Toward a normalized XML Schema for the GGP data archives

Alban Gabillon, J.P. Barriot, Youri Verschelle, Bernard Ducarme

TOWARD A NORMALIZED XML SCHEMA FOR THE GGP DATA ARCHIVES

Alban Gabillon[1*], Jean-Pierre Barriot[1], Youri Verschelle[1] and Bernard Ducarme[2]

[1]Université de la Polynésie Française. BP6570, 98702 Faa'a. French Polynesia*
Email: {alban.gabillon,youri.verschelle,jean-pierre.barriot}@upf.pf
[2]Royal Observatory of Belgium, Av. Circulaire 3, B-1180 Brussels, Belgium
Email: bf.ducarme@gmail.com

ABSTRACT

Since 1997, the Global Geodynamics Project (GGP) stations use a text-based data format. The main drawbacks of this type of data coding is the lack of data integrity during the data flow processing. As a result, metadata and even data must be checked by human operators. We propose in this paper a new format for representing the GGP data. This new format is based on the eXtensible Markup Language (XML).

Keywords: GGP data, XML schema

1 INTRODUCTION

Since 1997, GGP stations use a text-based data format known as PRETERNA. The main drawbacks of this type of data coding is the lack of data integrity during the data flow processing. As a result, metadata and even data must be checked by human operators. We propose in this paper a new format for storing and disseminating the data coming from the worldwide GGP network of superconducting gravimeters, in order to streamline the data processing and to enable the scientific community to access these data and their ancillary metadata through distributed, integrated information technology systems and virtual observatories. This new format is based on the eXtensible Markup Language (XML, Bray et al., 2006) that ensures the consistency, reliability and integrity of the data over the Internet and between any data processing platforms. Section 2 of this paper reviews the GGP network of superconducting gravimeters, section 3 outlines the main drawbacks of the current text-based GGP data format. Section 4 presents our new data format based on an XML *schema* (Thompson et al., 2004). Section 5 concludes this paper.

2 THE GGP NETWORK OF SUPERCONDUCTING GRAVIMETERS

The Global Geodynamics Project (GGP) is an international network of 25 superconducting gravimeters (Crossley et al., 1999) in operation since July1997, under the umbrella of the International Association of Geodesy (IAG). The continuous monitoring of timevariable gravity from seconds to years is a tool to investigate many aspects of global Earth dynamics and to contribute to other sciences such as seismology, oceanography, earth rotation, hydrology, volcanology, and tectonics. Another promising application is the use of SG subnetworks in Europe and Asia to validate time-varying satellite gravity observations (GRACE, GOCE) due to continental hydrology and large-scale seismic deformation. GGP plays a small but important role in the Global Geodetic Observing System (GGOS), a primary program of the IAG to coordinate the recording and dissemination of all geodetic data for Earth monitoring, namely the recording of the gravity field and especially its time variations (Crossley & Hinderer, 2009). GGP was incorporated into the IAG as Inter-Commission Project #3.1 in 2003; it is a joint project between Commission 3 (Earth Rotation and Geodynamics) and Commission 2 (The Gravity Field). It is expected to become a full Service of IAG in 2014.

3 THE CURRENT GGP DATA FORMAT

All GGP stations use the data format proposed by Wenzel (1996), known as PRETERNA, in which every value (predominantly gravity and pressure), are time tagged in the original units (volt).The only processing is a decimation filter from the original samples to 1-minute values, but no other corrections are done. The full signal is saved with a precision of 7.5+ digits, ensuring that the tides are adequately recorded as well as the smallest

tidal waves. A full discussion of data treatment is given in Hinderer et al. (2007). Users should realize that gaps, spikes and offsets still have to be treated if a clean continuous time series is required, or otherwise avoided if the series is processed as non-contiguous blocks. These 1-minute raw data files are stored at GFZ Potsdam (http://isdc.gfz-potsdam.de/). The International Center for Earth Tides, a Service of IAG, provides corrected minute data (i.e. manually cleaned for gaps, spikes and offsets) on their website (http://www.bim-icet.org/), but this treatment is designed for tidal analysis and may not be suitable for all purposes, especially long period studies. A GGP 1-minute file is a column-driven file made up of 2 sections, each section being subdivided into 3 parts:

