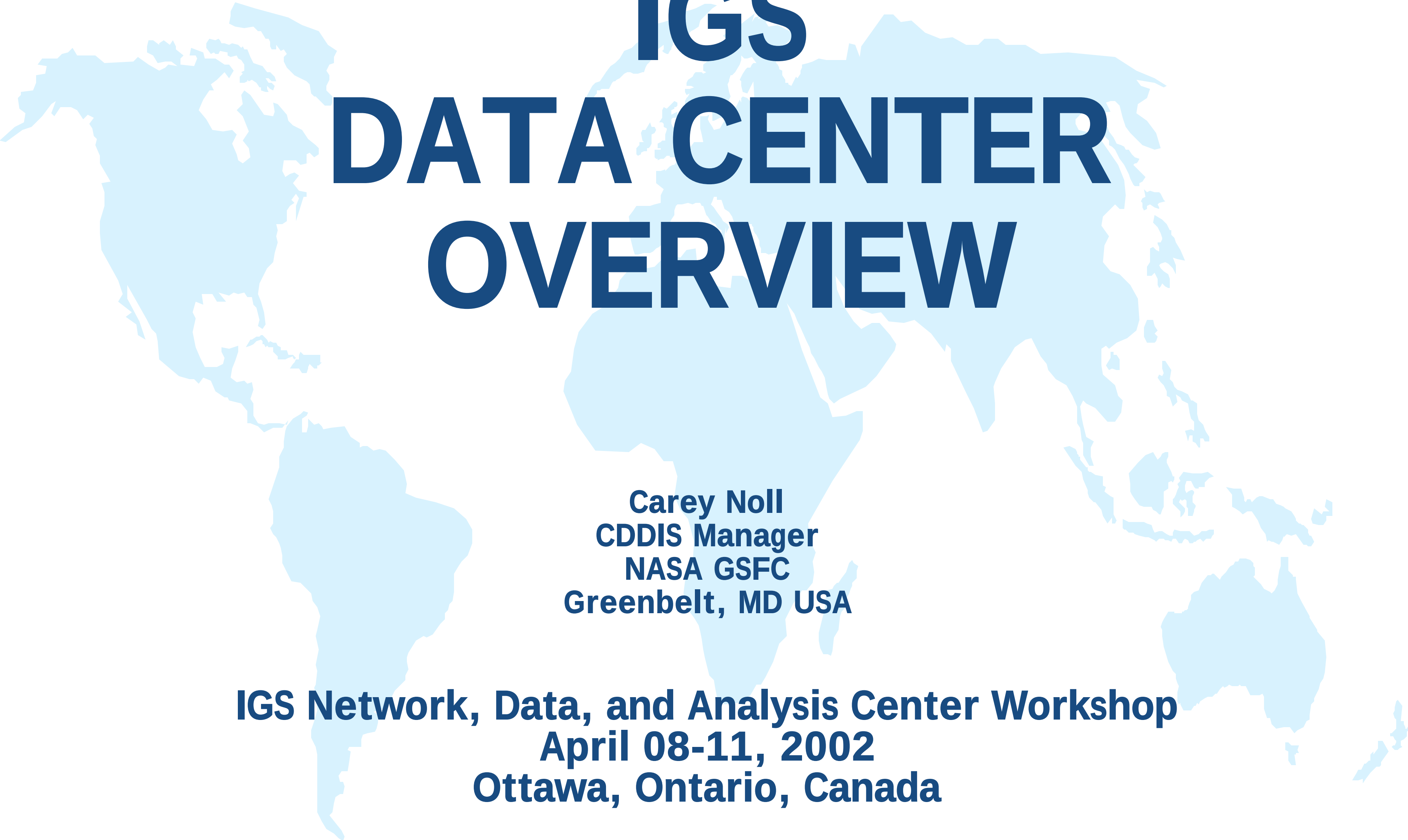
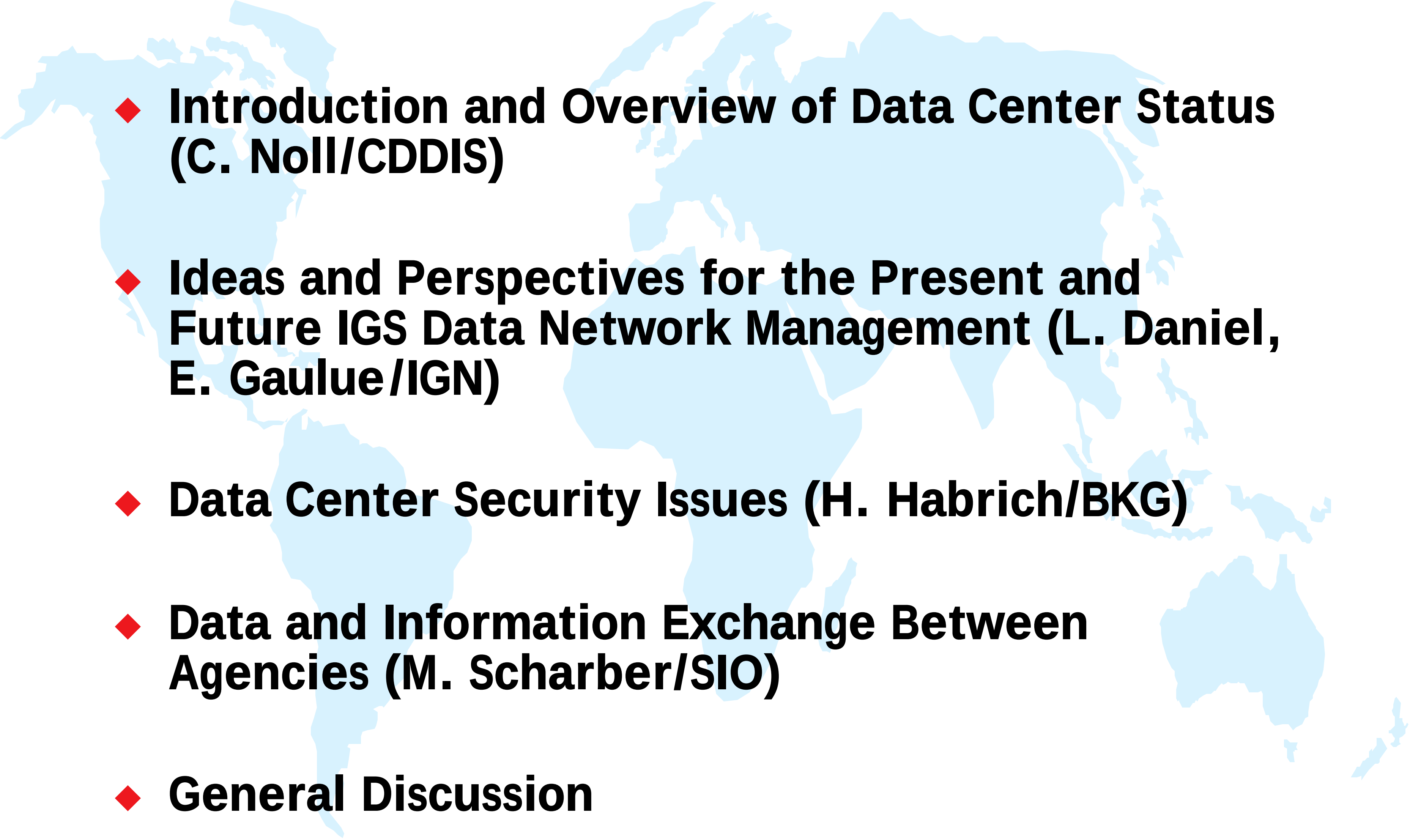


IGS DATA CENTER OVERVIEW

Carey Noll
CDDIS Manager
NASA GSFC
Greenbelt, MD USA

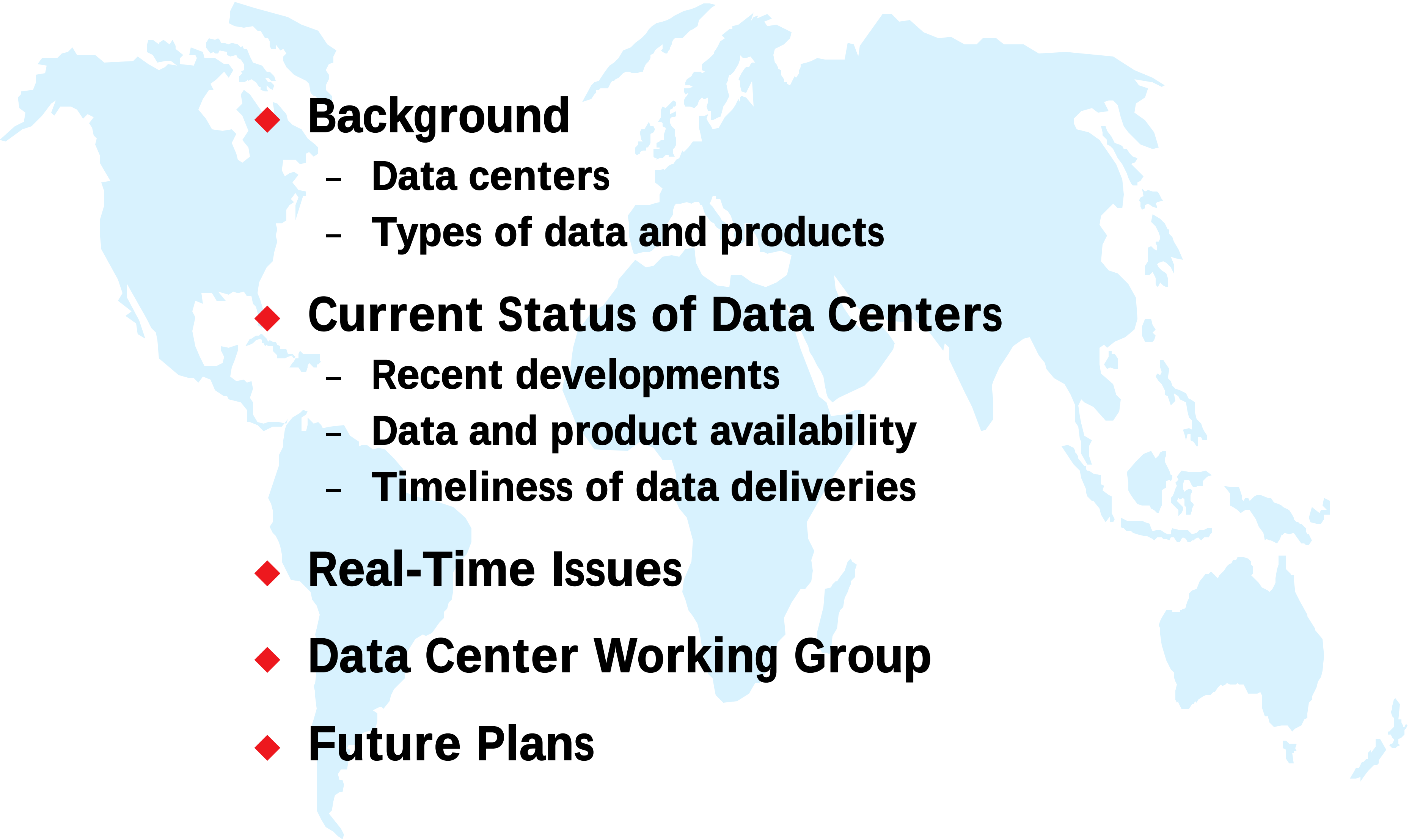
IGS Network, Data, and Analysis Center Workshop
April 08-11, 2002
Ottawa, Ontario, Canada

DATA CENTER SESSION



- Introduction and Overview of Data Center Status (C. Noll/CDDIS)
- Ideas and Perspectives for the Present and Future IGS Data Network Management (L. Daniel, E. Gaulue/IGN)
- Data Center Security Issues (H. Habrich/BKG)
- Data and Information Exchange Between Agencies (M. Scharber/SIO)
- General Discussion

