GECA: ESA's Next Generation Validation Data Centre

Y.J. Meijer[1], T. Fehr[1], R.M. Koopman[2], A. Pellegrini[1], G. Busswell[3], I. Williams[3], M. De Mazière[4], S. Niemeijer[5], R. van Deelen[5]

[1]ESA/ESRIN, Via Galileo Galilei, 00044 Frascati (RM), Italy, yasjka.meijer@esa.int
[2]GEO Secretariat, Case Postale 2300, CH-1211 Geneva 2, Switzerland
[3]Logica, Keats House, Springfield Drive, Leatherhead, KT22 7LP, United Kingdom
[4]BIRA-IASB, Ringlaan 3, B-1180, Brussels, Belgium
[5]S[&]T, Olof Palmestraat 18, 2616 LR Delft, The Netherlands

ABSTRACT

In the coming decade the availability of satellite data from Earth Observation (EO) platforms will exhibit a significant growth. The dataflow of the Sentinel 1-5 series, contributing to the Global Monitoring for Environment and Security (GMES), will start in 2013. Their flow will be much larger than the one of their preceding satellite missions, like ERS and ENVISAT, now also providing data to GMES. In addition, ESA develops a continuous series of Earth Explorer satellite missions of which currently ADM-Aeolus and EarthCARE focus on atmospheric profiling. As geophysical validation of these EO data remains a high priority, ESA has initiated a project to develop a Generic Environment for Calibration/validation Analysis (GECA), which is considered to become the next generation validation data centre. The evolution part of GECA is in the interoperability between various validation data centres through standardisation of (meta-) data, catalogue and data exchange. In addition, GECA will offer several functionalities facilitating validation analysis with full traceability. One of these functions is the collocation engine which matches satellite data to correlative data and provides the option to download selected sub sets. It will also be possible to compare satellite and correlative data using (best practice) analysis functions either via internet on the dedicated GECA server or locally at the user. Currently data centre interoperability has started with the Aura Validation Data Centre (AVDC), ENVISAT Validation Data Centre (EVDC), EARLINET, GAWSIS, GEOmon and NDACC.

1. INTRODUCTION

This paper describes the "Generic Environment for Calibration/Validation Analysis" (GECA), which is a service and functionality currently under development in an ESA project. It describes the scope, context and objectives of the project. The project consists of two distinct, but connected activities:

1. evolution of ESA's validation data centre,
2. definition of quality information & action protocol.

This project has as prime Logica and is scientifically led by BIRA.

1.1 Data Validation

Geophysical validation of satellite data is required to obtain confidence in the data quality. It involves comparison of satellite data products with independent measurements. Accurate characterisation of diurnal, geographical, seasonal and other dependencies are required. Figure 1 gives an example for GOMOS. Historically, ESA has placed major emphasis on the geophysical validation of Earth Observation (EO) products, in particular for atmospheric chemistry. To support the complex validation process, ESA has set up validation data centres that host the correlative data acquired for calibration and validation of satellite sensors. Its 'first-generation' data centre implemented for the GOME-1 instrument had been derived from the infrastructure hosted by NILU for the Arctic science campaigns organised by the European Commission. Its 'second-generation' data centre has been implemented for the Envisat sensors AATSR, GOMOS, MERIS, MIPAS and SCIAMACHY, and contains specific functionality to further facilitate inter-comparisons and independent analyses during the validation process. Its metadata standard and format guidelines have been an evolution from the initial deliverable of the EU project COSE. The Envisat validation metadata and file structure guidelines [2] have become a *de facto* standard used by several compatible data centres from, among others, NASA (Aura Validation Data Centre), and the European Commission. It has also been endorsed by the Atmospheric Chemistry subgroup of the CEOS Working Group on Calibration and Validation. The current project consists of the definition and implementation of a further evolution, the 'third-generation' validation data centre, providing significantly more specialised functionality in support of the validation process.