1. The header
1.1 - first ten required lines (ancillary information about the GGP station and instrument)
1.2 – optional text lines inserted by SG group (free comments)
1.3 – two required text lines

2. The data
2.1 – one required introductory line
2.2 – lines of timetagged gravity and pressure data
2.3 – last required termination line

An example of data file is given in Table 1, and the complete data format descriptor (last updated 10 December 2008) is available for download at http://www.eas.slu.edu/GGP/ggpnews19a.pdf. This format, in use since 1997, is based on Hollerith punched cards style formats, as FORTRAN character fields (A descriptor), integer fields (I descriptor) and float fields (F descriptor). The main drawbacks of this type of data coding is the lack of data integrity during the data flow processing as described at http://www.eas.slu.edu/GGP/ggpnews5.pdf, and the lack of a strict enforcement of data field lengths. As a result, metadata and even data must be checked by human operators. Moreover, this data format includes text-based tags like `77777777` or `99999999` without implicit semantics.

Table 1. Current GGP data format

```
Filename          :  H2050300.GGP
Station           :  Bad Homburg, Germany
Instrument        :  GWR CD030_U
Time Delay (sec)   :  45.0      2.0    estimated
N Latitude (deg)   :  50.2285    0.0001 measured
E Longitude (deg)  :   8.6113    0.0001 measured
Elevation MSL (m)  : 190.0000    0.1000 measured
Gravity Cal (uGal/V): -67.92      0.02   measured
Pressure Cal (hPa/V):   1.0      0.001  nominal
Author            :  P. Wolf (peter.wolf@bkg.bund.de)
yyyymmdd hhmmss gravity(V) pressure(V)
C*********************************************************
77777777       0.0       0.0
20050301 000000 -0.504559 993.78749
20050301 000100 -0.502637 993.79867
20050301 000200 -0.500711 993.81193
..
20050320 042800 -1.1410631001.19516
20050320 042900 -1.1415471001.19009
20050320 043000 -1.1420611001.18142
99999999
77777777       0.0       0.0
20050320 161100 -0.151548 998.28556
20050320 161200 -0.146616 998.29147
20050320 161300 -0.141674 998.30143
..
20050331 235700 -0.8851071004.02740
20050331 235800 -0.8876941004.03534
20050331 235900 -0.8902831004.04113
99999999
```

4 THE NEW XML DATA FORMAT

Writing GGP files in XML has several advantages:

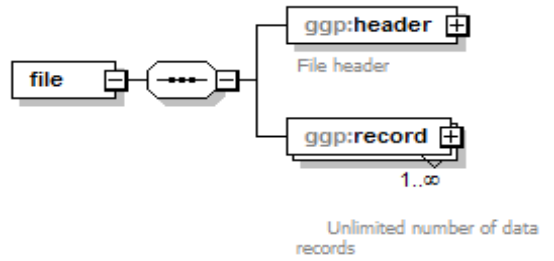
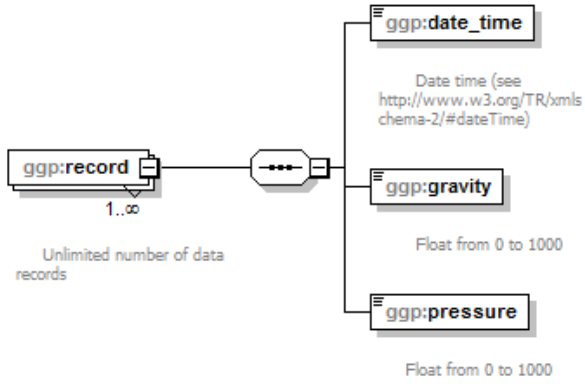
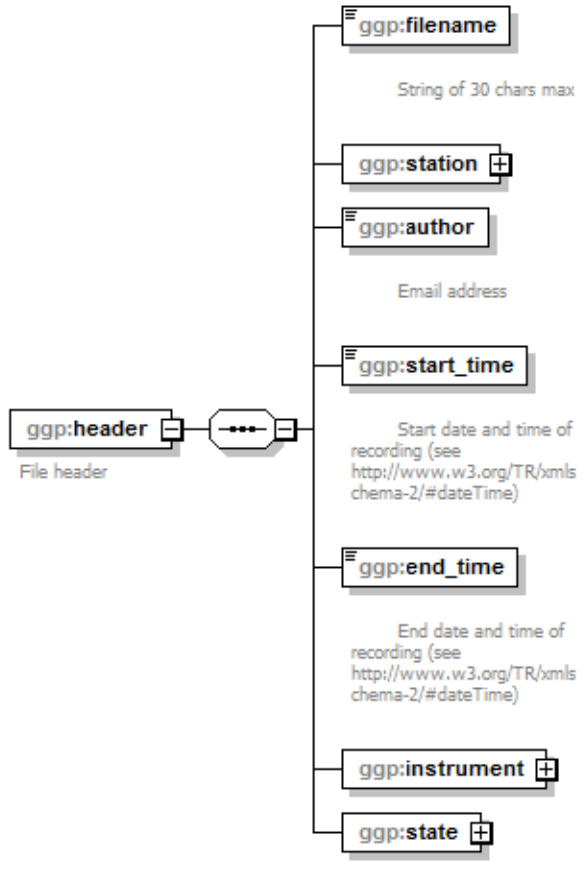
- XML is a markup language. Data fields are clearly separated by *tags*.
- Since tags are user-defined (XML is not restricted to a predefined limited set of tags like HTML), tags convey semantics specific to the application domain.
- XML files can be automatically analyzed for data treatment/presentation with an XML *parser*.
- An XML file can be checked against an XML schema. An XML schema is a special XML file that specifies a vocabulary identified by a *namespace* (Bray et al., 2006), and some grammatical rules. An XML file that respects the rules dictated by a particular XML schema is said to be *valid*. Checking the validity of an XML file is an *automated* process.

