Grid-Enabling the Global Geodynamics Project: The Introduction of an XML-Based Data Model

L. Ian Lumb
Platform Computing Inc.
3760 14th Avenue
Markham, Ontario, Canada L3R 3T7
Email: ilumb@platform.com

Keith D. Aldridge
Earth & Space Science and Engineering
York University
Toronto, Ontario, Canada M3J 1P3
Email: keith@yorku.ca

***Abstract*— The Global Geodynamics Project (GGP) provides a reasonable representation of the scientific collaboration evident in small-to-medium-scale initiatives. GGP also provides data management challenges that are different from those typically expressed by other areas of the physical sciences - e.g., High Energy Physics. These distinctions make GGP an interesting candidate for assessing the challenges and opportunities associated with technically enabling collaborative science via Grid Computing. Emphasis is placed here on the introduction of an XML-based data model into the GGP. Although it is concluded that Earth Sciences Markup Language (ESML) is highly effective and efficient in introducing a new data model, and paves the way for structural transformations on data, challenges and opportunities are also identified. Metadata (data about data) provides the gravest concern and therefore the most-important focus for further research.**

Citation—**Lumb, L. I. and K. D. Aldridge, Grid-Enabling the Global Geodynamics Project: The Introduction of an XML-Based Data Model, in** *Proceedings of The 19th International Symposium on High Performance Computing Systems and Applications, HPCS 2005*, **I. Kotsireas and D. Stacey (editors), The IEEE Computer Society, 216–222, 2005.**

I. Introduction

The Global Geodynamics Project (GGP) was established to allow Earth scientists the ability to leverage a network of globally distributed instruments for operational and research activities into Earth tides ([1], [2]). Now in its second phase, the GGP is proactively engaging non-traditional disciplines - i.e., those outside its original Earth tides community. For example, Lumb & Aldridge ([3]) seek to better understand Earth's rotational spectrum at periods of about a half-day, and the potential role of rotationally induced responses in generating and sustaining Earth's magnetic field. As another example, whose impact is underscored by the recent, devasting 26 December 2004 Sumatra-Andaman earthquake ([4]) and resulting tsunami ([5]), seismologists seek to make use of GGP data to better predict, catalog and interpret seismic activity (e.g., [1]). Even with this compelling interest in GGP data, geodynamicists, seismologists, and others, are faced with practicalities which inhibit their engagement as 'non-specialists' ([6]). For example:

- Temporal and/or spatial alignment of GGP data is challenging - The requirement to correlate data, in time and space, is currently a very manual process that requires geodynamicists and seismologists to specify temporal (e.g., a period of time, an event in time) and/or spatial (e.g., global, regional, specific instruments) details to allow for further analysis.
- There are undesireable, yet significant signals to contend with - To geodynamicists and seismologists, tidal, atmospheric, hydrologic and oceanic signals are all unwanted. This means that the processed GGP data must undergo further, non-trivial reductions before it is useful for geodynamic and seismic purposes. Traceable and reproducible reductions are critical as research efforts often involve complex modeling close to ambient noise levels of the instruments involved.

Typically, science employs technology for a purpose. However, when the technology is itself in its infancy, a reciprocity exists - i.e., the scientific use can shape the evolution of the technology. This is precisely the current case with Grid Computing - i.e., intersections between it and (in this case) the GGP have the potential for this reciprocity. Because so much of Grid Computing has been stimulated by the 'Big Science' needs ([7]) of High Energy Physics, this is a crucial juncture to motivate requirements from a broader scientific base that represents different disciplines as well as small-to-medium-scale science ([6]). Phrased bluntly, the applicability of Grid Computing in the case of small-to-medium-scale science comprises one of the key drivers for this investigation from the technology perspective.

With respect to Grid-enabling the GGP, Lumb & Aldridge ([6]) concluded that:

- Leveraging GGP as it exists today is a key consideration - This is especially true for GGP instrumentation and data standards plus bilateral agreements.
- There are numerous opportunities for Grid-enabling the GGP - These opportunities range from instrumentation to data to analysis to end users.

The purpose here is to initiate progress opposite the first of the scientific motivators - i.e., addressing the challenge of temporal and/or spatial alignment. Addressing this motivator also has the desireable side effect of initiating progress on one of the identified opportunities for Grid-enabling the GGP. This investigation is organized into five sections in addition to this

introductory section. The following section (§II) provides an overview of the existing GGP data model. Before introducing a data model for the GGP based on the Earth Sciences Markup Language (ESML) in §IV, a generic evaluation of XML-based data models is provided in §III. The application of ESML to the GGP underlines and amplifies the role of metadata in the GGP, and Earth Sciences in general. Metadata in the context of the GGP, and in general, receives consideration in §V. This paper closes (§VI) with conclusions derived from the current study, and identifies prospects for further research.