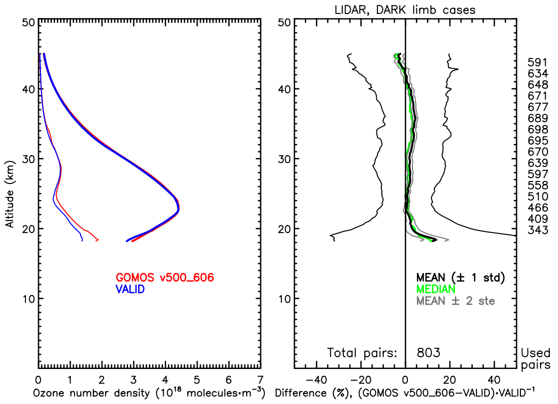

Figure 1. *Example of GOMOS ozone profile validation results in an intercomparison with ground-based lidar data (picture credits: RIVM). Similar results can be found in [1].*

1.2 Using Satellite Data Quality Information

Measurements are only significant if their quality is specified. Traditionally, accuracy of ESA EO data products is assessed during validation campaigns, and recorded in product quality statements. Transient deviations from quality level are described in subsequent quality degradation reports and characterisation of

© Proceedings of the 8th International Symposium on Tropospheric Profiling, ISBN 978-90-6960-233-2
Delft, The Netherlands, October 2009. Editors, A. Apituley, H.W.J. Russchenberg, W.A.A. Monna

complex quality dependencies is reported in cyclic, e.g., monthly or bi-monthly, quality assessment reports. Currently the sources of *a-posteriori* (i.e., formulated after product generation) information are not uniform, and are only available in (textual) documents. Users must interpret this information and adapt their analysis approach and tools to manually perform filtering, adapt the weighting, or correct parameter values. After reprocessing, the users will have to manually undo some of the earlier corrective actions in their code. These operations are error prone and often not traceable. As a result, subjective interpretations enter the objective analysis, adversely affecting the outcome of scientific results derived from ESA products.

It is the purpose of the GECA project to introduce a Quality Information and Action Protocol (QIAP) concept. This will allow the integration of available product quality information with data ingestion software resulting into uniform and automated action. The project includes an implementation for operational ESA missions. The definitions and architecture will be fully generic allowing further proliferation of the standard within the context of EO data harmonisation, e.g., in the frame of GMES.

2. GECA VALIDATION DATA CENTRE (GVDC)

The GECA Validation Data Centre (GVDC) will be a generic correlative data centre with significant evolution steps compared to the traditional validation data centres. The evolution part is in the availability of specific functionalities facilitating the process of validation analysis. One of these functions is the collocation engine which matches satellite data to correlative data and provides the option to download selected sub sets. It will also be possible to compare satellite and correlative data using (best practice) analysis functions either via internet on the dedicated GECA server or locally at the user. More evolution is in the interoperability between various (validation) data centres through standardisation of (meta-) data, catalogue and data exchange offering a wider range of data.

2.1 Generic Correlative Data Centre

Instead of hosting correlative data related to a particular satellite mission, the GVDC will host data in a generic (meta-) data format suitable to support any EO mission and covering multiple domains. In the first phase of the project, the database will be covering data related to ERS and ENVISAT including domains like atmospheric composition, land cover, ocean (waves, Sea Surface Temperature (SST), etc.) and SAR imaging.

GVDC compliant files can be generated with dedicated tools that will be available. The standard that is being developed, the GECA Data Format or short GDF, will be an evolution the Envisat metadata standard 4. Tools will be shared between compliant data centres. As some data providers are used to submit data to more than one data centre, a service for centralised submission is foreseen that, upon request, will automatically distribute files to various databases.

2.2 Data Centre Inter-Operability

Access to a wider range of correlative datasets is foreseen through interoperability. Currently data centre interoperability has started with the Aura Validation Data Centre (AVDC), ENVISAT Validation Data Centre (EVDC), European Aerosol Research Lidar Network (EARLINET), Global Atmospheric Watch Station Information System (GAWSIS), Global Earth Observation and Monitoring (GEOmon) and Network for the Detection of Atmospheric Composition Change (NDACC). In addition, there are plans to register all catalogue metadata with the WMO Information System (WIS), which is in turn connected to the GEOSS clearinghouse. See Table 1 for the full list of data centres.