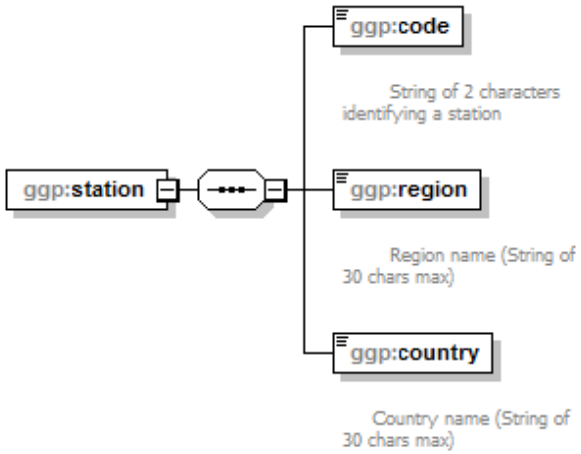
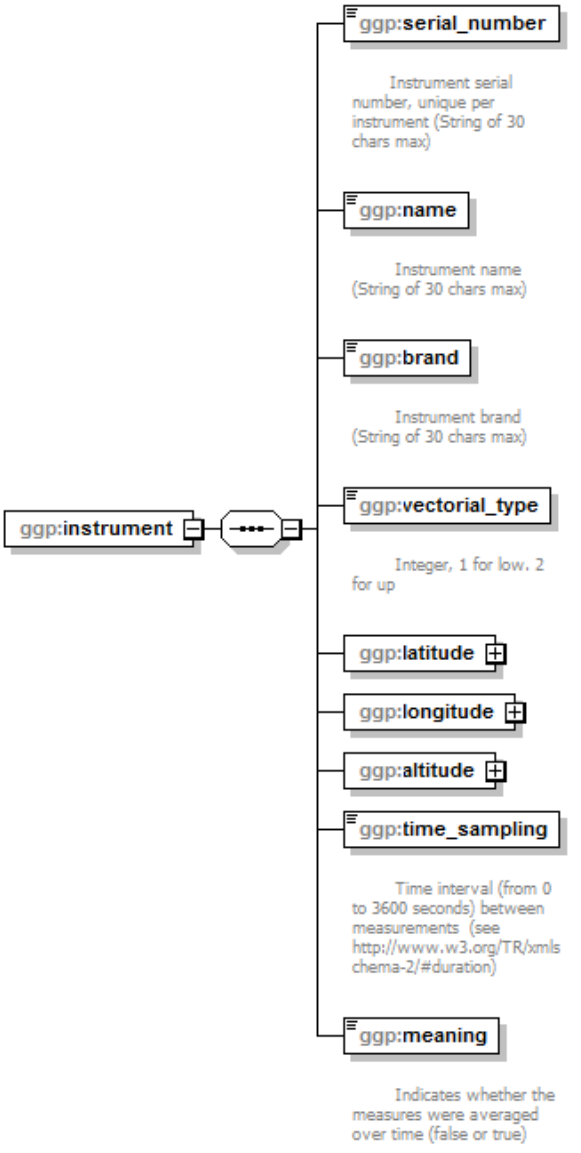
In this section, our objective is to propose an *XML GGP schema*. Our XML GGP schema defines the legal building blocks of the XML GGP files. Our XML GGP schema defines its own namespace identified by the GGP web page URL: http://www.eas.slu.edu/GGP/ggphome.html. The schema itself can be accessed at the following URL: http://pages.upf.pf/Alban.Gabillon/ggp/ggp.html**.**

