電離圏全電子数データベースを例とした研究者グループによる研究データの公開・維持について

Development and maintenance of scientific databases by a group of “non-data-center” scientists

A. Saito (Dept. Geophysics,
Graduate School of Science, Kyoto University)

4 Phases: An ideal scenario for “homemade” database developed by non-data-centre scientists (～ IT start-up business scenario)

A. Obtain/produce and maintain new data set.

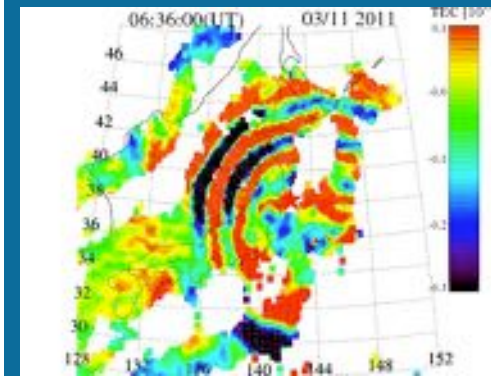
B. Make them available for collaborating scientists/colleagues.

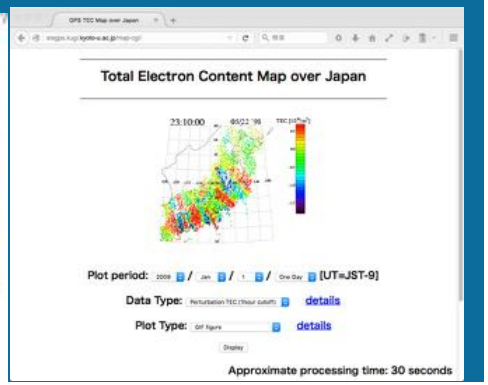
C. Make them available for public.

D. “Big” institute acquires the database and maintains it permanently.

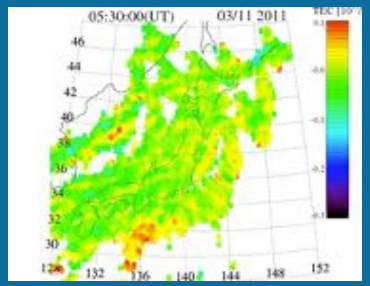
GPS-Total Electron Content database: Phase-D

- Ionospheric plasma content derived from the GPS data obtained by GEONET of Geospatial Information Authority of Japan (GSI)
- GSI does not provide the plasma data as their data service.
- 4 G byte/day (uncompressed) since 1997.
- Data is processed in quasi-realtime.
- 7 T byte disk space is used now.

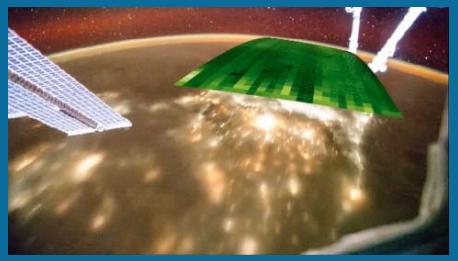




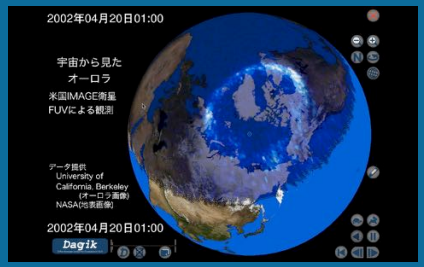
Three databases that we have developed and maintained