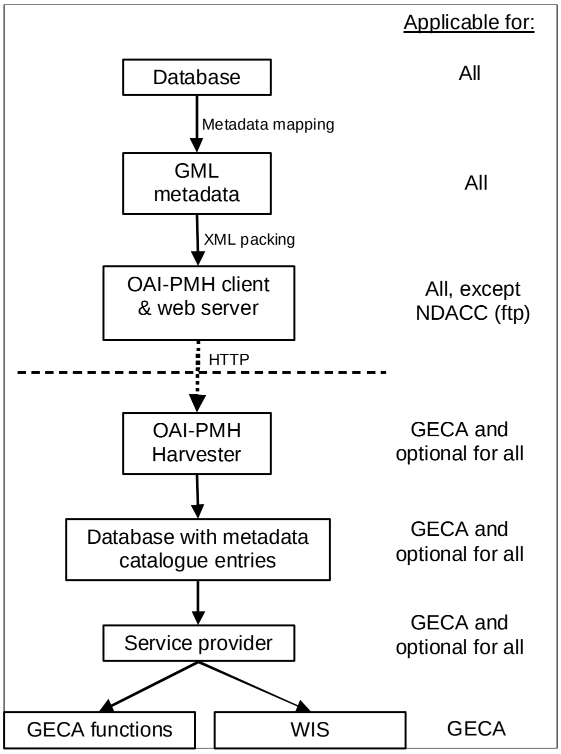

Figure 2. *Data Centre Inter-Operability concept for catalogue-metadata exchange of GECA with other data centres using OAI-PMH. Note that NDACC will make their XML-packed metadata available via FTP without a request functionality.*

Exchange of catalogue metadata between the centres is currently operational as a prototype using the Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH v2, see [3]) (see Figure 2). GVDC will host a metadata catalogue of the data available in the vari-

Table 1. Data centres interoperating with GECA.

Data Centre	Main focus
AVDC (NASA)	Satellite validation
Earlinet (European)	Research and monitoring
EVDC (NILU/ESA)	Satellite validation
GAWSIS (Inter.nat.)	Station information
GeoMON (European)	Monitoring; data exchange/exploitation
GEOSS (Inter.nat.)	Data exchange/exploitation
NDACC (Inter.nat.)	Long-term monitoring/ Support to validation
WIS (Inter.nat.)	Data exchange/exploitation

ous peer data centres to which it is connected. In this way users can query datasets outside the GVDC. In the next phase of the interoperability, some data centres will agree to exchange actual data files (respecting the intellectual property rights). For the other data centres the user will have to access the relevant database directly using personal credentials to obtain the dataset of interest.

2.3 Campaign Management

The GVDC will support campaign managers and Cal/Val coordinators. It will empower the validation teams to review plans for correlative data acquisition and assess their adequacy for their requirements. During or after campaigns they can assess the campaign performance by comparing planned versus actual data acquisition.

2.4 Satellite Data Access

Access using ESA's multi-mission facility interface (MMFI) to satellite-data archives will allow validation scientists an easy way to obtain their required datasets. They will be able to query the product or satellite mission of their interest. The GVDC will perform a query in the relevant metadata using the Heterogeneous Missions Accessibility (HMA) standard [4]. The satellite data subset can then be selected for download for further analysis either via the GVDC server or (locally at) the user. In each of the validation steps (see 2.6), it will be possible to obtain the corrected or pre-processed satellite data in the GECA Data Format. The ESA datasets related to atmospheric profiling are listed in Table 2.

Table 2. *Current and future atmospheric profiling instruments and missions covered by GECA.*

Satellite mission	Main profiling products*
Envisat-GOMOS	O_3, NO_2, NO_3, H_2O, aerosol (strato- and mesospheric)
Envisat-MIPAS	O_3, p, T, H_2O, CH_4, NO_2, N_2O, HNO_3 (strato- and mesospheric)
Envisat-SCIAMACHY	O_3, NO_2, BrO, H_2O, CO, SO_2, AAI, CH_4 (strato- and mesospheric, tropospheric columns)
EarthCARE	Aerosol and cloud properties (tropo- and stratospheric)
ADM-Aeolus	Windspeed, aerosol properties (tropo- and stratospheric)
Sentinel-4	Tropospheric (sub-) columns of trace gases
Sentinel-5 precursor	Tropospheric (sub-) columns of trace gases
Other GECA domain missions/ instruments	AATSR, Cryosat II, ERS-2, MERIS, RA2, (A)SAR, Sentinel 1-3, etc.

* Note that this is a selected list only addressing products providing either profiling or sub-column information, which is of interest to the ISTP community.

2.5 Collocation Engine

A major evolution, compared to previous data centres, is the possibility to search for collocated pairs between satellite and correlative datasets. The GVDC user will be able to set certain basic, and if desired more advanced, criteria matching a satellite measurement under assessment with a correlative dataset. This other dataset can either be:

1. a measurement from another satellite,
2. a measurement from a non-satellite platform (i.e., ground-based, balloon-borne, air-borne, ship-borne or drifting), or
3. related to a fixed site.

