effective than not using the Hatanaka software. In other words, there is no common compression tool available that will provide better compression than using Hatanaka + another compression tool.

Charts 2-4 summarize tests on one day of daily 30-second GNSS data. A total of 1,586 files were compressed and uncompressed using each compression scheme and available compression level. The selected directory included compact (Hatanaka) and un-compacted RINEX observation, navigation, meteorological data, and teqc summary files. Chart 2 shows the compression factor as a percentage of the original file for each compression scheme at all available levels. The higher the level, the more compressed the resulting file. The larger the bar in this chart the better the compression; as can be seen, xz results in a better compression factor overall. Bzip2 provides improved compression over gzip and is rather constant on this data set regardless of the compression level.

However, a higher level of compression can come at the price of the time required to compress and decompress the files. As shown in Chart 3, xz compression time increases at higher compression levels; bzip2 compression time is relatively constant and gzip increases significantly using higher levels.

Chart 4 presents the time required for decompression of the files; compression level has little effect on the time required for decompression. Here bzip2 performance is worse than the other two compression schemes. Taken together, Charts 2-4 show that xz level 3, gzip level 6, and bzip2 level 9 provide the best compression in size with reasonable computing time.

Chart 5 summarizes compression for different types of GNSS data, daily 30-second files, hourly 30-second files, and sub-hourly, high-rate 1-second files. The compression levels used for this chart are gzip level 6 (default), bzip2 level 9 (default), and xz level 3. Xz consistently provides the best compression for all types of GNSS data.