Our schema is described in table 2. Regarding this description we can make the following comments:

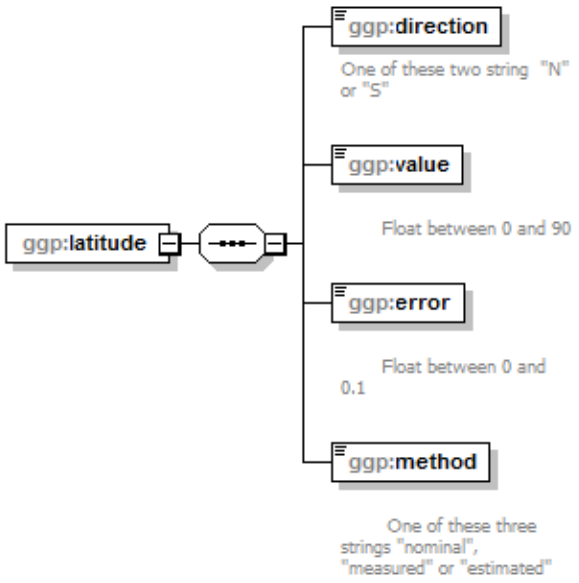
- Our schema is a preliminary version of what should become a normalized XML GGP schema officially approved by the IAG.
- Sample GGP files should can be validated online by using the W3C validation service: http://validator.w3.org/
- Our schema uses the standardized W3C built-in data types (Biron & Malhotra, 2004).
- We are planning to improve our schema by referring to already official schemas and vocabularies defined by international organizations like the Open Geospatial Consortium (OGC) (http://www.opengeospatial.org). Such already existing vocabularies could be used to define some concepts like latitude, longitude etc.
- We are also planning to refer to the Sensor Model Language (SensorML) that provides standard models and an XML encoding for describing the process of measurement by sensors and instructions for deriving higher-level information from observations (Botts & Robin, 2007).
- Checking the validity of a time series and its associated metadata can be done statically from the corresponding XML GGP file. It can also be done dynamically during the data flow processing.