OVERVIEW

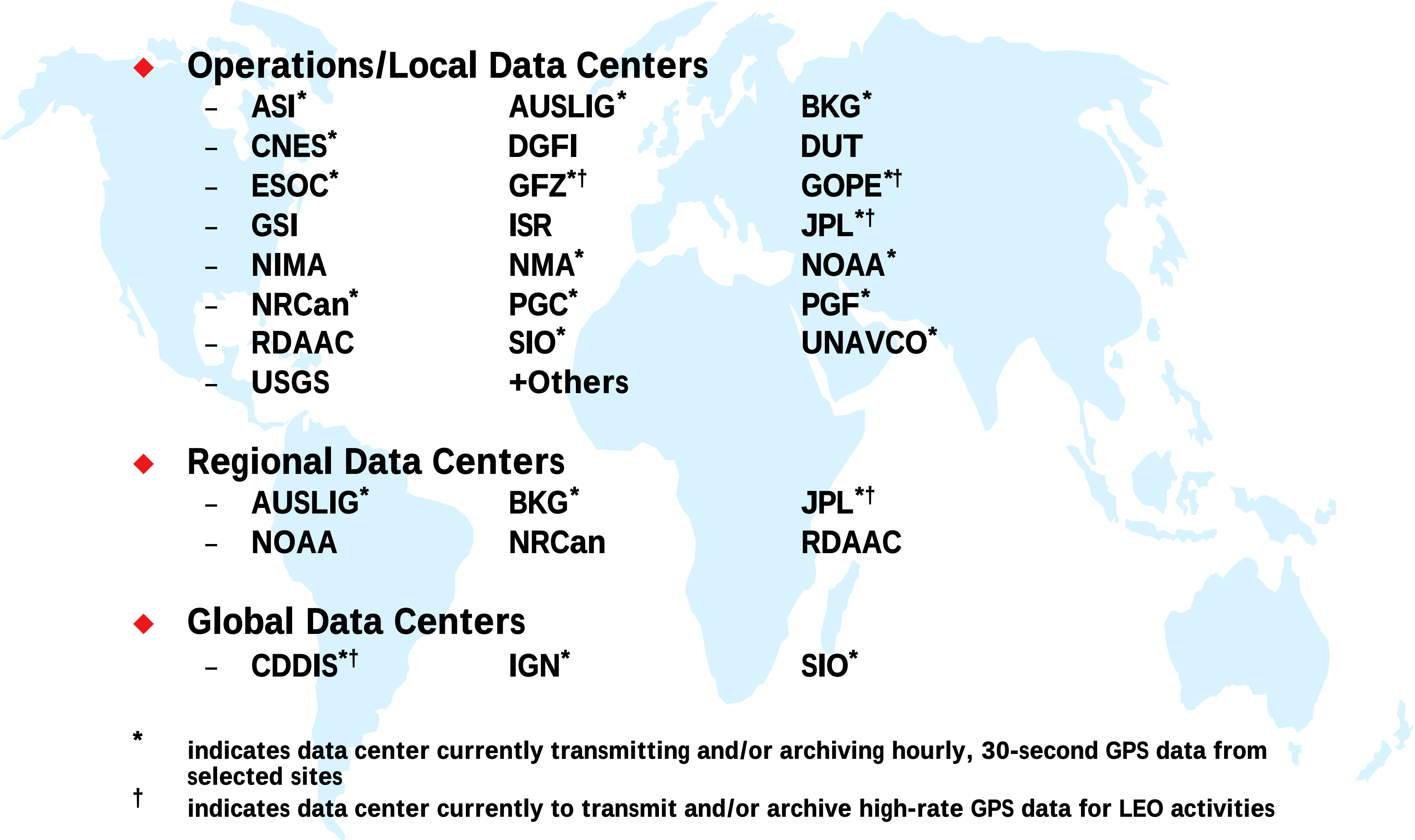


- Background
  - Data centers
  - Types of data and products
- Current Status of Data Centers
  - Recent developments
  - Data and product availability
  - Timeliness of data deliveries
- Real-Time Issues
- Data Center Working Group
- Future Plans

BACKGROUND

- IGS has hierarchy of data centers
  - Operational or local data centers
  - Regional data centers
  - Global data centers
- Operational or local data centers (OCs or LDCs) interface to receiver, download, and QC data
- Regional data centers (RDCs) gather data from OCs and maintain archive of stations in a particular region
- Global data centers (GDCs):
  - Receive/retrieve data (IGS global sites, at a minimum) from OCs and RDCs
  - Equalize data holdings for key sites
  - Provide archive of data and products for ACs and user community

IGS DATA CENTERS



- **Operations/Local Data Centers**
  - ASI* AUSLIG* BKG*
  - CNES* DGFI DUT
  - ESOC* GFZ*† GOPE*†
  - GSI ISR JPL*†
  - NIMA NMA* NOAA*
  - NRCan* PGC* PGF*
  - RDAAC SIO* UNAVCO*
  - USGS +Others

- **Regional Data Centers**
  - AUSLIG* BKG* JPL*†
  - NOAA NRCan RDAAC

- **Global Data Centers**
  - CDDIS*† IGN* SIO*

* indicates data center currently transmitting and/or archiving hourly, 30-second GPS data from selected sites

† indicates data center currently to transmit and/or archive high-rate GPS data for LEO activities

GPS (AND GLONASS) DATA SETS

- GPS (and GLONASS) data (daily files)
  - 30-second sampling
  - ~300 GPS stations (~50 GLONASS) at CBIS
  - Average 2-hour delay
  - File types:
    - O (RINEX observation data)
    - D (RINEX observation data, Hatanaka compression)
    - M (RINEX meteorological data)
    - N (RINEX broadcast ephemeris data)
    - S (output from teqc)
- Near real-time GPS (and GLONASS) data (hourly files)
  - 30-second sampling
  - ~100 regularly submitting
  - Average 5-15 minute delay
  - Retained for three days
  - Since mid 1998
- High-rate GPS data
  - 1-second sampling
  - 39 stations currently (from JPL, GFZ, ASI, and GOPE)
  - Data in 15 minute files (*ssssdddhmi.yyt.Z*)
  - Since mid 2001

IGS PRODUCTS

- **Orbit, clock, ERP products**
  - Seven ACs
  - Since GPS week 0649
  - Weekly precise combination, daily predicted and rapid combinations from AC Coordinator (AIUB)
- **SINEX products (station positions)**
  - Standard IGS and working group products
  - Seven ACs, two GNAACs, three RNAACs (currently)
  - Since ~GPS week 0840
  - Weekly combination from Reference Frame Coordinator (NRCan)
- **Ionosphere products (global ionosphere maps of total electron content, TEC)**
  - Working group product
  - IONEX format
  - Daily files
  - Five AACs
  - Since June 1998
- **Troposphere products (combined zenith path delay, ZPD)**
  - Working group product
  - Seven AACs
  - Weekly files
  - Weekly combination (from GFZ)
  - Since January 1997

IGS GLOBAL DATA CENTER HOLDINGS

Data Type	CDDIS	IGN	SIO
Data			
GPS daily (D format)*	X	X	X
GPS daily (O format)	X		X
GPS hourly (30-second)*	X	X	X
GPS hourly (high-rate)	X		
GLONASS daily (D)[†]	X	X	
GLONASS daily (O)[†] format)[†]	X		
Products format)			
Orbits, etc.*	X	X	X
SINEX*	X	X	X
Troposphere[†]	X	X	X
IONEX[†]	X	X	

* Official IGS data set/product

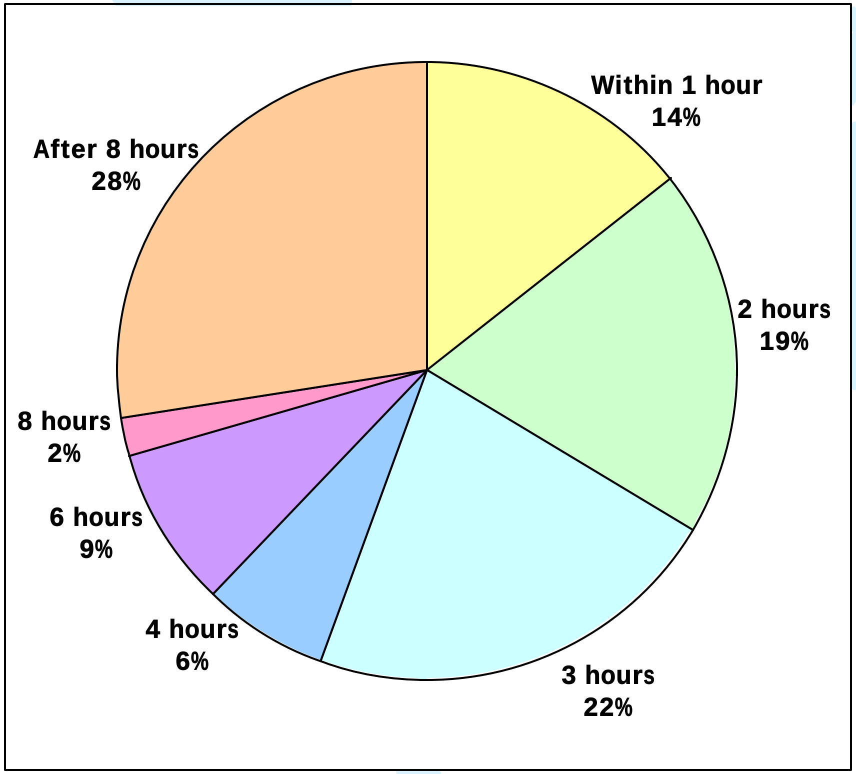
[†] Pilot project/working group data set/product

RECENT DEVELOPMENTS

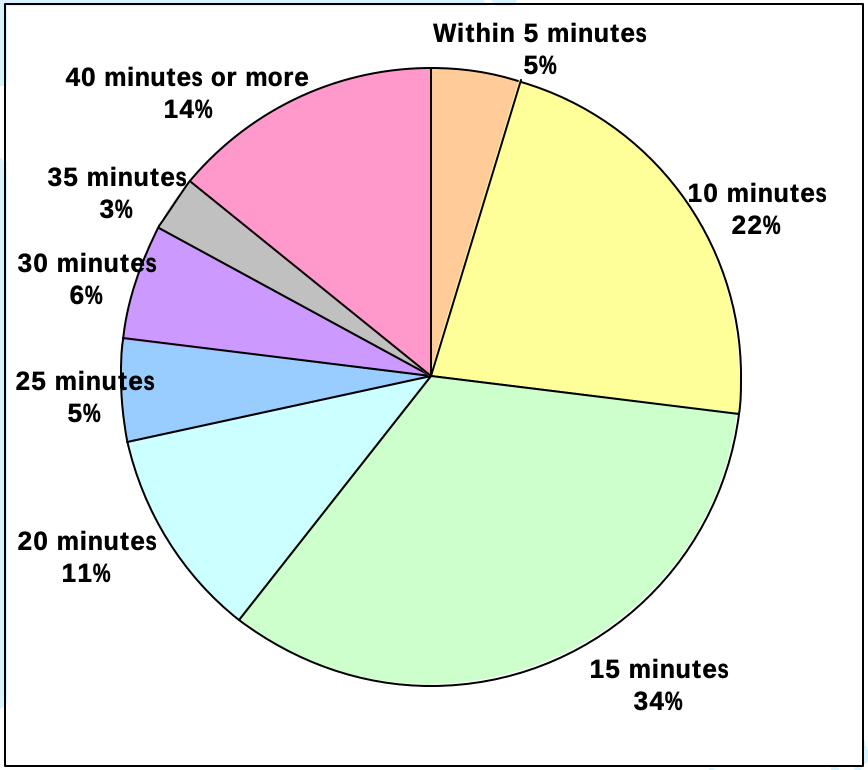
- **General:**
  - Integration of GLONASS data into IGS data stream to begin in 2002 in support of IGLOS-PP
  - Approximately 60% of daily data files delivered within three hours
  - Approximately 60% of hourly 30-second data files delivered within fifteen minutes
- **IGN:**
  - Begun “revitalization” of global data center in Jun-01
- **SIO:**
  - Archiving observation data in compact RINEX as of Apr-2001
  - Began archive of hourly, 30-second RINEX data in May-2001
- **CDDIS:**
  - Began archive of 1-second RINEX data in 15-minute files in May-2001 in support of LEO-PP; also archiving analysis products for test campaign (13 received thus far)
  - Archive of LEO receiver data (CHAMP and SAC-C) since Jan-2002 in compact RINEX (V2.0) format; retrieved from GENESIS archive
  - Supported HIRAC/SolarMax campaign in Apr-2001; archived 13 Gbytes of high-rate data from 104 sites

DATA LATENCY (CDDIS)

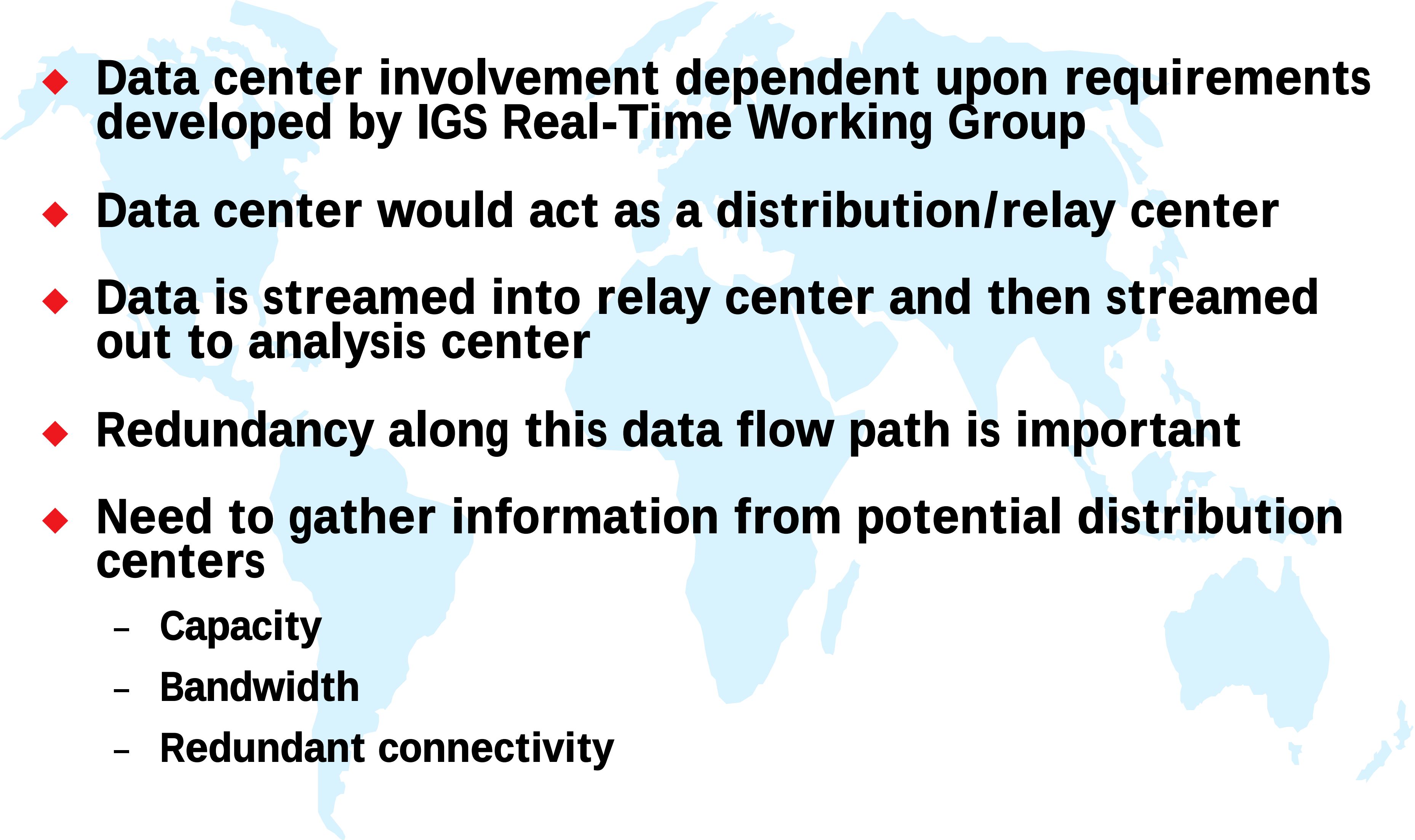
Daily Files



Hourly (30-Second) Files



REAL-TIME ISSUES



- Data center involvement dependent upon requirements developed by IGS Real-Time Working Group
- Data center would act as a distribution/relay center
- Data is streamed into relay center and then streamed out to analysis center
- Redundancy along this data flow path is important
- Need to gather information from potential distribution centers
  - Capacity
  - Bandwidth
  - Redundant connectivity