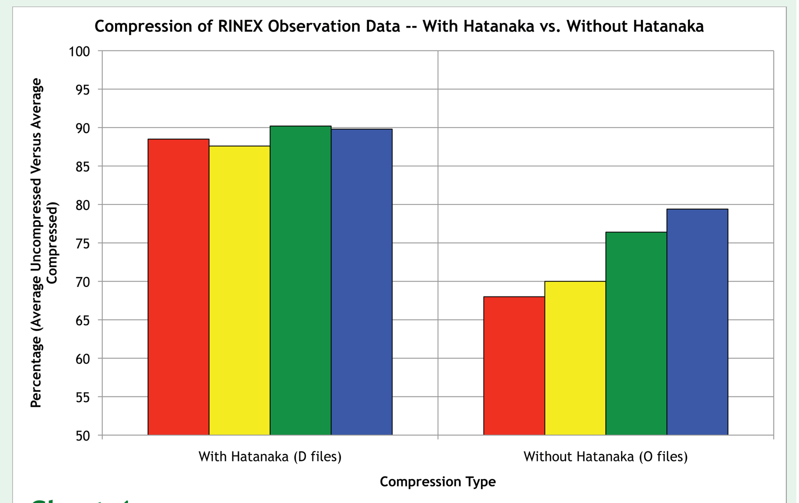


Chart 1

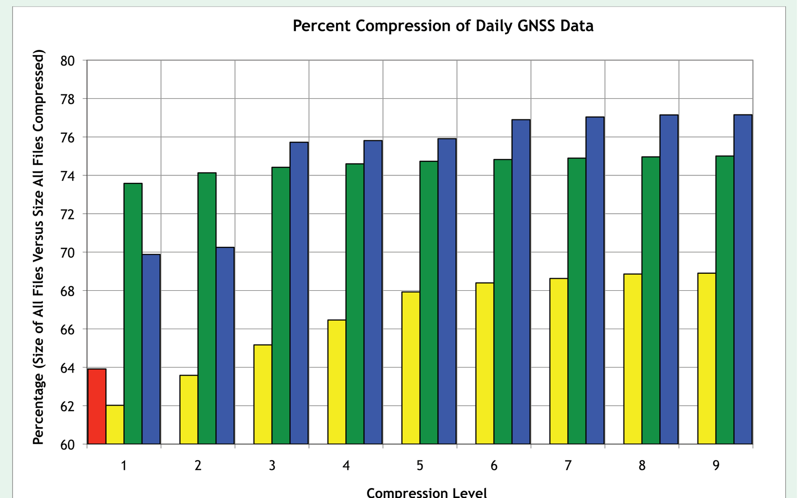


Chart 2

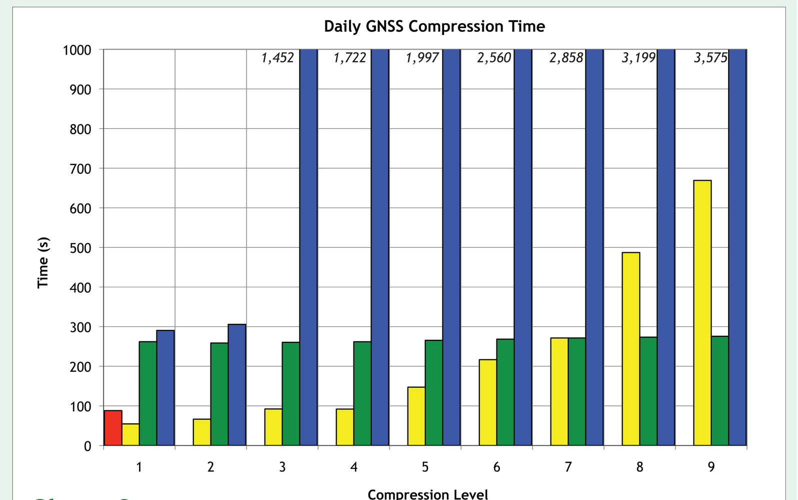


Chart 3

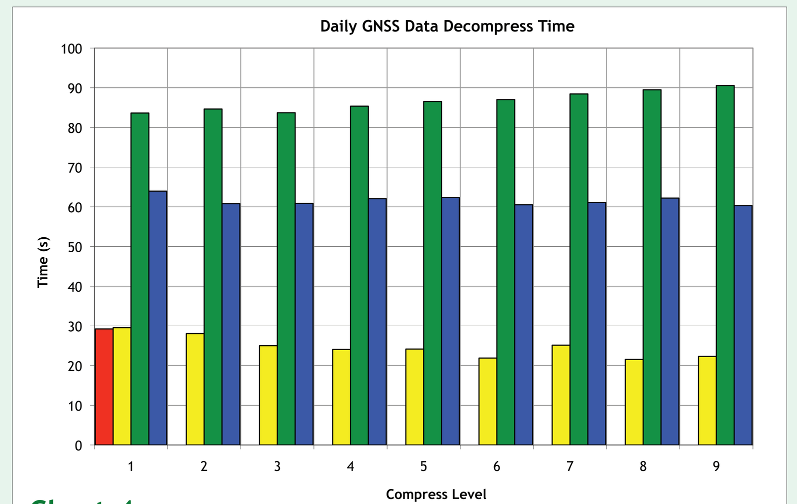


Chart 4

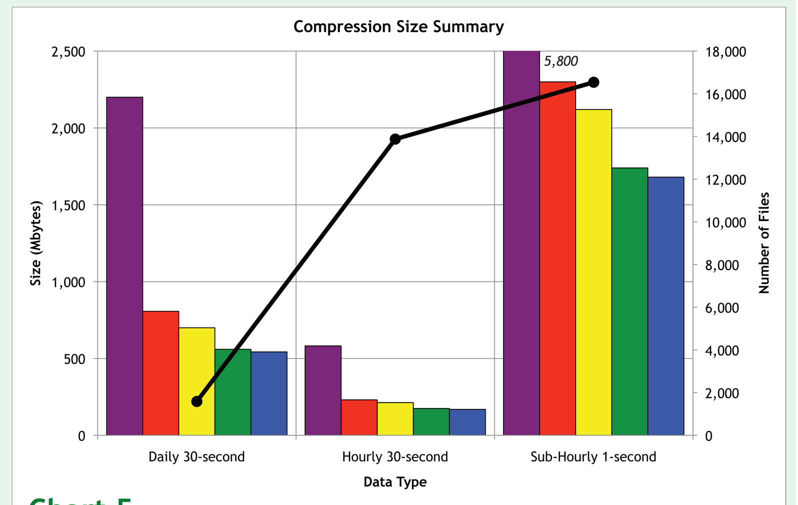


Chart 5

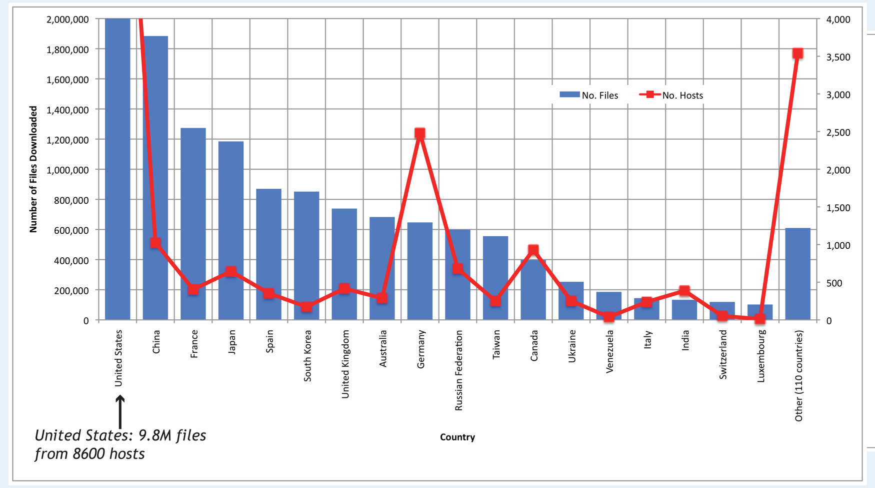
Capturing IGS GDC Usage Metrics

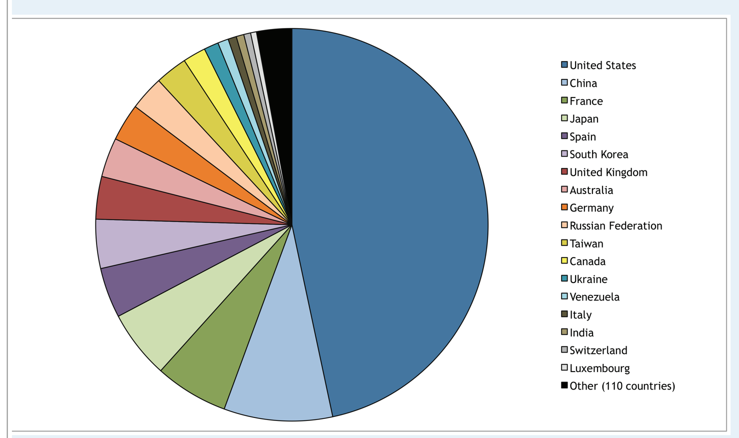
Recommendations for Monitoring GDC Archive Usage

Current situation:
- Many organizations provide data and derived products to IGS data centers
- IGS should periodically evaluate the effectiveness of DC operations and the scope of their content distribution

Proposed solution:
- Develop process to capture and analyze distribution metrics at IGS GDCs
- Report results by data/product type, amount of data retrieved (number/volume of files), and destination

Advantages:
- Organizations can use metrics to understand the extent of the utilization of their contributions to the IGS