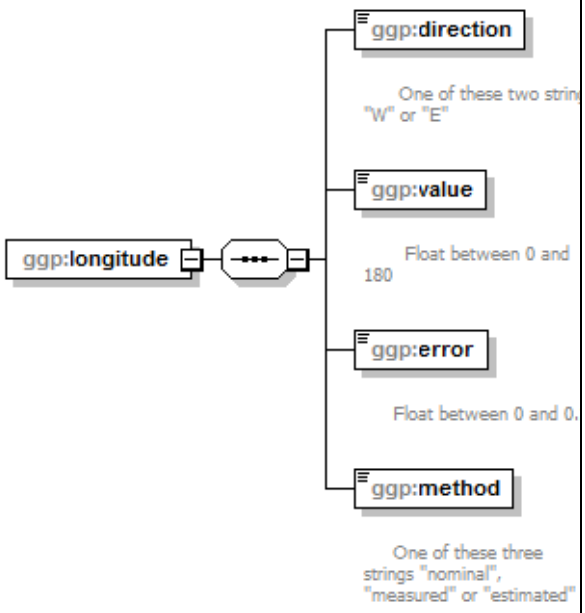
Table 2. Schema GGP.xsd Description

	Our schema divides GGP files into 2 blocks: The *header* which consists of a set of header fields and the data block which corresponds to the *time series* and which consists of an unbounded number of data *records*.
	Each data *record* consists of 3 fields. The first field records the *date* and *time* of the measure in the format specified by the W3C. The second field is the *gravity* measure. The last field is the *pressure* measure. Specified bounds (from 0 to 1000) correspond to physical limitations.
	The header includes several data fields, *filename*, *author*, *start_time* and *end_time*. *filename* and *author* are self described. *start_time* correponds to the date and time of the first data record, *end_time* corresponds to the date and time of the last data record. Other fields (*station*, *instrument* and *state*) are complex elements containing nested subfields. *instrument* contains some data regarding the instrument that recorded the time series. *station* contains some data referring to the station which hosts the instrument. *state* contains some data which are specific to the version of the time series obtained after a given processing step. Indeed, a given time series can follow a processing chain and each step in the processing chain outputs a new state of the time series.

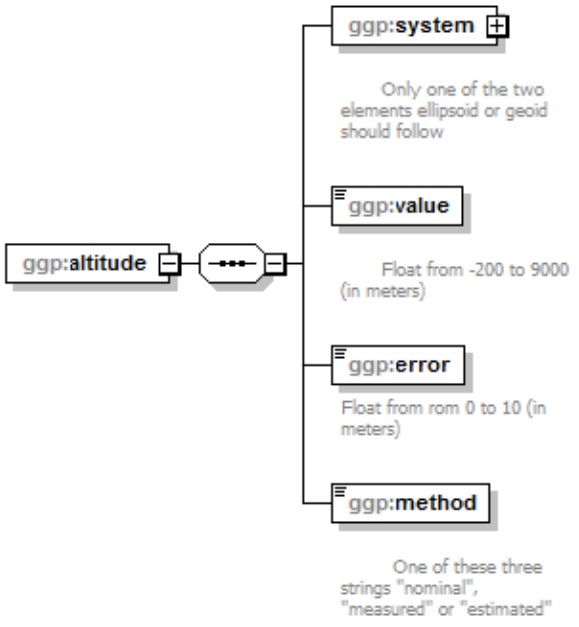
	The station element contains 3 fields *code*, *region* and *country* identifying and locating the station.
	The instrument element contains several fields and complex elements describing the instrument which produced the time series. *serial_number*, *name*, *vectorial_type* and *brand* are the simple data fields whereas *latitude*, *longitude* and *altitude* are the complex elements containing nested subfields. All these instrument parameters remain unchanged over the various versions (states) of a given time series. *serial_number*, *name* and *brand* are self described. *vectorial_type* indicates the levitating ball (1 for low, 2 for up). *latitude*, *longitude* and *altitude* contain some data about respectively the latitude, the longitude and the altitude of the instrument at the time it produced the time series. The latitude element (respectively the longitude element) contains some self described fields related to the latitude (respectively the longitude) of the instrument.

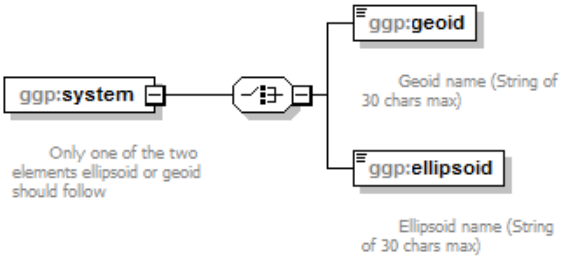
The latitude (respectively longitude) element contains some self described data fields related to the latitude (respectively longitude) of the instrument (see below)