GPS-Total Electron Content (2001-): Phase-D

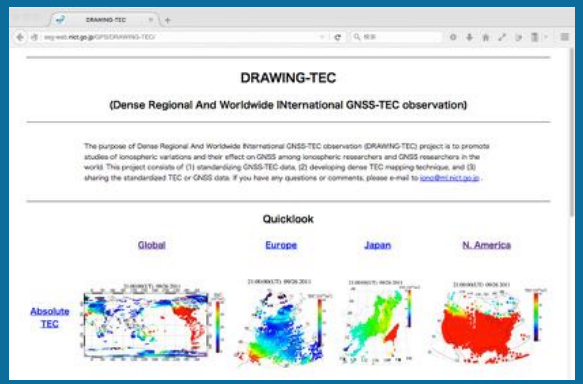


ISS-IMAP (2012-): Phase-B



Dagik Earth (2009-): Phase-C

Phase-A：1998-1999：TEC algorithm was developed.
Phase-B：2000-2001：Database from 1997 was developed
Phase-C：2001-：Web service developed and maintained with A JSPS grant, Kakenhi-Database, 科研費 研究成果公開促進費「データベース」, between 2002 and 2009.
Phase-D：2010?：NICT starts TEC database service using our TEC algorithm with global GPS data set.



http://seg-web.nict.go.jp/GPS/DRAWING-TEC/

Lessons learned in GPS-TEC database project

Why did we start it?

It is efficient to make database for usage of collaborators because the data size was huge for the computer environment at that time. GSI and NICT were not able to make the GPS-TEC database. Technical difference between Phase-B (closed database) and Phase-C (open database) was relatively small.

What are good points?

- Data usage gets easy for us.
- Several collaborating researches, several papers as co-author, and establishment of research community of GPS-TEC users. But some of them might be accomplished even in Phase-B.
- “Long-tail” usage is the advantage of Phase-C, but it is mostly invisible.
- Increase of HDD capacity is faster than that of the data size

Lessons learned in GPS-TEC database project (Continued)

What are difficulties?

- Data backup is necessary. Hardware troubles are inevitable. Backup server and disk-system are supposed to be on the shelf.
- Budget is limited. JSPS grant cannot be used for server and disk-system. Obtaining the budget for 10 years with one-year base application is not easy.
- Update of softwares, and daily maintenance for data error are necessary.
- Scientific target shifted in 10 years. Priority of the data has changed in time because of the other projects.

Lessons learned in GPS-TEC database project (Continued)

How to reduce the load of database maintenance?

- “Cloud” can reduce the hardware maintenance tasks to make trouble shooting and prepare backup systems.
- Continuous budget is more difficult to obtain than one-time budget.

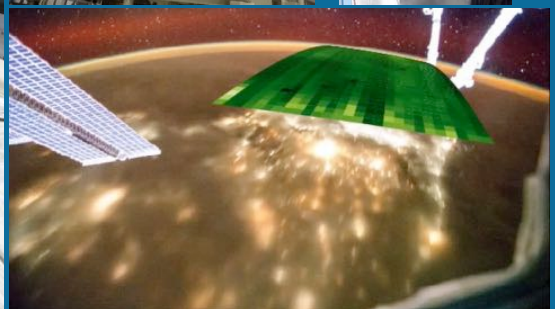
Conclusions

- It is difficult to maintain Phase-C database (open database) for more than 10 years by “non-data-center” scientists.
- Phase-D database (data/technology transferred database) by NICT is very helpful although not all the data is served by the NICT database.

ISS-IMAP data: Phase-B

- Imaging observation of the airglow and the aurora from International Space Station from 2012 to 2015.
- Joint project of Kyoto Univ., JAXA, Tohoku Univ., Univ. Tokyo and other institutes.

[http://www.nasa.gov]



Phase-A : 2011-2015 : Data processing system was developed.
Phase-B : 2012-2015 : Database on a server in Kyoto Univ. for collaborators.
Phase-C : 2016- (plan) : Level-1 and Level-2 data will be open for public
Phase-D : 2016- (plan) : JAXA will host ISS-IMAP database in DARTS/C-SODA, permanently.

Lessons learned in ISS-IMAP database project

Why did we start it?

- Open data policy was determined in the early stage of the project.

What are difficulties?