IGS DATA CENTER WORKING GROUP

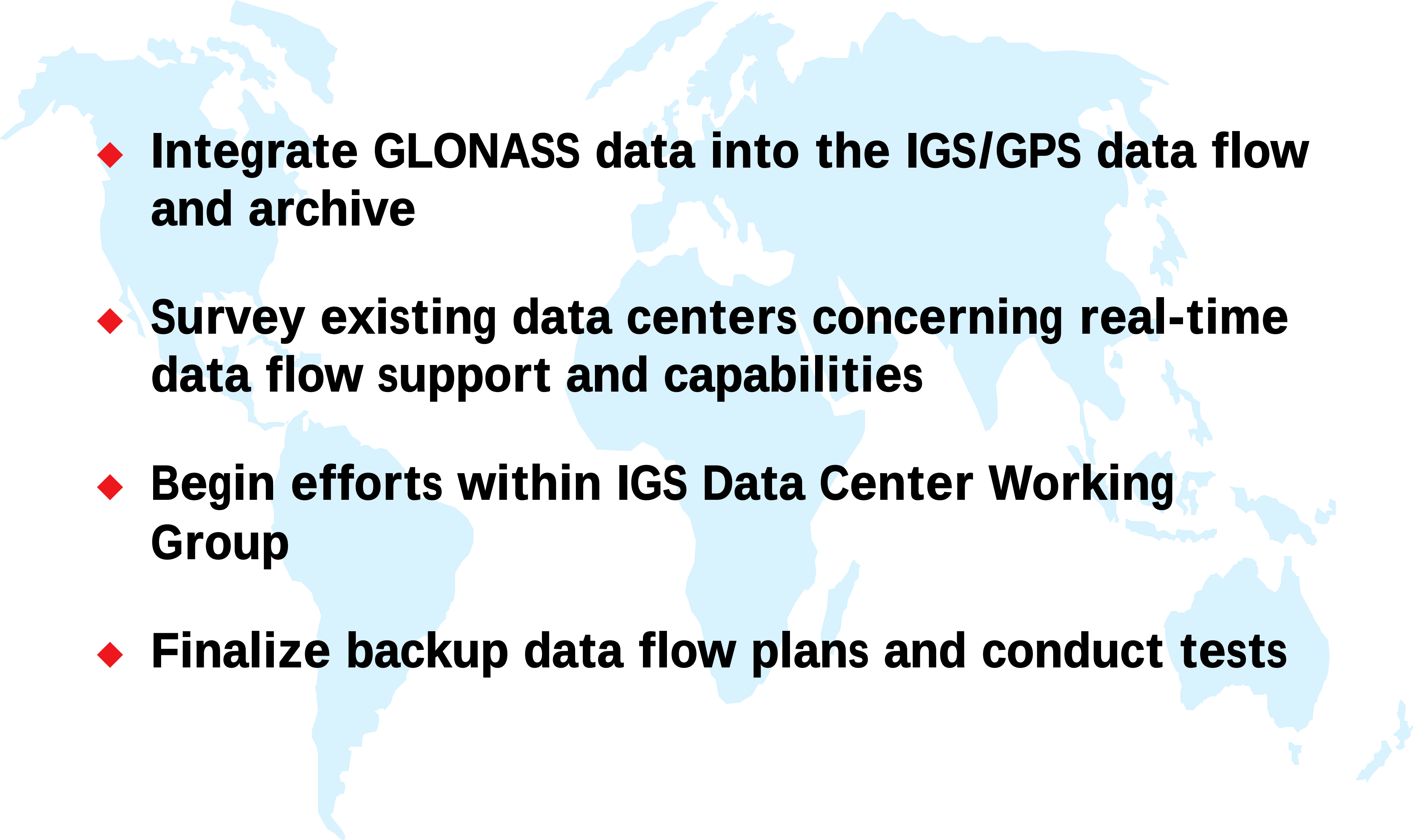
- Direction of IGS has changed; time to re-visit data center requirements
- Many projects and working groups have been created since the inception of the IGS
- Address issues and challenges at the IGS data centers (all levels) in order to improve service for users both internal and external to the IGS
  - Effective data flow
  - Backup of the operational data flow
  - Security
  - Consistency of data holdings among centers
  - Timely archive and dissemination of data in a real-time scenario
- Membership consists of main IGS data center contacts plus other experts
- Immediate plan:
  - Contact potential members (done)
  - Develop charter (done)
  - Present draft charter at next IGS Governing Board meeting (April 11) for approval

IGS DCWG ACTIVITIES

- **Near-term activities:**
  - Ensure data center information at IGS CB is current
  - Create web site for working group
  - Establish DCWG email exploder
  - Develop a user survey form and compile results
  - Develop a topology of the current IGS data flow
  - Develop data flow redundancy procedures at key data centers
  - Develop procedures for identifying and notification of problem/replacement data sets

- **Long-term activities:**
  - Study of web-based enhancements to data center information
  - Interface with real-time working group and assess requirements
  - Study seamless archive and how effort can be utilized by the IGS and provide a benefit to the data centers

FUTURE PLANS



- Integrate GLONASS data into the IGS/GPS data flow and archive
- Survey existing data centers concerning real-time data flow support and capabilities
- Begin efforts within IGS Data Center Working Group
- Finalize backup data flow plans and conduct tests

IGS Workshop 2002
Ottawa, Canada

Data Centers, ideas and issues

Loïc Daniel, IGN, France Edouard Gaulué, IGN, France

In this paper, I'll try to address some of the points that seem of importance to me, this will inevitably be biased by what we see from our position at IGN as a Global Data Center (GDC). I won't try to encompass all the aspects to be dealt with in this session or about data centers in general. This paper is intended mostly as a repository of present or potential problems that I have identified and an incentive to discuss and find ideas among the community.

1. Present situation

First I'll describe the raw characteristics of the data management at IGS data centers the way we do it presently, I'll try to summarize our activities and evaluate the impact in terms of computer and network loads.

Basic data management

These are the tasks that should be operated on the fly, with a minimal additional delay induced at each step. The data moving operations rely upon a layered structure of Data Centers. The goal is to propagate observations of the stations and products from analysis centers to final places where they are easily available to everyone. Most if not all of the files should end at the GDCs.

Two types of data flows may be considered:

1) file transfers between data or analysis centers as part of the general scheme defined by IGS in order to ensure the best performance in data availability to analysis centers and users. This is the "IGS data flow". It represents the main part of the day to day activities of the data centers. The objective is to put the data and product files at places where they can be downloaded by users. This is the part of the data flow that can be controlled, optimized and supervised because all actions are triggered by an identified component of the data network and following a predefined time table.

2) file transfers initiated by users of the service, this is the "users data flow". Users can be IGS analysis centers and any other kind of user of the IGS. This is much less controlled by data centers, the files are provided for download on an ftp server and users get them as they want without registering or making a special agreement whatsoever. In some cases, a user will issue a request for offline data and the data center will have to restore data and provide access but this tends to be the exception since most data centers (at least the global ones) strive for putting online all the files created since the beginning of the service.

Data and product files are managed by a network of operational, regional and global data centers following current guidelines and specifications. The main constraints are timeliness and completeness, the files should be propagated throughout the network as quickly as possible and no file should be missed. The highest timeliness constraint is found in the Near Real Time (NRT) data flow : these are hourly data files that need to be distributed within minutes after the last observation. The secondary constraints are quality check and reliability. The reliability constraint means that no file should be lost after having reached the data center, this can be ensured by short term backups, redundant hardware, disk arrays or any other kind of adequate computer technique.

The typical IGS daily data flow at a regional or global data center sum up to a few hundred Mbytes/day at maximum: in an optimal situation a data center should collect the observation files from other centers or directly at the stations and propagate all or part of them after quality checks to the next neighbor in the network, hence moving roughly twice the amount of data corresponding to the files size. In the case of a global data center the amount of observation data to be moved through the center can be bigger due to the necessary equalization of data holdings among the 3 GDCs.

As an example, the present daily amount of data collected by IGN consists of 200 "daily" files = 60 Mbytes + 100 NRT files = 50 Mbytes + product files = 6 Mbytes

Some of these files are redistributed to other centers such as CDDIS and SIO. In average the daily IGS data flow at IGN is around 250 Mbytes, which is equivalent to 30 Kb/s in average network load. If we take into account optimal theoretical timeliness constraints of completing the "daily" stations data set in less than an hour and all the "hourly" files 5 minutes after time of last observation then we would have a network load of 340 Kb/s between 0-1 UT and a network load of 140 Kb/s every hour. These are rough estimates, but they are representative enough. An important factor in these figures is that at times when part of the data center network is not up for some reason, a single data center may have to cope with a much augmented data flow.

In addition to this rather predictable data flow, GDCs have to distribute data to IGS users (analysis centers or "unknown" users), this traffic has a "random" shape, there is no easy way to tell when there will be a peak or a low. At IGN, this amounts to 40 Gbytes/month, or 150 Kbit/s on average.

At each data center the routines and procedures used to manage the data flow have been developed, improved and maintained over many years, yielding to a rather satisfactory mature state.

In general, the data center files management requires moderate processing power. The tasks requiring some are file formatting, quality checks, reports generation, database requests, and miscellaneous data up keeping scripts. This is not a major problem. Should the network expand, the processing power needed would increase linearly w.r.t, to the number of stations. There is some users access induced processing load (ftp accesses and jukebox management), but this is marginal at the moment. This would change if web search and data browsing tools are implemented, most of the user requests will hit the web server (server side scripts) and the database system, some of them heavily.

In summary what can be said about the current situation w.r.t, day to day tasks at data centers is that the amount of data and the kind of data flow we have to deal with fit well into the hardware and software resources available at the data centers. Disk space is getting cheaper every year (100 Gbytes of storage capacity is equivalent to 3 years of IGS data and products) and network bandwidth capacity is not a problem considering that to my knowledge no RDC or GDC is connected to Internet with less than a TI/E1 (1.5-2Mbit/s).

Background data management:

These tasks are archiving, quality checks, historical archives update, special requests handling, software development (user access and management tools), report generation, ancillary data and documentation finding. The timetable of these operations can be out of sync w.r.t, the data flow. Actually there are time constraints on these tasks, but they are large enough to allow for some flexibility. No other IGS center or IGS users will suffer if some of these tasks are delayed for some time at a data center. These are as important as the daily data flow management, especially archiving and quality checks but the time table is mostly driven by the center's choice rather than by external availability constraints. There should not be any problem with that kind of operations even if we add RT data into the flow. RT data is not supposed to be archived so should have a low impact on these routines.

I won't detail these activities, this is not the main theme of the workshop. Furthermore, I don't see any limit there either. Each center has developed its own computer architecture and procedures to cope with that and it seems to work well enough up to now.

2. Present issues

Overall the situation is pretty satisfactory: data centers are up to the task most of the time. There are no technical barrier preventing us to deal with the present data flow and related activities. But this is only the big picture, I have detected issues and space for improvement in some areas. It is even more critical to sort out the problems attached to the "classical" IGS now that we want to integrate RT activities, I think this is the right time to clean up the present data flow operations in order to be able to extend them or coordinate them well with RT data flow. Many of the issues I'll address will have a negative impact on the RT data flow if not taken into account. I'm aware that the RT working group may have already addressed some of them but there is no harm in listing them from an other point of view, is there ?

Specifications and documentation

The specifications and documentation concerning the data flow are sparse. There is no detailed information about the operations taking place at each data center and how it should be done, the coordination between centers is done on a peer-to-peer basis without a global point of view. This surely leads to sub optimal data network performance: repeated transfers of the same file over the same link occur, delays are artificially kept higher than they could be due to lack of synchronization between centers, backup data paths are not used when they should be or used when they should no more, transfers are duplicated, wrong files are transmitted.