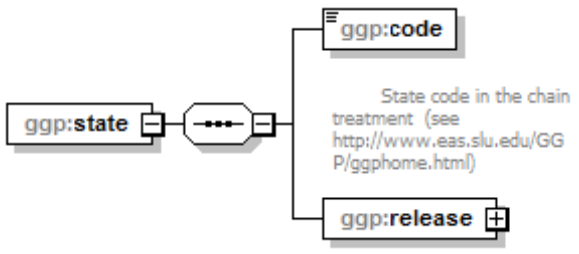




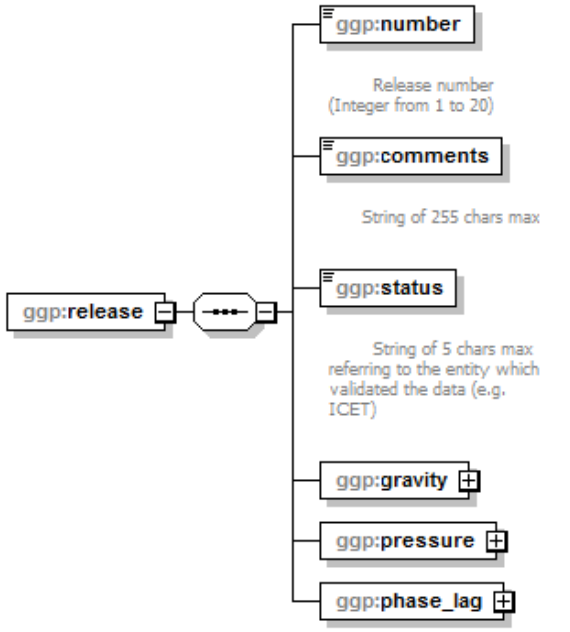
The altitude element contains some self described fields related to the altitude of the instrument. Note that *system* is a complex element which *either* includes a *geoid* data field or an *ellipsoid* data field (see below).





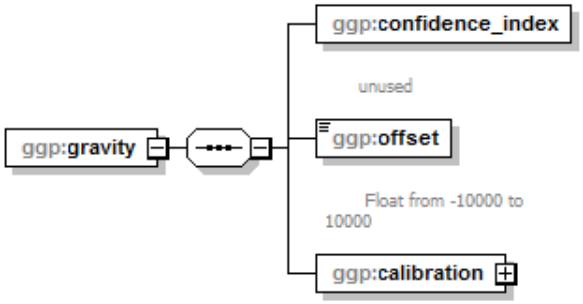


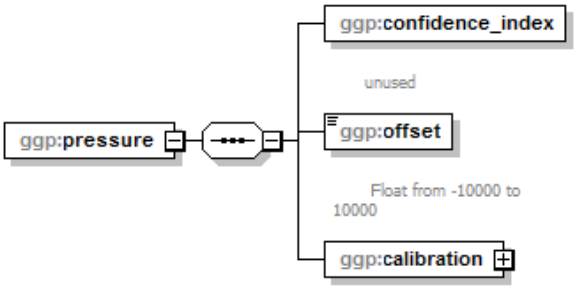
The state element consists of one data field *code* and one complex element *release*. *code* identifies the chain processing step (see http://www.eas.slu.edu/GGP/ggphome.html). Each state (i.e. version of the time series) can be subject to several releases (at least one). *release* records some data which are specific to each release.

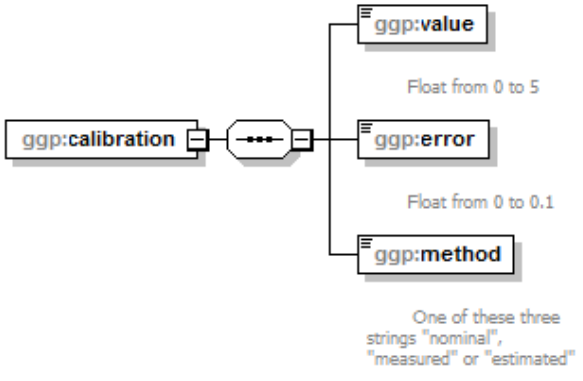


The release element contains several fields and complex elements describing some parameters which are specific to a given release of a time series state. *number*, *comments* and *status* are self described. Other fields (*gravity*, *pressure* and *phase_lag*) are complex elements containing nested subfields. They all contain some specific instrument parameters which were specifically set for the release.