II. Existing GGP Data Model

The GGP began as an international consortium of Earth scientists involved in operational and research activities into Earth tides (e.g., [1], [2]). Although Earth tides have a substantial and colorful history, dating back centuries in the use of primitive tide gauges, modern efforts make use of ultrasensitive Superconducting Gravimeters (SGs). In general, gravimeters measure relative variations in Earth's acceleration due to gravity using a variation on the theme of a mass-spring balance. In the case of the SG, however, the mass is a niobium sphere whose spring is superconductivity-induced levitation ([8]). Because many Earth scientists were already involved with organizations having their own SGs, GGP's initial role was to facilitate the creation of a globally distributed network of these instruments ([1]). By establishing standards around SG instrumentation and data, in concert with various bilateral agreements, the GGP ensured progress without having to compromise scientific and organizational integrity.

GGP data is described in detail elsewhere ([9]). To summarize the salient points:

- On a monthly basis, each site reports gravity (via the SG), pressure (via a barometer), and a number of environmental parameters that are captured in three files - an example of a typical file for gravity and pressure is provided in Listing 1.
- These three files share a common filename but differ in their filename extensions. Collectively they are referred to as "GGP Data".
- Each of the three files follows an ASCII-based format specific to Earth tides.

Listing 1. Example GGP data for gravity and pressure ([9]).

```
Filename                    ST970910.GGP
Station              Strasbourg , France
Instrument                      GWR C026
Phase Lag (deg/cpd)     0.1500   0.0100   nominal
N Latitude (deg)       48.6220   0.0010   estimated
E Longitude (deg)       7.6840   0.0010   estimated
Height (m)            180.0000   1.0000   estimated
Gravity Cal (mgal/v) -792.0000   1.0000   measured
Pressure Cal(mbar/v)  200.0000   1.0000   nominal
Author              (jhinderer@eost.u-strasbg.fr)
yyyymmdd hhmmss    gravity(V) pressure(V)
C****************************************
77777777
19970901 000000   0.075913   0.420192
...
```

The intention here is to introduce an XML-based data model to the GGP. In other words, the objective is to introduce an XML-based representation for GGP data as described above and shown in Listing 1. Before presenting the specifics of this representation (§IV), the results of a generic evaluation of XML-based data models is presented in the next section (§III).

III. XML-Based Data Models

Riding on the success of HTML (HyperText Markup Language) on the World Wide Web, the eXtensible Markup Language (XML) continues to experience rapid adoption. This is understandable because XML ([10]):

- Creates application-independent documents and data - It can be processed by any application, yet is human-readable.
- Has a standard syntax for metadata - It effectively describes the structure and purpose of data.
- Has a standard structure for both documents and data - It organizes data into a hierarchy.
- Is not a new technology - It is a subset of 30-year-old Standard Generalized Markup Language (SGML).

Before focusing on data representation via XML, it's important to note some of the complimentary XML-related offerings:

- XML Path Language (XPath) - An expression syntax used to create location paths.
- XML Query Language (XQuery) - A comprehensive data query language that works with XPath, and is analogous to the Structured Query Language (SQL) used in traditional database contexts. Notably, XQuery searches can span multiple XML repositories in a single query statement ([11]).
- eXtensible Stylesheet Language Transformations (XSLT) - A vehicle for structural (i.e., the conversion of one XML document type into another) and aesthetic (i.e., the formatting of one XML document type into another) transformation.
- XML Inclusions (XInclude) - A relatively new vehicle ([12]) for merging XML documents into a single, composite XML document.

XML Document Type Definitions (DTDs) represent an established, traditional approach to XML data representation and validation ([11]). Although XML DTDs aren't completely obsolete, they have been superceded by XML Schema Definition Language (XSD) for the following reasons ([11]):

- Industry support increasingly favors XSD.
- Support for multiple data types and highly customizable validation rules is only present in XSD.
- XSD can handle requirements for complicated data representations and integrates well with relational databases.
- XSD is natively supported by SOAP - SOAP is the *de facto* standard messaging protocol for Web Services.
- XSD is highly extensible.
- Data-binding libraries increasingly leverage XSD.

Of course, XSD has its own limitations around ([11]):

- Conditional constraints.
- Inter-element dependencies.
- Cross-document validation.
- Null values for attributes.
- Validation of large numeric data values.

The XSD specification is currently being revised, so it's likely that many of these issues will be resolved in due course.

The above issues with XSD, the inherent verbosity of XML, along with the key requirement of effectively representing scientific data, lead to a short list of three candidates - BFD, BinX and ESML. (A recent and comprehensive assessment is provided in [13].)

Binary Format Description (BFD, [14]) is based on the eXtensible Scientific Interchange Language (XSIL, [15]). It was developed as part of a Scientific Annotation Middleware (SAM) project at the United States, Department of Energy's Pacific Northwest National Laboratory (PNNL, [16]). Although BFD does have native support for ASCII-format data files, its dependency on XSIL also implies a dependency on XML DTDs.