Most of the data centers are not newcomers so this may not strike most of us that there is indeed a lack of information. The present texts from the CB define a general framework but don't go into much detail. There is nothing wrong with interpreting the standards and guidelines but I think the directives should go into more details and be scrutinized globally by the IGS.

Here is a tentative list of documents that can be improved or added, the new ones are designed by a + sign

data center description form (.DCN files)
data center procedures description form (+)

This would complement the description form, focusing on technical aspects of the computer systems of the data center. Things to specify: software environment, scripting languages, database, operating system, hardware description and capacity (CPU, disk space, archival systems, jukeboxes), transfers time table, data sources, detailed list of files handled at the center, data checks description, regular neighbors, backup neighbors, network link capacity and redundancy
hourly and daily primary and backup data paths map (.NET files)

These maps exist in a synthetic form at the CBIS, but the information is not up to date and it is not clear whether the information is a description of how things are done or how they should be done. There is no apparent discrimination between daily/hourly data flow, the present CBIS map is mostly a daily data flow description. This document should be expanded to present a functional view of the data flow and be a reference document to define where a center should primarily get or put its files (concept of neighbors). The backup map will define the available backup paths to use when a DC or a link is down.

station site log

The site logs are used by data centers to refer to ancillary information at different steps in the data handling procedures. Only part of the information contained is used but still, the format is not really adapted to automated procedures and the contents are not always reliable.

station status table (+)

This table would give the status of each station in the network. It would specify whether a station is global, if it provides hourly or daily files, what operational center takes care of it. This information is presently partially available at the CBIS, and it is in html format thus not optimal for importing in a database.

data center guidelines
data center requirements

These two documents together would form a complete guide to help everyone run their data center. These would provide precise definitions of the requirements, with more details than is available presently. This would include requirements about data transfers coordination, data checks, reports, bandwidth, availability, downtime, directory structure, users request management, archivals. The guidelines document would present methods and software solutions than can be deployed to run a data center. Part of it could be a detailed technical description of tools and recipes developed and used at the present data centers. I think everyone would benefit from sharing this knowledge and expertise.

data holding report

Some of the holding reports available at the CBIS do not seem to be updated on a regular basis, yet there is a need to know precisely what files each data center holds. I think the overall distributed generation and assimilation process can be improved. These files are not documentation files per se, they are reports. They present some of the elements needed to assess the overall performance of the data network. As such they should be integrated in a more global and comprehensive supervision system (this is developed further in this document).

Regardless of their current status, all existing documentation files should be regularly reviewed to see if they are still adequate or need to be modified/expanded/dismissed.

Some of these files are no meant to be available to the average user of the IGS, they are only for private use within the IGS active components.

Some of these documentation files should be suitable for exchange among data centers, they should be human readable as well as computer readable. They should integrate flawlessly in procedures to upload their contents in a database. Most of the data centers copy information from the CBIS relevant to their activities into their own database in order to keep metadata information at hand during processing. I guess most of the analysis centers do the same. The present format of the metadata files is formatted ASCII, some of them are in HTML (global sites table), this is not optimal for content validation and exchange.

XML is an emerging technique that seems very well suited for that task. XML is a markup language that implements syntax and content validation. It provides a framework to help define data formats and has features that allow easy transfer and use between heterogeneous applications. It is an open standard and tends to spread out in many computer and network techniques. With XML the information stays in ASCII human readable form but is tagged so as applications can easily recover the data and validate the contents, there is also a standard mechanism that translates the XML file in any commonly used presentation format like formatted ASCII, HTML, PostScript, PDF. We may benefit from using XML for some of our metadata files. Good candidates are station site logs, data paths maps, data center description forms, stations status table, IGSMails. The adoption of XML would not have to break the view users have of these files, an IGS XML metadata file could be exported in any other format including the present ASCII formatted one, both formats could be provided at the same time ensuring continuity.

Monitoring and managing the data center network

The data network is a complex system, but not so complex as being out of control and impossible to manage and monitor. To do so, we would have first to identify the parameters that best characterize the data flow and then find a way of measuring them on a regular basis. This is much similar to what is commonly done in network and computer systems supervision.

Many *parameters* are valuable, here is a (non comprehensive) list to help clarify what I am talking about :

> basic knowledge about what part of the network is not operating as expected, what nodes are up/down, for what reason, is the latency ok everywhere, data sets completeness track down any particular file and see the path it went through to reach its location, bandwidth used, data flow benchmark, are unnecessary file transfers occurring, what is the current users download activity, be able to track down data path for each file to be able to verify that the path complies with the plan and help identify the cause when a problem arises

The addition of a RT data flow would call for even more and higher sampled parameters.

Some of these parameters are already available or can be deduced from the reports at CBIS.

There is a need for short term and long term reporting, short term metrics encompass all the elements that describe the situation of the network a short while ago, this is what would help raise a flag and take action when something bad is happening and quick reaction is needed. Long term metrics quantify the performance over a longer period of time, the parameters are basically the same only integrated or resampled and presented in a different way. The output of the long term supervision would be somewhat similar to the statistics presented by the data centers in the annual reports, with a higher sampling and greater freshness.

Access to the supervision system will be provided through a web interface. It will present the statistics in tables and graphics and provide raw metrics data in XML files for cross checking and other tasks.

Implementing such a tool is not an easy task, this will require a strong cooperation between data centers, each of us will have to generate statistics and send them to a central place for integration. The sampling may be high and the number or parameters to measure too. To do that some of the centers will have to develop programs and modify their routines if their own log files are not dense and detailed enough. A team will have to commit itself to the task of developing and implementing the supervision system. The final place where it will actually run is not really important as long as it is well connected and reliable.

Fortunately many tools are available and can be adapted to do the job, the web presentation and browsing features won't have to be written from scratch, existing software platforms in the open source world would do. As a consequence, most of the work will have to focus metrics definition and statistics generation and collection.

On another but related matter, a web based bug tracking system would be helpful in registering and resolving current and recurring problems with the data flow. This system could also take care of other component of the service. The bug tracking system would be operated at some web site, IGS CB behind the obvious choice. This would complement the IGSMail mailing list or rather the mailing list would complement it. Again free software solutions to do that are available (bugzilla, keystone), we actually use one such system at IGN.

Yet another item worth mentioning: there is a need to define a procedure to switch backup data paths. When a link fails, the problem should be identified as soon as possible and the backup data path for that part of the network should be used by all affected data centers. Presently the procedure is not well defined, should it be announced by the CB or decided among some data centers and reported later? Should it be automated or not? What exact conditions should be met to switch to the backup procedures? Also the backup data paths should be tested on a regular basis.

Data exchange techniques

It may be worth using file versioning techniques. The problem to address is how to uniquely identify a data or product file and discriminate whether it is worth propagating it if it is a new version or not if it the same as already stored. Most of the data centers will inspect a remote file or a local one in an upload area and download it only if the size or the date has changed, or even systematically propagate it to other centers. In some cases the file contents have not changed, no transfer is needed but the only way to know is by unambiguously identifying the contents. Checksumming and versioning techniques should be evaluated, new solutions to be developed can be considered too.

Common sense practices for data transfer Should be explained and suggested, they should be part of the data center guidelines document. Nobody is immune from error, a technique can seem perfectly well from one center's point of view and can be harmful to an other. Examples of what could be presented : the pros and cons of getting versus putting files, the importance of synchronization with neighbors, use of a Is-IR file, security aspects, anonymous versus named ftp, bandwidth saving techniques, fine tuning of the schedule, mirroring techniques, generating useful logs, getting data at the right place.

User segment

The user community of the IGS is expanding, every component of the IGS can see this. As far as data centers are concerned we see an increasing data download activity and receive more questions, requests and suggestions. Many users seem to be in need for more documentation. Progress is being made in this area, the CB has already produced a lot of documentation, and more is still to come I'm sure. I consider this is not really the job of the data centers to provide that kind of information, we should rather complement the guides and tutorials where data operations are explained. Another thing we can do to help users, especially at GDCs is providing a consistent and straightforward interface to the files and documentation.

Web interfaces should be developed to assist users and provide access to search tools and utilities. Many centers already provide some web tools with various levels of features and capacity. This should be encouraged and developed further. At a minimum, every Global Data Center should provide a common set of web tools. The CBIS too should provide these tools, there is presently no search tool over the CBIS web site which makes it a bit cumbersome to locate information sometimes.

Data centers are the main interface between users and products of the IGS. As such they can log a great deal of information about them and contribute to feed a global IGS user database. It is important to get to know the users of the IGS and to encourage feedback, this should be a contributing factor in the evolution of the service. Many techniques can be joined together to collect information about our users. The ftp and web access log files contain some information about the users, even if it is not really fine grained. A well thought web interface can motivate users to fill in a form and send us information. An IGS users mailing list and a forum can be set up in order to provide assistance and information.

Archive integrity

Historical data of the IGS are secured in every GDC archive. These data are getting more and more precious to many analysts because of the typical slow evolving nature of the phenomenons they study. Some groups are planning recomputations of all the observations. There is no easy way to tell whether all the files are present and coherent among the archives. The metadata may be incorrect or missing too. It is not easy either to decide what to do when something is wrong in an archive or a particular file. Archive integrity checking is a daunting task. It would be very interesting to hear from those that have plans or are doing it in order to define a policy and evaluate what can be done over the 3 GDCs.

Uncategorized items

Define a procedure to let data centers know when a station changes name or a new site opens. Stations changes have occurred in the recent past without any announcement made via IGSMaiI. This kind of information seems to circulate via alternate circuit without always reaching an official IGS resource.

Check import files not practical as a source for knowing whether a center stores the observations of a particular station. They are not easy to import in a database (most of our metadata is database driven, no flat files all over the place). Yet we need to know that when we want to add a station to our set. We need to know where is the best place to get the data and what are the backup sources in case the primary one fails.

Some files are not described by the IGS yet available at some data centers. These are files stored in the products directories but their naming conventions are not clear enough to help assess what should be done with them. A rule should be established requiring that any file due to storage at a GDC should be described and specified.

A simple thing that could help users, would be to adopt a common directory structure at all GDC. There is no need to really change existing directory structures, just construct another one with symbolic links. What this structure should be has to be discussed and agreed upon.

System date in UT at all centers would be more convenient, and a common time reference would even be better (ntp), if not possible have each center specify its time zone.

3. Future aspects

Real Time

RT data management will likely imply an additional load on all the data centers participating in it. The details of what will be required are not defined yet, at least to my knowledge. The workshop will help define *specifications* for that. Probably not all existing data centers will be required to be part of the RT data flow, those that will have to implement additional *procedures* and comply with new constraints. New RT dedicated data centers will be *created* or incorporated in the IGS. This may lead to a situation where two data network run in parallel within the IGS, there is no major problem with that expect that it should be well coordinated in order to be able to locate data easily (for users as well as for analysis *centers)* and flow either type of data efficiently through the network.

This is a broad view based upon the things I know about the future RT activities. I feel that Data centers are up to now in a standby position vis-à-vis RT, we expect that specifications and guidelines will come out from the workshop.

More services/products/users/centers/stations

Some of the contributing factors in the evolution of the IGS are predictable, some less. From what is presented in this text about the present situation, basically that there should be not technical reason why we couldn't cope with the current tasks. Also, the historical record of the IGS shows that we have been able to adapt to the continuous growth of the service. Thus one can boldly extrapolate that chances are that we will continue to do well in the future.

This will be true only if we prepare and anticipate the changes. This is why I consider important to monitor all factors contributing to modify the load the data network, I have already exposed that previously. Most of the evolutions listed in this section's header won't constitute a major problem, the induced load on the data network will increase gently whereas hardware and network capacity will increase exponentially to some extent.

More pressure may come from the users and the development of new services and products. This is actually what happens with the RT activities, they have a potential of quite suddenly multiply the requirements on data centers and other components by an important factor. We may even have to consider at some time in the future liability and economic constraints.

The load RT activities would put on the data network was considered huge a few years ago, it can now be re-evaluated as important but certainly not unmanageable, partly due to the progress in computer and network technology. Yet, many unknowns remain: there is a potential of demanding too much from the existing network thus requiring an additional RT dedicated network partly overlapping the present one.

There should be no need to archive RT data so the main issue that remains is the feasibility of putting through the network the data files from the RT components. As a worst estimate, this can amount to 30 times what we deal with presently and at the same time require much more stringent latency constraints.

Either the current data network will we required to take in charge the RT data flow or a separate network will be decided to be set up and interconnected with the present data network, the choice that will be made will undoubtedly have an effect over our current data flow practices.

We'll also have to take into account what present data centers will be willing and able to do regarding the RT activities. Whatever the decisions that will come up from the workshop, the impact will have to be estimated and the data network adapted in accordance.