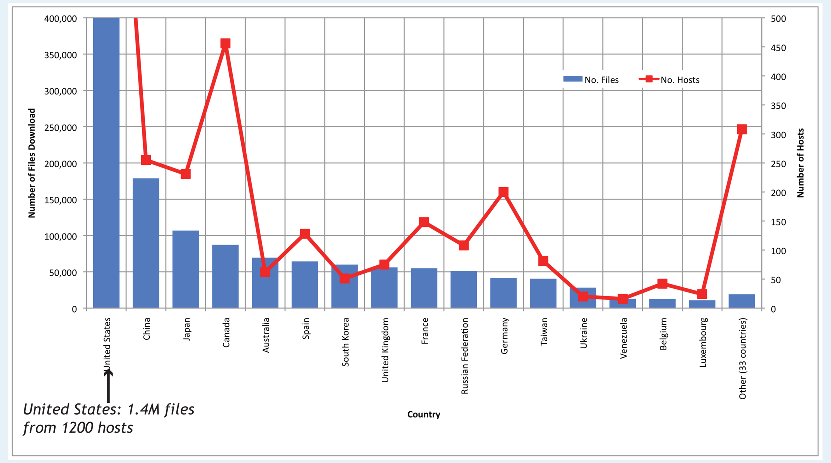


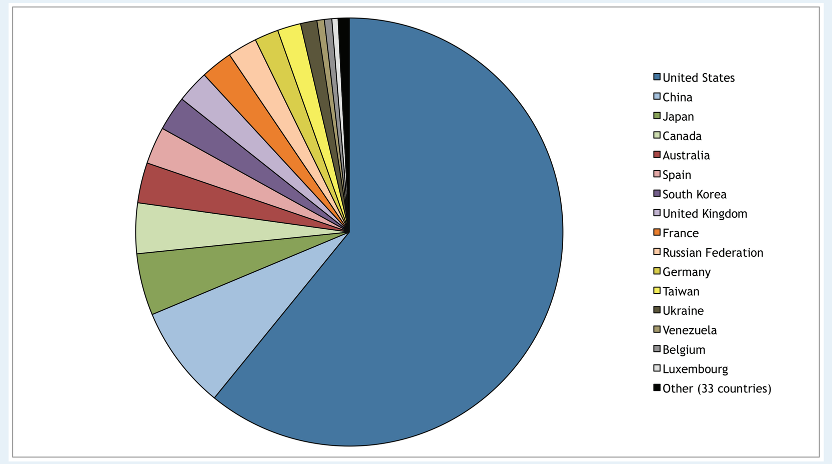


As an example, the CDDIS has loaded the server's ftp distribution logs into a MySQL database. Queries were constructed to determine the number of file downloads based on a particular data or product type. Using freely available IP host information, the location information (i.e., country) of downloaded files can be determined. The first set of charts (above) shows the total number of daily 30-second GNSS files downloaded from June through November 2010, sorted by country. The second set of charts presents the downloads for the same time period, for 30-second GNSS files from NRCan sites only.

NRCan Daily Data Downloads from CDDIS
June-November, 2010

Country	No. Files	No. Hosts
Argentina	8	3
Australia	69521	62
Austria	1	1
Belgium	12810	42
Brazil	20	7
Brunei Darussalam	24	8
Canada	87249	456
Chile	1	1
China	178820	255
Colombia	200	7
Czech Republic	2	2
Denmark	16	2
Egypt	1	1
Finland	8	2
France	54936	148
Germany	41300	200
Greece	5	3
Hong Kong	202	13
Iceland	68	7
India	933	44
Indonesia	5110	13
Iran	333	7
Italy	3393	35
Japan	106812	231
Kenya	5	4
Luxembourg	10828	24
Malaysia	317	14
Netherlands	132	5
New Zealand	6	1
Norway	2252	25
Panama	1	1
Poland	205	8
Portugal	3556	37
Romania	4	4
Russian Federation	51083	108
Serbia	1	1
Singapore	565	7
South Africa	2	1
South Korea	59971	51
Spain	64378	128
Sweden	5	5
Switzerland	1694	29
Taiwan	40510	81
Turkey	4	3
Ukraine	28195	20
United Kingdom	56121	75
United States	1392326	1192
Venezuela	12887	16
Viet Nam	25	7





Concluding Remarks

This poster has presented ideas for modifying IGS data center operations to improve efficiency and ensure a robust data and product archive infrastructure for the user community. However, any ideas shown here will require further research in order to develop concrete recommendations and schedules on ways forward. Any major modifications to data center operations must be introduced under close coordination with the IGS Infrastructure Committee, the IGS Data Center Working Group (DCWG), the IGS data centers, receiver manufacturers and, of course, with the user community. The authors would like to encourage feedback on the proposals presented here. Please send any comments, suggestions, or questions to Carey Noll (*Carey.Noll@nasa.gov*).

12/07/2010