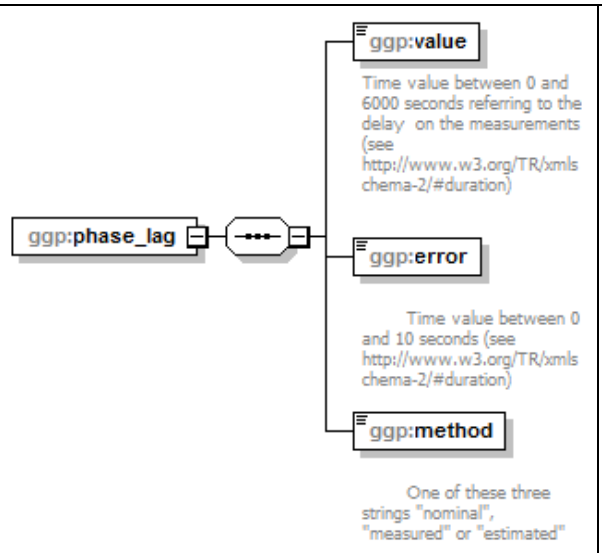
The gravity (respectively pressure) element (see below) contains 2 data fields and 1 complex element. *confidence_index* is unused and *offset* (float from –10000 to 10000) indicates a general offset on gravity for the considered data. Note that, contrary to the previous format, there should be only one possible offset value for each time series. *calibration* is the complex element and contains nested subfields (see below).







The calibration element which is nested in both the gravity and the pressure element contains 3 data fields *value*, *error* and *method*. These fields are self described.

	The phase_lag element contains 3 data fields *value*, *error* and *method*. These fields are self described.

5 CONCLUSION

We hope that the format we proposed in this paper will serve as a base for the future official GGP data format. We are currently developing a toolbox to allow easy back and forth conversion between the old and our new xml format. We are also writing several XSLT (Kay, 2007) style sheets for visualization of the XML GGP data.

6 REFERENCES

Biron, P. V., & Malhotra, A., (2004). XML Schema Part 2: Datatypes Second Edition. W3C Recommendation 28 October 2004. Retrieved March 3, 2012 from the World Wide Web: http://www.w3.org/TR/xmlschema-2/

Botts M., & Robin, A., (2007). OpenGIS Sensor Model Language (SensorML) Implementation Specification. 2007-07-17. 2007-07-17. OGC 07-000. Retrieved March 3, 2012 from the World Wide Web: http://www.opengeospatial.org/

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau,F., & Cowan., J. (2006). Extensible Markup Language (XML) 1.1 (Second Edition). W3C Recommendation 16 August 2006, edited in place 29 September 2006. Retrieved March 3, 2012 from the World Wide Web: http://www.w3.org/TR/xml11

Bray T., Hollander, D., Layman, A., & Tobin, R. (2006). Namespaces in XML 1.1 (Second Edition). W3C Recommendation 16 August 2006. Retrieved March 3, 2012 from the World Wide Web: http://www.w3.org/TR/xml-names11

Crossley, D., & Hinderer, J., (2009). The Contribution of GGP Superconducting Gravimeters to GGOS. In Sideris, M.G., (Ed), *Proceedings of the IUGG, IAG Symposia 133, Perugia 2007, Observing our Changing Earth*: Springer Verlag.

Crossley, D., Hinderer, J., Casula, G., Francis, O., Hsu, H. T., Imanishi, Y., Jentzsch, G., Kääriaïnen, J., Merriam, J., Meurers, B., Neumeyer, J., Richter, B., Shibuya, K., Sato, T. & Van Dam, T., (1999). Network of superconducting gravimeters benefits a number of disciplines, *EOS*, 80, 11, 121/125-126.

Hinderer, J., Crossley D., & Warburton, W., (2007). Superconducting Gravimetry. In Herring T. and Schubert G., (Eds), *Treatise on Geophysics*, Vol 3, Elsevier.

Kay M. (2007). XSL Transformations (XSLT) Version 2.0. W3C Recommendation 23 January 2007. Retrieved March 3, 2012 from the World Wide Web: http://www.w3.org/TR/xslt20/

Thompson, H.S., Beech, D., Maloney, M., & Mendelsohn N., (2004). XML Schema Part 1: Structures, Second Edition. W3C Recommendation 28 October 2004. Retrieved March 3, 2012 from the World Wide Web: http://www.w3.org/TR/xmlschema-1/

Wenzel, H.G. (1996). The nanogal software: Earth tide processing package ETERNA 3.30. *Bull. Information des Marées Terrestres*, 124, 9425-9439.