IGS Data Center Security Issues

Heinz Habrich

Regional IGS Data Center Europe

Bundesamt fuer Kartographie und Geodaesie
Frankfurt, Germany

IGS Workshop 2001, 15-18 October, Ottawa

Preliminary Remarks

- I´m not an IT-expert and I will talk with the background of an IT-user.
- Motivation: I have noticed general changes in the security concepts affecting the IGS Data Centers
- Primary Objective of this presentation is the exchange of experience (in follow-up discussions)

Introduction

- Review of the applied security concepts since the beginning of the IGS

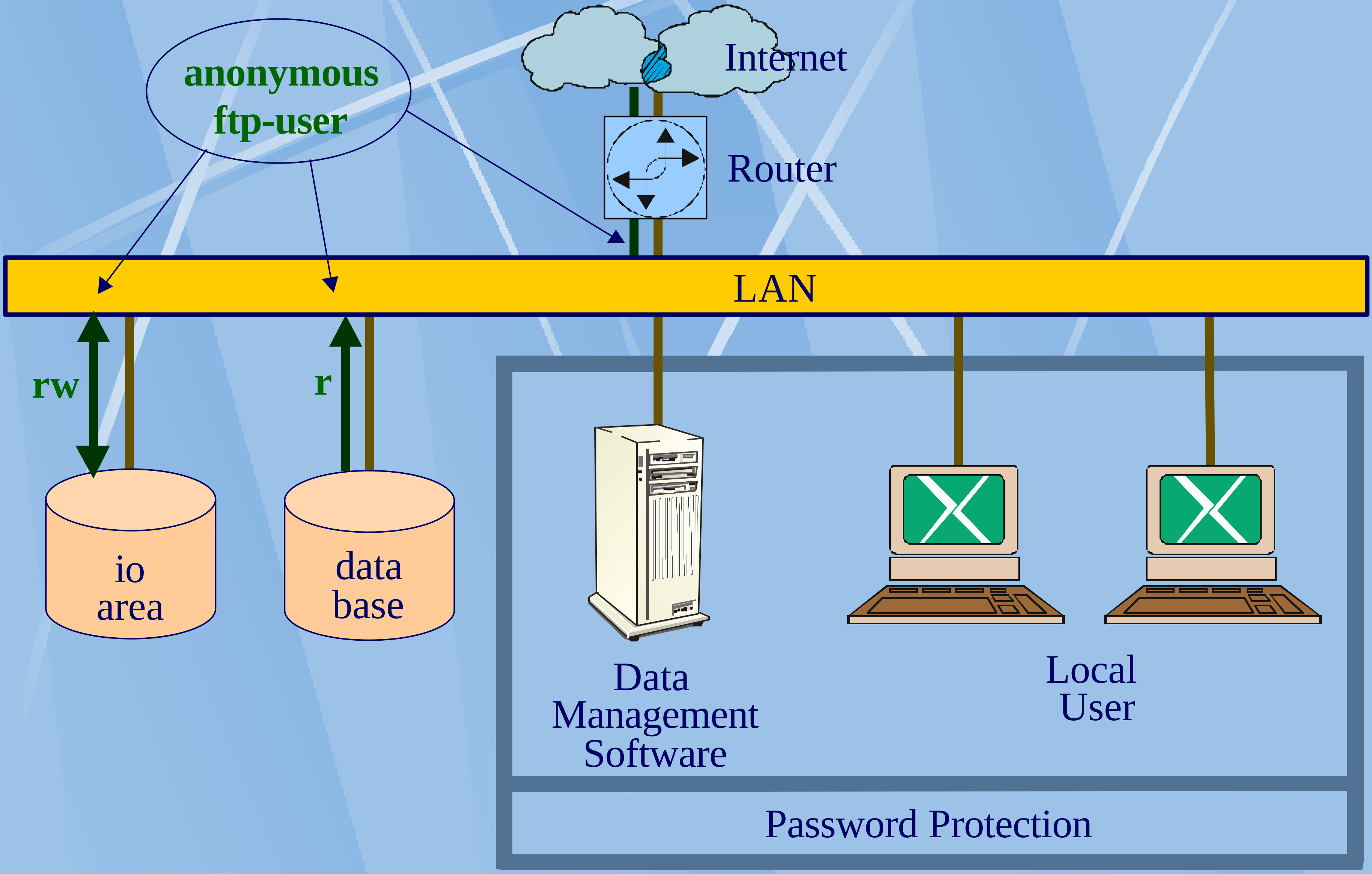
  Focus on components

  - Firewall
  - ftp-protocol

- Future security requirements for IGS data transfer

- Summary

Scenario at the beginning the IGS in 1992



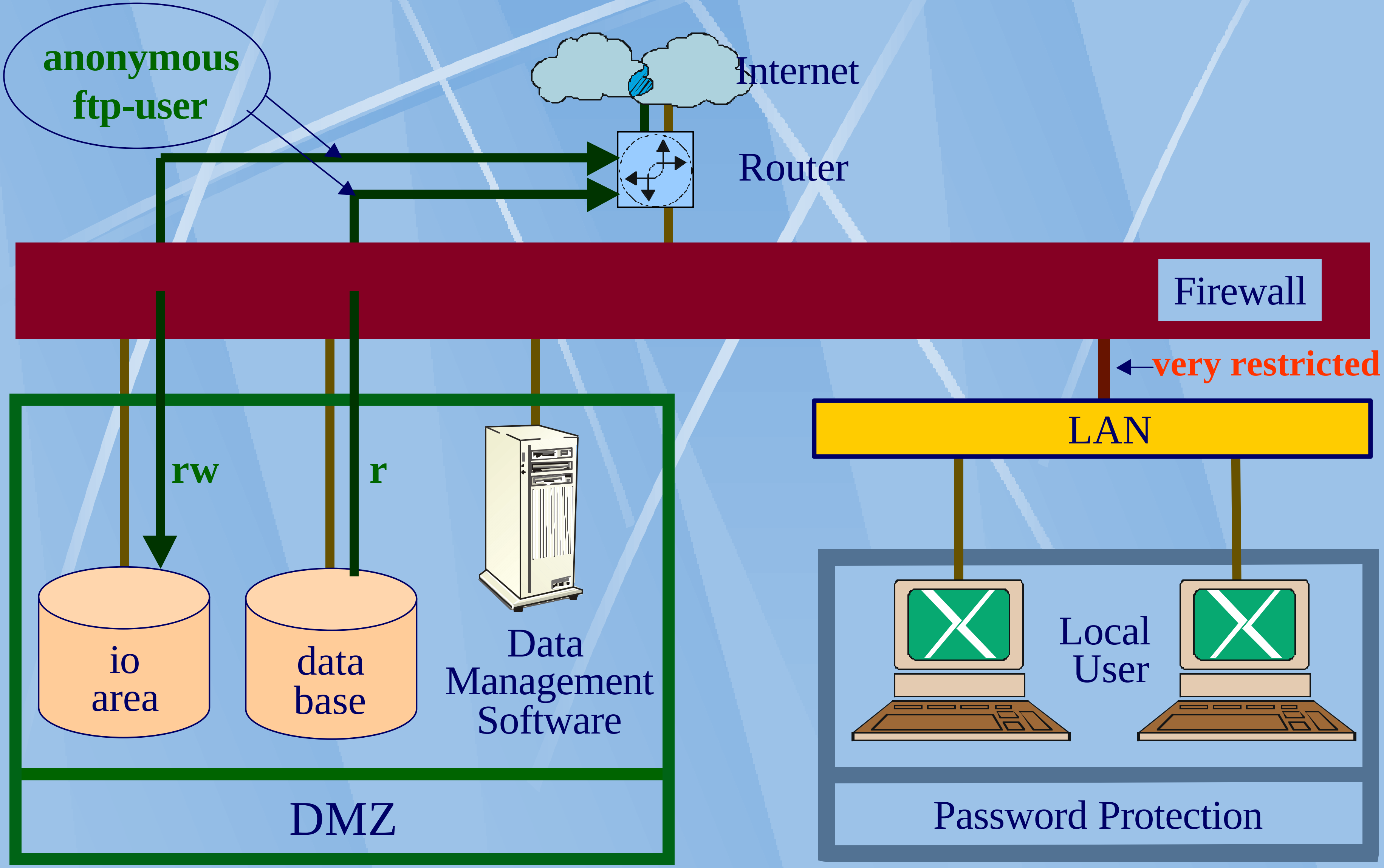
Scenario at the beginning the IGS in 1992

- ☹ Overflow of io-area, because of write permission for a-ftp
- ☹ Poor protection of local user area (password only)
- ☹ No protection of LAN; a-ftp could block the LAN

- ☺ Easy upload and download of IGS data by usage of „a-ftp login“
- ☺ Free access of local user to all areas

- ➲ Need for change!

Intermediate Scenario



Intermediate Scenario

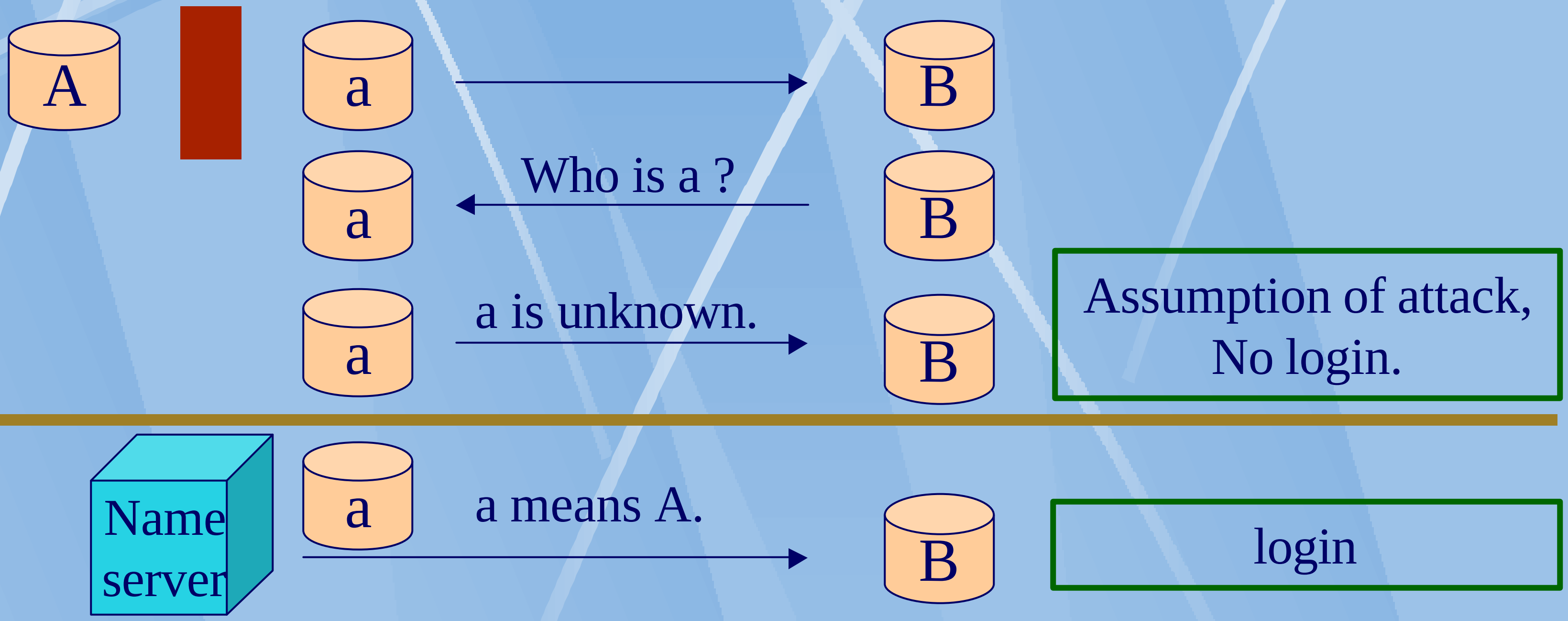
- ☹ Overflow of io-area, because of write permission for a-ftp
- ☹ Poor protection of DMZ; a-ftp could block the DMZ
- ☹ Restricted access of local user to DMZ; more difficult operation of the Data Center
- ☹ Different Firewall configuration for each IGS Data Center

- ☺ Easy upload and download of IGS data by usage of „a-ftp login“
- ☺ Good protection of LAN

- ➲ Need for change!

Firewall Configuration
- Reverse Name Resolution (RNR)-

- **Functionality:**

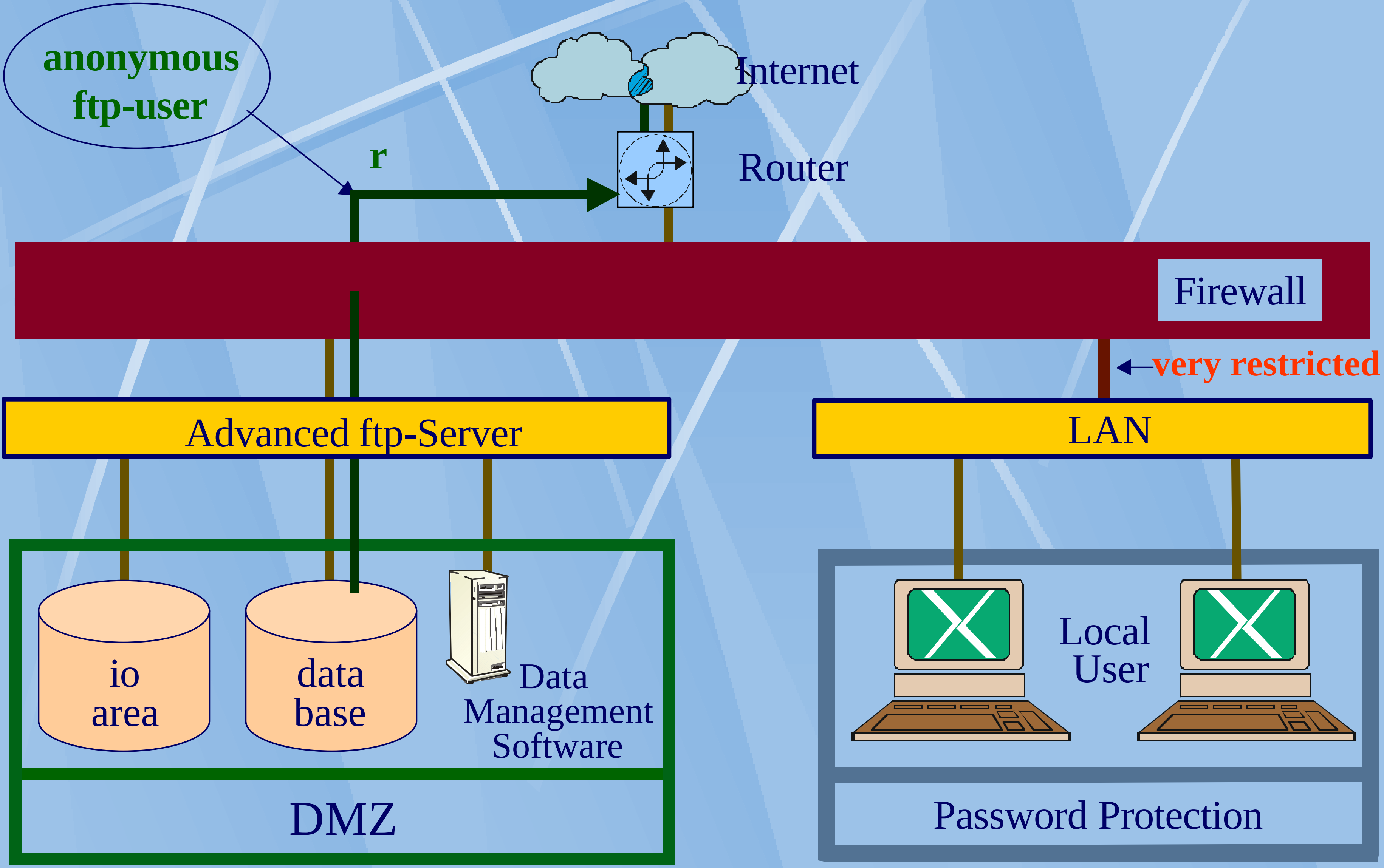


- **Problem:** RNR at a DC requires a relaxation of security configuration for other DCs
- **Action Item:** Recommendation of IGS would help