The query output result will be presented together with collocation quality information related to the defined criteria. For example, time difference and measurement area overlap (or distance) between the collocated observations. As some queries will require a long computing time, the user can logon at a later time to view the results of the scheduled or repetitive collocation query. The query results can then be viewed and manipulated before deciding which datasets to download for further analysis.

2.6 Analysis Environment

GVDC will allow users to directly perform intercomparisons based on 'best practice' analysis functions. These best practises have been compiled by the scientists within the GECA consortium. This involves work currently performed within the Quality Assurance Framework for EO which, among others, establishes best practices for each community. An assessment of all data handling and manipulation operations was performed that are commonly used during comparative calibration/validation analysis. Based on this, software support functions will be developed. Some examples of common functionalities include conversion to common format, inclusion of auxiliary data (e.g., meteorological parameters from ECMWF), parameter conversion (e.g., mixing ratio to concentration), regridding and smoothing of data. The functions set up for GECA will be based on generic building blocks allowing the GVDC to be expanded to cover data from any new satellite mission.

A very interesting feature of GVDC is the set of preprocessing functions. In the pre-processing stage the data are brought into a comparable state (same unit, same resolution, non-matching data removed, etc.). It should be noted that the data resulting from each preprocessing stage on the server is also available for download by the user and will use as much as possible the GECA Data Format.

The user has the choice to analyse the data either on the dedicated GECA server, which prevents the need of downloading the required datasets and exploits the available computing power, or the user can apply the GECA toolbox locally on downloaded data. The toolbox consists of the same building blocks as available on the server. An intercomparison performed locally at the user is executed as a series of command line executables with required input parameters. GECA will maintain full traceability of the whole validation process improving the credibility of EO products.

On the GECA server the user will be presented with the option to generate a report for each intercomparison. These reports have a default setup, related to the intercomparison type, which can be customized by

- adding/removing report components, and
- setting parameters for each report component (see Table 3 for optional list).

Optionally the user can choose to retrieve ASCII files with numerical data for each report component, which can then be used for further (local) processing or for ingesting them into other plotting programs.

Table 3. *Optional report components which can be generated using the analysis functions of GECA.*

Report components
Plots:
Measurement location plot
XY-scatter plot
Generic X,Y plot or one of the following special cases: - Time-series plot - Plot of vertical profiles - Plot of averaging kernel matrices - Plot of difference in parameter vs other parameter
Histogram
Raster image plot
Surface plot
Tables with statistics, composed of:
max, mean, median, interquartile range, percentage within range, standard deviation, median absolute deviation, count, variance, root mean square.

3. QUALITY INFORMATION & ACTION PROTOCOL (QIAP)

The Quality Information and Action Protocol (QIAP) shall enable electronic transmission of information on product quality. The validation support functionality developed in the context of the GVDC shall include retrieval of quality information and actions, and further processing of this information for validation analysis.

Initially the QIAP implementation will target data from operational ESA missions, and not correlative validation data. Nevertheless, the concept will be sufficiently generic that it can also be applied to future ESA missions and possibly also non-satellite data.

4. CONCLUSION

The project is currently passed its system requirements review and has now started with the first phase of the implementation. An initial version of GECA will be operational early 2010, which will allow validation analysis of selected data sets from ERS and ENVISAT. By the end of 2010 the system will be fully operational and accessible for external users.

REFERENCES

[1] Van Gijsel, J. A. E., D. P. J. Swart, J. -L. Baray, H. Claude, T. Fehr, P. Von Der Gathen, S. Godin-Beekmann, G.H. Hansen, T. Leblanc, I.S. McDermid, Y.J. Meijer, H. Nakane, E.J. Quel, W. Steinbrecht, K.B. Strawbridge, B. Tatarov, E.A. Wolfram, 2009: Global validation of ENVISAT ozone profiles using lidar measurements, *International Journal of Remote Sensing*, **30**, 15, pp. 3987—3994.

[2] Bojkov, B.R., M. De Mazière, R.M. Koopman, 2002: Generic Metadata guidelines on atmospheric and oceanographic datasets for the Envisat Calibration and Validation Project, ESA document, pp 1-65.

[3] OAI, 2002: The Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm.

[4] OGC, 2006: OGC Catalogue Services Specification 2.0.0 – EO Application Profile for CSW 2.0 version 0.1.4.

ACKNOWLEDGEMENTS

The authors are grateful for the contributions of the scientific partners in the consortium of the GECA project. These partners are from the following institutes: BIRA-IASB, CLS, CLS (BOOST), KNMI, RAL and University of Leicester.