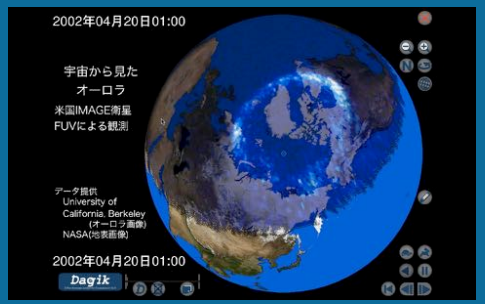
- Data calibration needs longer time than planned.
- Project budget cannot cover whole of the data processing system. Only the system in JAXA was covered.
- The resource in JAXA that can be used for ISS-IMAP data is limited.
- The role of servers in Kyoto Univ. and JAXA should be determined in 2016.

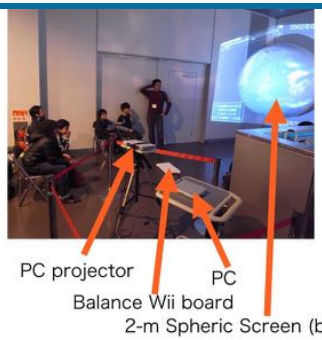
Conclusion

- The “life-plan” of ISS-IMAP database is easy to be determined. Phase-D is promised.

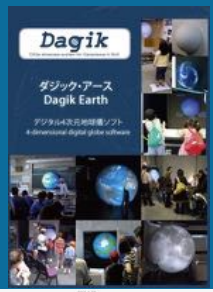
Dagik Earth: Phase-C

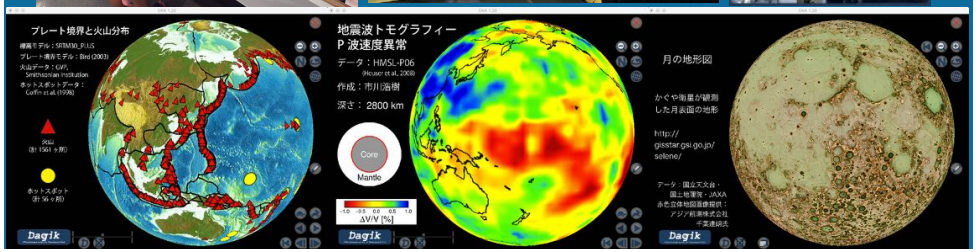
- Database of Earth science data visualisations that are projected on a spherical screen for 3-dimensional presentation.
- About 1,000 registered users. Dagik Earth is used in classrooms and science museums.
- It is supported by the ministry of education, MEXT.





Phase-A, B : 2007-2008 : At first, Google Earth was used. Original software was developed for Windows and Mac. It was used for open campus and exhibition of Kyoto Univ.
Phase-C : 2008- : WWW, Facebook, Mailing List, DVD and news letter.





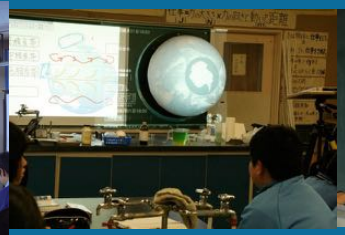
Lessons learned in Dagik Earth database project

Why did we start it?

Nobody did it. Miraikan and NOAA are focused on the large-scale system.

What are good points?

- MEXT support enables us to distribute DVD and manual to more than 3,000 users.
- It is used in many schools. Teachers are developing teaching plans using it in classroom, and hardware in low cost.


Conclusions

- Concept of four phases for “non-data-center” databases
- Phase-D, “Exit”, is important for this eco-system.
- It is good that research institutes and data centers acquire grass-root databases.
- “Cloud” system reduces the hardware maintenance task. However, it is almost impossible to get continuous budget for its long time usage.
- NICT’s Science Cloud and Institutional Repositories could be a good platform for Phase-A (and Phase-B) databases.
- There is Phase-E database in which no new data is added, and can be relocated to any systems.

Lessons learned in Dagik Earth database project (continued)

What are difficulties?

- To be used by school teachers is not easy. They tend to hesitate to use new tools. To be known is the first step.
- International collaboration is not wide enough. In Taiwan, we have collaborators.

Conclusions

- As the database for general public, the way to be known and used is very important.
- Long time plan for the database maintenance is not easy to make. Phase-D should be considered.