Firewall Configuration
- Passive Ftp Mode -

- **Functionality:** Only port 21 (control port) is opened by the firewall, no additional port for data transfer.
- **Problem:** Old ftp-programs do not support „passive mode“.
  - Typical error: Login successful, but „dir“ command fails.
- **Action item:** Others than port 21 data ports should be permitted by the firewall.

Current Popular Scenario



Current Popular Scenario

- ☹ No write permission for a-ftp
- ☹ Restricted access of local user to DMZ; more difficult operation of the Data Center
- ☹ Different Firewall configuration for each IGS Data Center
- ☹ Different configuration of the new ftp-server for each IGS Data Center
- ☹ Upload IGS data by login with user-id and password
- ☺ Good protection of LAN
- ☺ Good protection of io-area against write-overflow
- ➲ Need for change!

Washington University (WU) Ftp-Server

- **Standard ftp-server:**
  - Common root-directory for all ftp-users
- **WU ftp-server:**
  - Separate root-directories for each ftp-user
  - Example:

Project	Incoming Directory	User
IGS	IGS/igsin	igspush
	IGS/nrtin	igsnrt
	IGS/mirrorin	igsmirr
EUREF	...	

  - Advantage:
    - Password protection for „write“ command
    - Separation of data types while login procedure, e.g., data which are submitted for backup could easily be handled

Usage of http for IGS

- **Anonymous user:**
  - Download of observation data, products and information by everyone (*standard today*)
- **Restricted access:**
  - Exchange of auxiliary information by IGS associates
    - Backup control
    - Mirror control
    - Usage of http as an administration tool
  - Problem: Firewall does not allow for all http features, e.g., „Java“ denied

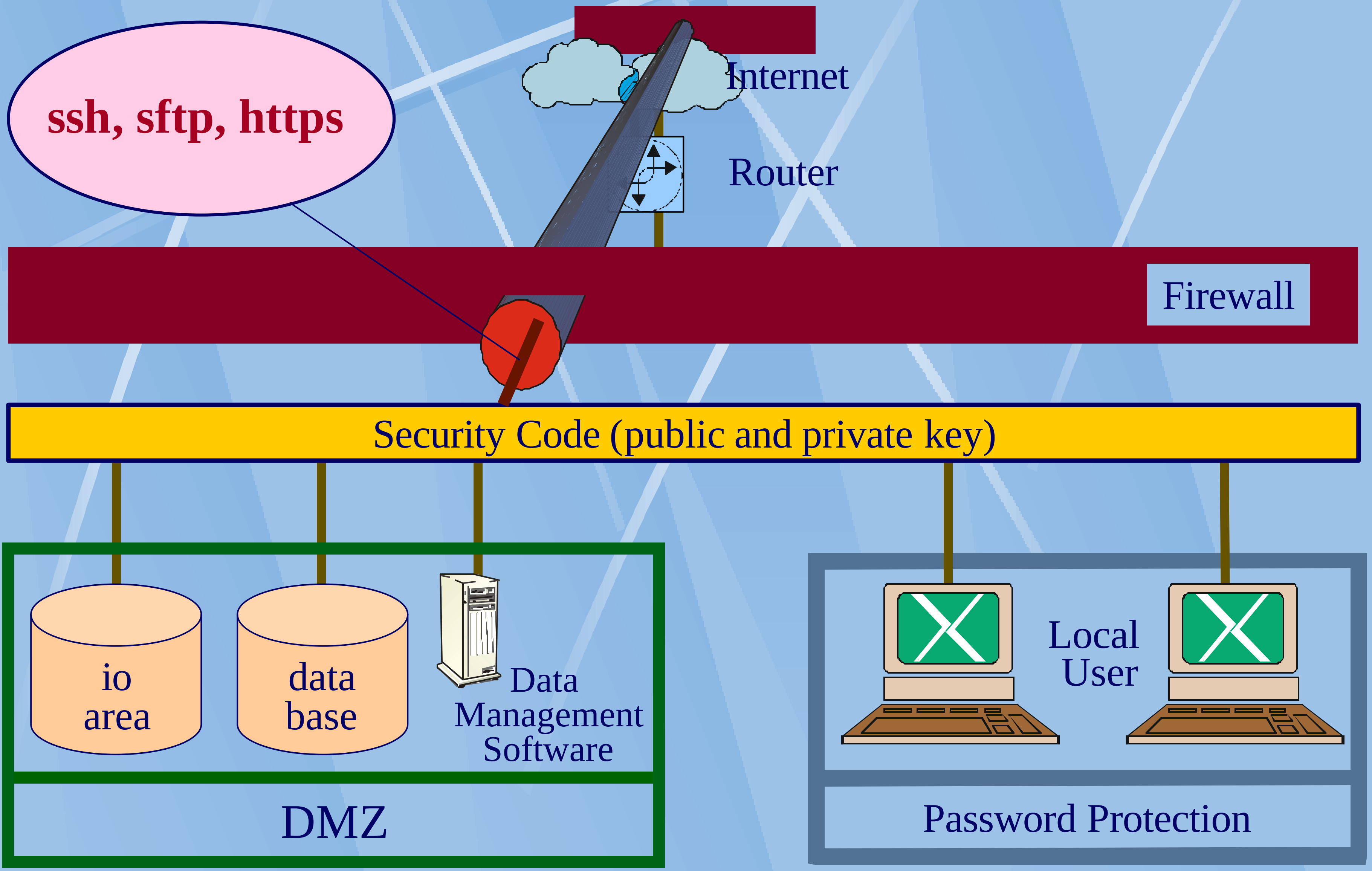
Requirements for Future Scenarios

- Protection of the data base and the connected LAN against attacks from the Internet has to be guaranteed
- Usage of a-ftp reduced to download purposes
- Data center specific firewall configurations may not affect the IGS data flow
- Others than port 21 data ports should be opened, e.g., for real time data transfer
- It is recommended to establish a unified functionality for the data transfer at all data centers (equal „put“ and „get“ commands and and equal directory structure)

Future Scenario ?

- Secure protocol applications:
  - ssh / sftp / https
- Requirement:
  - Exchange of security code between members of this „Virtual Private Network (VPN)“
  - Not all station operators could become member of the VPN
- Advantages:
  - Realization of a „state of the art“ security policy
  - All communication ports could be used
  - Secure protocol will be accepted by most firewalls
  - Real time and mirror programs could be started

Future Scenario ?



Possible User Classes

- (1) Secure protocol:
  - Global and regional data centers and analysis centers
  - ssh, sftp and https
- (2) User und password protection:
  - operational data centers and station operators
  - Upload of tracking data
  - ftp
- (2) Anonymous user:
  - everyone
  - ftp and http

Summary

- Review of internet connections of IGS Data Centers shows a development towards improved security concepts
- Attempt to formulate requirements for future scenarios
- It should be discussed, whether IGS needs a common security concept
- Exchange of security experiences between Data Centers should be improved

GPS Seamless Archive Centers (GSAC)

A UNAVCO project

Streamlining data/metadata exchange in the GPS community

Michael Scharber, Scripps Orbit and Permanent Array Center

Contents

- General Definition
- GSAC as a Network
- What makes the GSAC useful?
- Status of the GSAC
- Example Client-Retailer Interaction
- Development Plans
- GSAC Participation
- Contact Information

General Definition

Loosely defined, the GSAC is a network of agencies, users, data/metadata exchange formats and communication protocols facilitating the expeditious flow and availability of GPS-related information among it's participants.

Contents

- General Definition
- GSAC as a Network
- What makes the GSAC useful?
- Status of the GSAC
- Example Client-Retailer Interaction
- Development Plans
- GSAC Participation
- Contact Information

GSAC as a Network

The GSAC is composed of a network of:

- Providers
- Wholesalers
- Retailers
- Client Applications (software)
- End Users

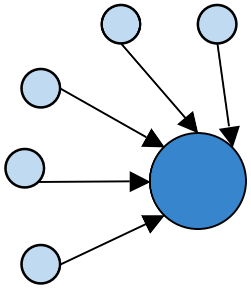
GSAC as a Network

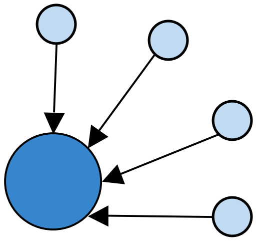
GSAC Providers

Act as the source for GPS-related data/metadata in the GSAC.

Though providers need not be directly involved in the GSAC, without their data the GSAC cannot function.

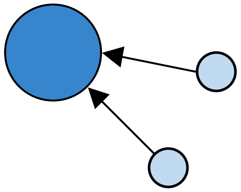
GSAC as a Network

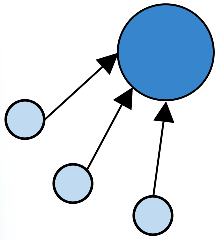


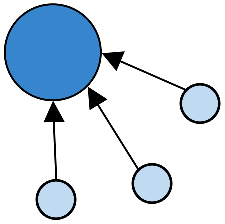


GSAC Wholesalers

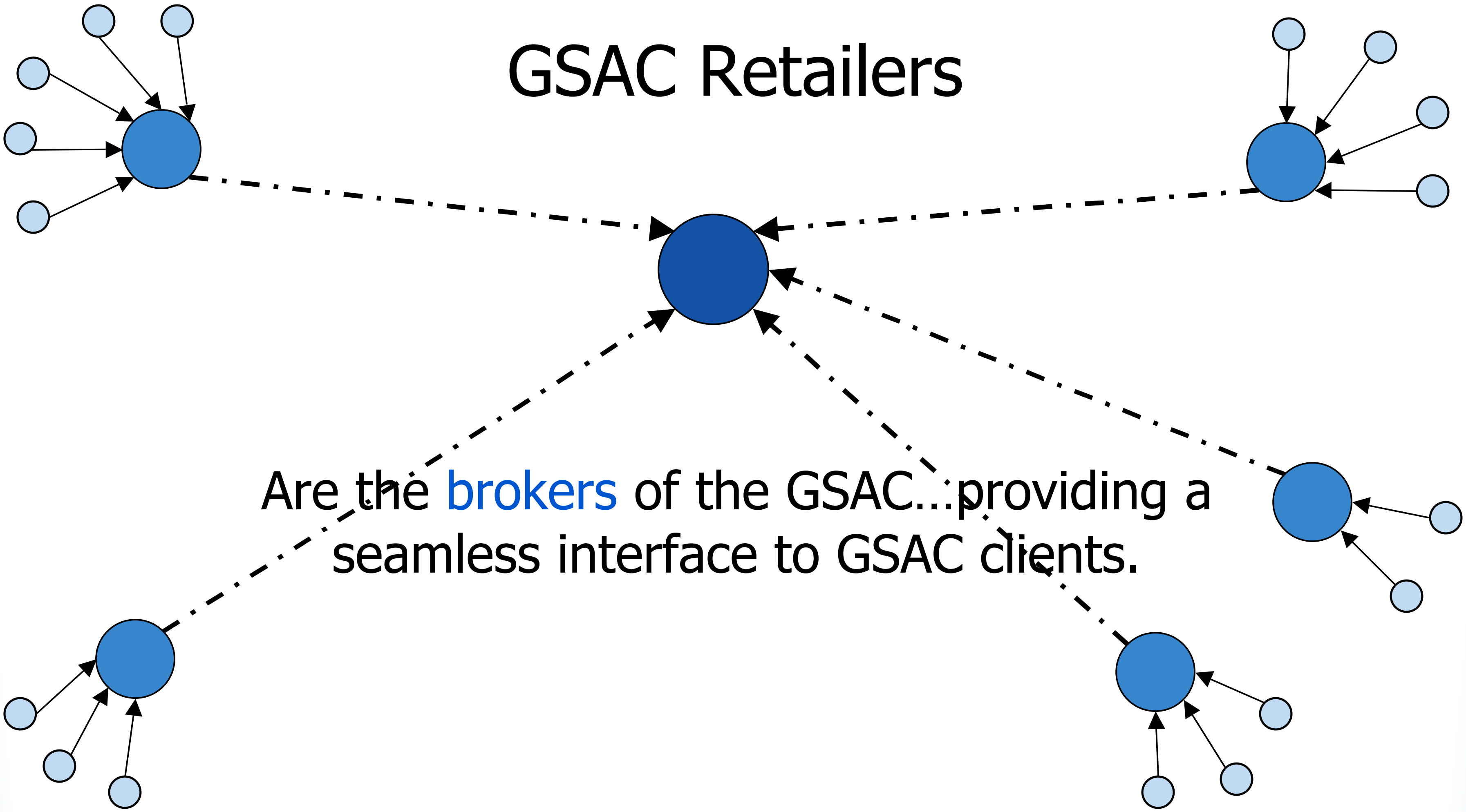
Provide a residence for GPS-related data/metadata in the GSAC. They collect/ receive data/metadata from a set of providers…and then make that data available to the GSAC via a prescribed publication protocol and data availability policy.







GSAC as a Network



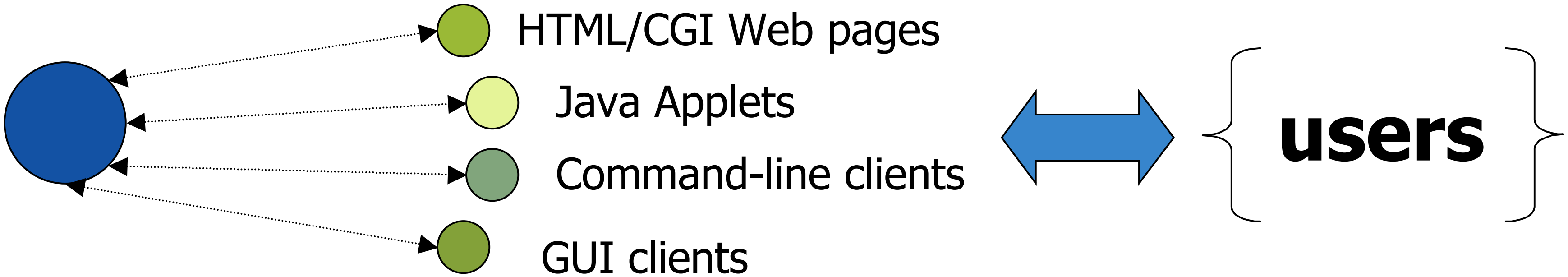
GSAC Retailers

Are the brokers of the GSAC…providing a seamless interface to GSAC clients.

GSAC as a Network

GSAC Clients

Communicate with retailer services on the behalf of end users…



HTML/CGI Web pages

Java Applets

Command-line clients

GUI clients

users

…facilitating data discovery, data collection and data sharing among GSAC participants.

Contents

- General Definition
- GSAC as a Network
- What makes the GSAC useful?
- Status of the GSAC
- Example Client-Retailer Interaction
- Development Plans
- GSAC Participation
- Contact Information

What makes the GSAC useful?

GSAC relationships foster benefits on several different levels:

- End User - Retailer (via Client)
- Client - Retailer
- Retailer - Wholesaler
- Wholesaler - Wholesaler
- Wholesaler - Provider

What makes the GSAC useful?

End User - Retailer Relationship

- Centralized and standardized data/metadata acquisition interface spans entire wholesaler network.
- Integrated query context allows mixture of descriptive, temporal and spatial constraints in GSAC queries.
- Standardized vocabulary for data types, dates and times (UTC), compression and file grouping types
- Clarified relationships between sites/data, responsible entities and originating source

What makes the GSAC useful?

Client - Retailer Relationship

- Retailer service protocol fosters GSAC client diversity.
- Independently-maintained clients reduce need for reinvention of data collection software.
- Streamlined data discovery benefits utility of GPS-related information.

What makes the GSAC useful?

Retailer - Wholesaler Relationship

- Publication process promotes conformance across wholesalers, facilitating centralized brokerage of information.

- Retailers can integrate wholesaler interchange records into a single database, and serve a metadata quality-control role for the GSAC in the meantime.

What makes the GSAC useful?

Wholesaler - Wholesaler Relationship

- Mirroring data among wholesalers is simplified and streamlined by a common cataloging process and metadata interchange format.

- Redundancy created through the data mirroring process allows retailers to redirect clients to alternate locations seamlessly, in the event of wholesaler downtime.

What makes the GSAC useful?

Wholesaler - Provider Relationship

- Providers benefit from increased visibility and recognition of their data in the GPS community, as well as data quality-control provided by wholesaler.

- GSAC community benefits from single point-of-entry for all published data, reducing confusion concerning the origin of GPS data.

Contents

- General Definition
- GSAC as a Network
- What makes the GSAC useful?
- **Status of the GSAC**
- Example Client-Retailer Interaction
- Development Plans
- GSAC Participation
- Contact Information

Status of the GSAC

- 1997-2000 - Preliminary investigation, design and development
- 2001 (initial funding) - Devoted primarily to development and transportability of wholesaler, retailer and client software…and retailer design specifications.
- Entered testing phase of network in March 2002.
- Will be operational in summer of 2002.

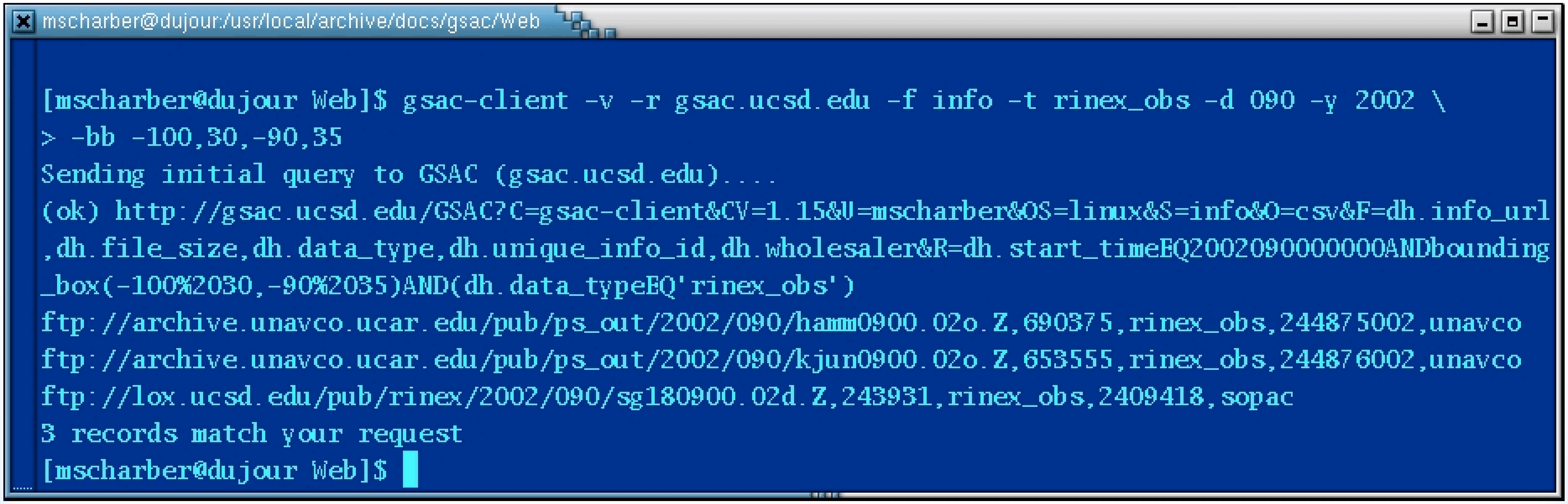
Contents

- General Definition
- GSAC as a Network
- What makes the GSAC useful?
- Status of the GSAC
- Example Client-Retailer Interaction
- Development Plans
- GSAC Participation
- Contact Information

Client-Retailer Interaction

Command-Line Clients

Ex. Linux gsac-client binary fetching *information* from GSAC

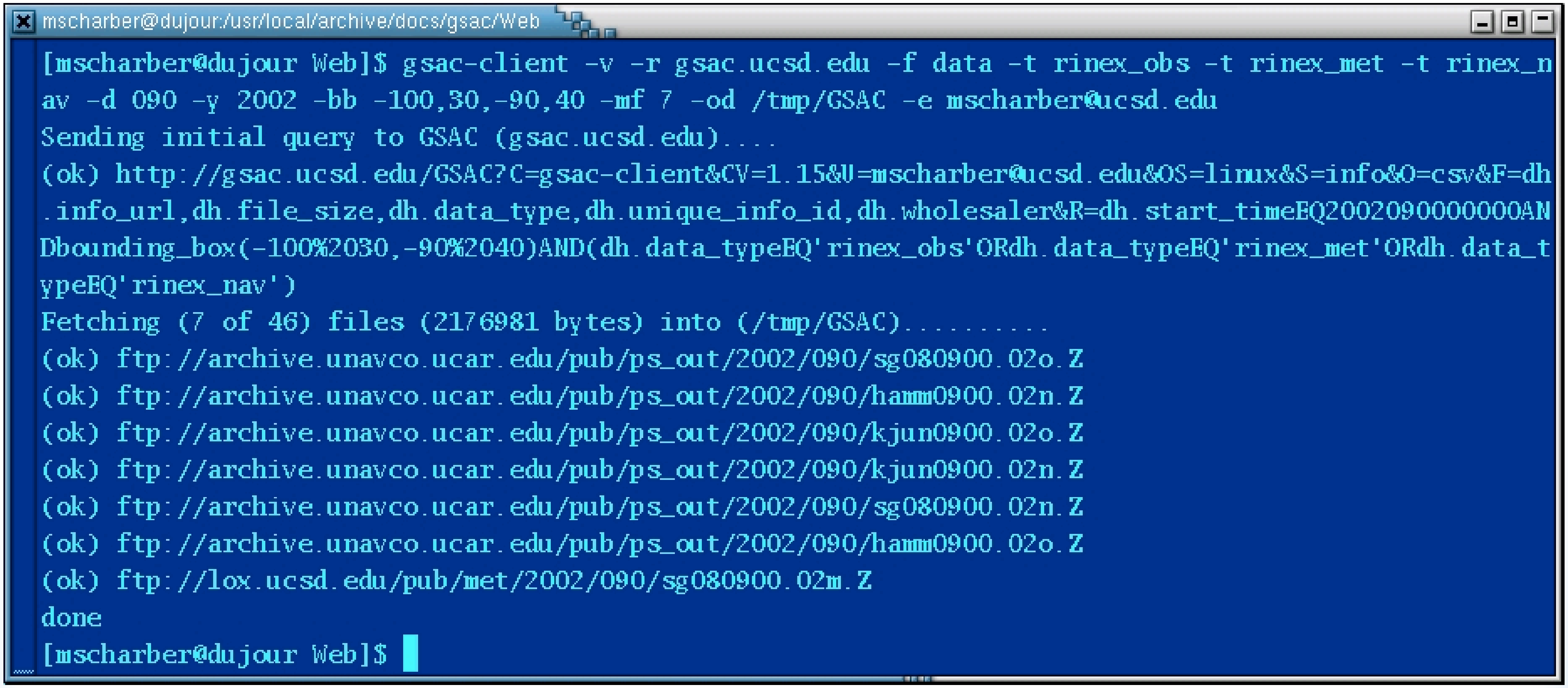


```
[mscharber@dujour Web]$ gsac-client -v -r gsac.ucsd.edu -f info -t rinex_obs -d 090 -y 2002 \
> -bb -100,30,-90,35
Sending initial query to GSAC (gsac.ucsd.edu)....
(ok) http://gsac.ucsd.edu/GSAC?C=gsac-client&CV=1.15&U=mscharber&OS=linux&S=info&O=csv&F=dh.info_url
,dh.file_size,dh.data_type,dh.unique_info_id,dh.wholesaler&R=dh.start_timeEQ2002090000000ANDbounding
_box(-100%2030,-90%2035)AND(dh.data_typeEQ'rinex_obs')
ftp://archive.unavco.ucar.edu/pub/ps_out/2002/090/hamm0900.02o.Z,690375,rinex_obs,244875002,unavco
ftp://archive.unavco.ucar.edu/pub/ps_out/2002/090/kjun0900.02o.Z,653555,rinex_obs,244876002,unavco
ftp://lox.ucsd.edu/pub/rinex/2002/090/sg180900.02d.Z,243931,rinex_obs,2409418,sopac
3 records match your request
[mscharber@dujour Web]$ 
```

Client-Retailer Interaction

Command-Line Clients

Ex. Linux gsac-client binary fetching *data* from GSAC

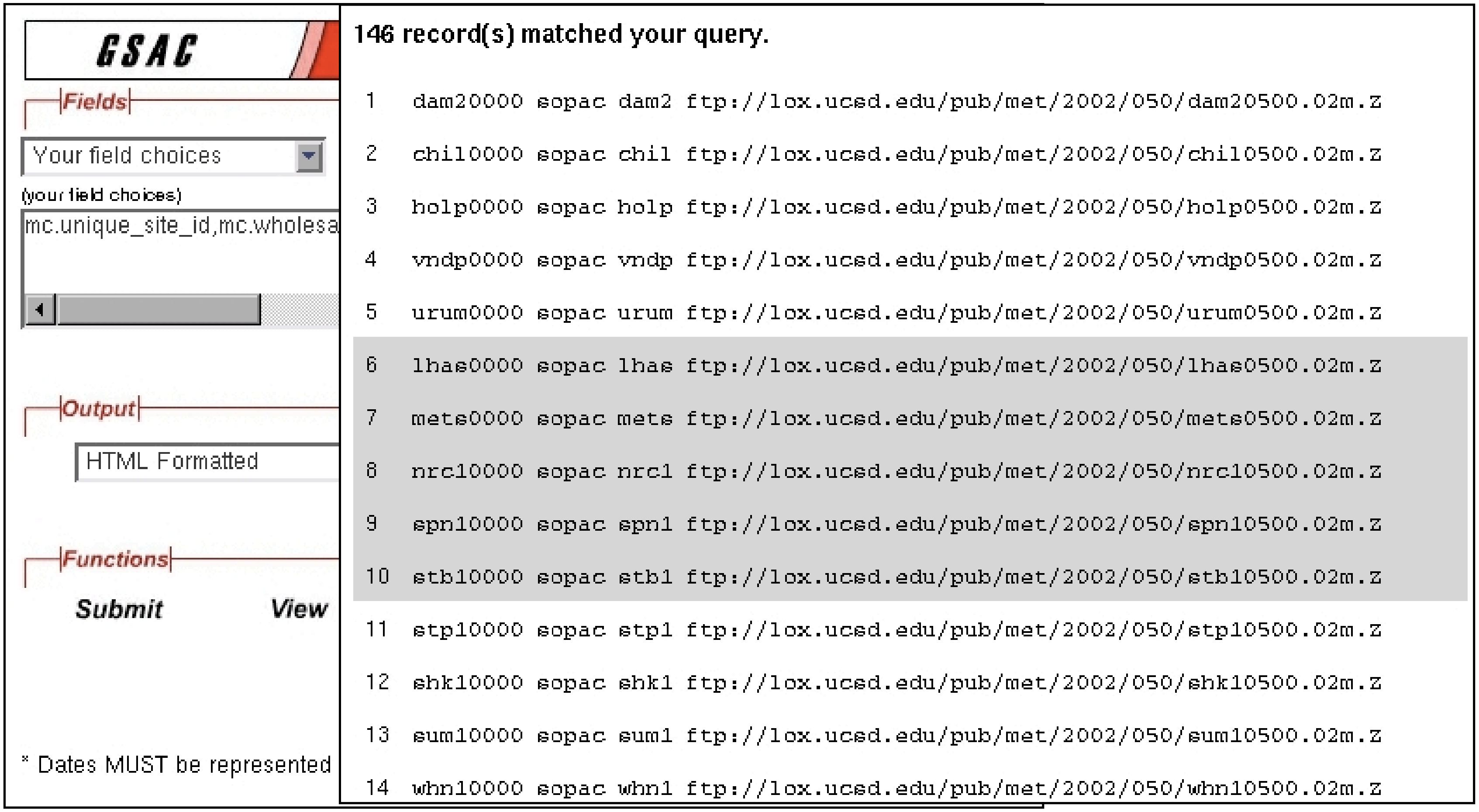


```
[mscharber@dujour Web]$ gsac-client -v -r gsac.ucsd.edu -f data -t rinex_obs -t rinex_met -t rinex_n
av -d 090 -y 2002 -bb -100,30,-90,40 -mf 7 -od /tmp/GSAC -e mscharber@ucsd.edu
Sending initial query to GSAC (gsac.ucsd.edu)....
(ok) http://gsac.ucsd.edu/GSAC?C=gsac-client&CV=1.15&U=mscharber@ucsd.edu&OS=linux&S=info&O=csv&F=dh
.info_url,dh.file_size,dh.data_type,dh.unique_info_id,dh.wholesaler&R=dh.start_timeEQ2002090000000AN
Dbounding_box(-100%2030,-90%2040)AND(dh.data_typeEQ'rinex_obs'ORdh.data_typeEQ'rinex_met'ORdh.data_t
ypeEQ'rinex_nav')
Fetching (7 of 46) files (2176981 bytes) into (/tmp/GSAC)..........
(ok) ftp://archive.unavco.ucar.edu/pub/ps_out/2002/090/sg080900.02o.Z
(ok) ftp://archive.unavco.ucar.edu/pub/ps_out/2002/090/hamm0900.02n.Z
(ok) ftp://archive.unavco.ucar.edu/pub/ps_out/2002/090/kjun0900.02o.Z
(ok) ftp://archive.unavco.ucar.edu/pub/ps_out/2002/090/kjun0900.02n.Z
(ok) ftp://archive.unavco.ucar.edu/pub/ps_out/2002/090/sg080900.02n.Z
(ok) ftp://archive.unavco.ucar.edu/pub/ps_out/2002/090/hamm0900.02o.Z
(ok) ftp://lox.ucsd.edu/pub/met/2002/090/sg080900.02m.Z
done
[mscharber@dujour Web]$
```

Client-Retailer Interaction

Web Page Clients

Ex. GSAC.cgi client fetching *information* from GSAC



GSAC

Fields

Your field choices

(your field choices)

mc.unique_site_id,mc.wholesa

Output

HTML Formatted

Functions

Submit View

* Dates MUST be represented

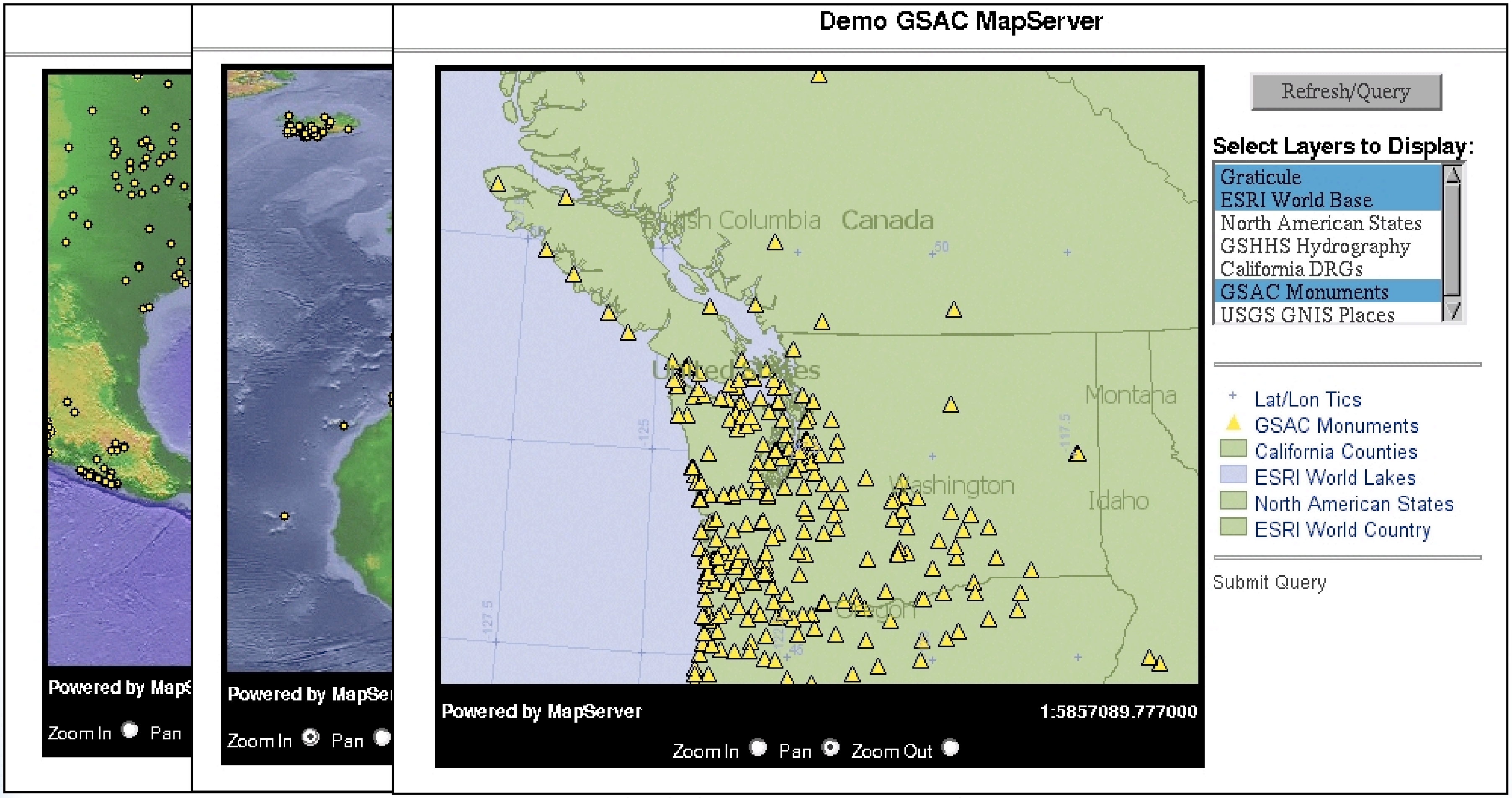
146 record(s) matched your query.

```
1   dam20000 sopac dam2 ftp://lox.ucsd.edu/pub/met/2002/050/dam20500.02m.Z
2   chil0000 sopac chil ftp://lox.ucsd.edu/pub/met/2002/050/chil0500.02m.Z
3   holp0000 sopac holp ftp://lox.ucsd.edu/pub/met/2002/050/holp0500.02m.Z
4   vndp0000 sopac vndp ftp://lox.ucsd.edu/pub/met/2002/050/vndp0500.02m.Z
5   urum0000 sopac urum ftp://lox.ucsd.edu/pub/met/2002/050/urum0500.02m.Z
6   lhas0000 sopac lhas ftp://lox.ucsd.edu/pub/met/2002/050/lhas0500.02m.Z
7   mets0000 sopac mets ftp://lox.ucsd.edu/pub/met/2002/050/mets0500.02m.Z
8   nrc10000 sopac nrc1 ftp://lox.ucsd.edu/pub/met/2002/050/nrc10500.02m.Z
9   spn10000 sopac spn1 ftp://lox.ucsd.edu/pub/met/2002/050/spn10500.02m.Z
10  stb10000 sopac stb1 ftp://lox.ucsd.edu/pub/met/2002/050/stb10500.02m.Z
11  stp10000 sopac stp1 ftp://lox.ucsd.edu/pub/met/2002/050/stp10500.02m.Z
12  shk10000 sopac shk1 ftp://lox.ucsd.edu/pub/met/2002/050/shk10500.02m.Z
13  sum10000 sopac sum1 ftp://lox.ucsd.edu/pub/met/2002/050/sum10500.02m.Z
14  whn10000 sopac whn1 ftp://lox.ucsd.edu/pub/met/2002/050/whn10500.02m.Z
```

Client-Retailer Interaction

Web Page Clients

Ex. GSAC Map Server client browsing *sites* in GSAC



Select Layers to Display:

- Graticule
- ESRI World Base
- North American States
- GSHHS Hydrography
- California DRGs
- GSAC Monuments
- USGS GNIS Places

+	Lat/Lon Tics
▲	GSAC Monuments
🟩	California Counties
▨	ESRI World Lakes
🟩	North American States
🟩	ESRI World Country

Contents

- General Definition
- GSAC as a Network
- What makes the GSAC useful?
- Status of the GSAC
- Example Client-Retailer Interaction
- Development Plans
- GSAC Participation
- Contact Information

Development Plans

- Attract additional wholesalers and retailers to the GSAC.
- Broaden diversity and functionality of GSAC clients.
- Extend spatial functionality of retailer databases.
- Streamline wholesaler publication processes.
- Shift from anonymous ftp servers and file-based interchange mechanism at wholesaler level to an http web service strategy.
- Add support for additional types of data, including realtime (stream-based) data contexts.

Contents

- General Definition
- GSAC as a Network
- What makes the GSAC useful?
- Status of the GSAC
- Example Client-Retailer Interaction
- Development Plans
- GSAC Participation
- Contact Information

GSAC Participation

The following organizations are either actively involved in the development of the GSAC or have expressed interest in participating:

- UNAVCO - Colorado, USA
- SOPAC - California, USA
- MIT - Massachusetts, USA
- CDDIS - Maryland, USA
- SCEC - California, USA
- NGS/CORS - Maryland, USA
- NCEDC - California, USA
- PANGA - Washington, USA
- WCDA - British Columbia, Canada

GSAC Participation

With the participation of each additional wholesaler/retailer the GSAC will hopefully gain utility in many different sectors of the GPS community.

Interested agencies/individuals are encouraged to contact any of the parties listed at the end of this presentation.

Contents

- General Information
- GSAC as a Network
- What makes the GSAC useful?
- Status of the GSAC
- Example Client-Retailer Interaction
- Development Plans
- GSAC Participation
- Contact Information

Contact Information

- **UNAVCO - http://www.unavco.ucar.edu/data_support/data/gsac/GSAC-1.html**
  - background
  - definition
  - history
- **SOPAC - http://gsac.ucsd.edu**
  - software
  - presentations
  - development support

Contact Information

- Massachusetts Institute of Technology (MIT)
  - Bob King - UNAVCO GSAC Community Coordinator rwk@prey.mit.edu

- University Navstar Consortium (UNAVCO)
  - Chuck Meertens - chuckm@unavco.ucar.edu
  - Fran Boler - fboler@unavco.ucar.edu

- Scripps Institution of Oceanography (SIO)
  - Yehuda Bock - bock@ucsd.edu
  - Michael Scharber - mscharber@ucsd.edu