BinX ([13], [17]) is a language with supporting tools and a library for describing binary data. It was developed as part of the eDIKT Project ([18]) at the United Kingdom's National e-Science Centre ([19]). In striking contrast to BFD, BinX does not have support for ASCII-format data files,[1] but is based on XSD.

Earth Sciences Markup Language (ESML, [20]) is another XML description system which has been developed at the Information Technology and Systems Center University of Alabama in Huntsville ([21]). ESML supports ASCII-format files *and* is based on XSD. In addition, ESML ([22]):

- Is a specialized markup language for Earth Science metadata based on XML.
- Offers a machine-readable and -interpretable representation of the structure of any data file, regardless of data format.
- Has description files that contain external metadata that can be generated by either the data producer or data consumer (at collection, data set, and/or granule level).
- Provides the benefits of a standard, self-describing data format without the penalty of data conversion.
- Is the basis for core interchange technology that allows data/application interoperability.

All of the above makes ESML the most-appropriate vehicle for representing GGP data.

Before applying ESML to GGP data (§IV), it's worth noting that BFD, BinX and ESML are all working towards compliance with the emerging Data Format Description Language (DFDL, pronounced 'daffodil', [23]) specification(s) - a standards activity under the auspices of the Global Grid Forum ([24]). Together, the members of the DFDL Working Group are building on BFD, BinX, ESML and other work to provide a general and extensible platform for describing data formats ([13]).

[1]BinX support for ASCII-format files is anticipated in 2005 (BinX Support, private communication).

IV. ESML-Based Data Model

ESML emerged (§III) as the most-viable choice for introducing an XML-based data model into the GGP. Listing 2 makes use of this choice in representing the GGP data for gravity and pressure provided in Listing 1. Details of the ESML Schema are available elsewhere ([25]). Here we note:

- Use of XML namespaces and schema instances - Explicit reference is made to ESML-specific namespaces and schema instances in addition to standard references.
- Built-in support for ASCII-format data - Along with binary format and several Earth-science-specific formats, ESML provides comprehensive support for syntactic (or structural) metadata in ASCII. This ASCII support includes:
  - In-file data structuring - In Listing 2, the entire file is cast as a single, logical `Structure`.
  - Multidimensional arrays - Repeated, and often nested, use of the simple `Array` element permits `Header` and `Field` data to be represented effectively in a variety of pre-defined formats (e.g., string, integer, floating-point, etc.). In the case of GGP data, an additional step is required to bound one of the `Array` elements. This is indicated as `FROM_PREPOCESSOR` in Listing 2. This is understandable from the perspective of ESML, and therefore should not be regarded as a weakness. In other words, there exists reasonable expectation to bridge data models through a preprocessing step.
- Rules-based constructs - Illustrated in the context of field-embedded sentinnel values (see e.g., `DataGapCode` in Listing 2), occurrence matching provides an example of a rules-based construct provided with ESML. Support for conditional constructs is also highly desireable in this context.

Listing 2. ESML representation of the GGP gravity and pressure data presented in Listing 1

```
<?xml version="1.0" encoding="UTF-8"?>
<ESML xmlns="ESML" xmlns:xsi="http://www.w3.org↘
→/2001/XMLSchema-instance" xsi:schemaLocation="↘
→ESML C:\MyFiles\...\Schema\ESML.xsd">
 <SyntacticMetaData>
  <Ascii>
   <Structure instances="1">
    <Array occurs="2">
     <Header name="_Filename" format="%20s" />
     <Header name="Filename" format="%20s" />
    </Array>
    <Array occurs="2">
     <Header name="_Station" format="%20s" />
     <Header name="Station" format="%20s" />
    </Array>
    <Array occurs="2">
     <Header name="_Instrument" format="%20s" />
     <Header name="Instrument" format="%20s" />
    </Array>
    <Array occurs="4">
     <Header name="_PhaseLag" format="%20s" />
```

```
    <Header name="PhaseLag" format="%10.4f" />
    <Header name="PhaseLagError" format="%10.4f"↘
    → />
    <Header name="PhaseLagComment" format="%20s"↘
    → />
   </Array>
   <Array occurs="4">
    <Header name="_Latitude" format="%20s" />
    <Header name="Latitude" format="%10.4f" />
    <Header name="LatitudeError" format="%10.4f"↘
    → />
    <Header name="LatitudeComment" format="%20s"↘
    → />
   </Array>
   <Array occurs="4">
    <Header name="_Longitude" format="%20s" />
    <Header name="Longitude" format="%10.4f" />
    <Header name="LongitudeError" format="%10.4f"↘
    → />
    <Header name="LongitudeComment" format="%20s"↘
    → />
   </Array>
   <Array occurs="4">
    <Header name="_Height" format="%20s" />
    <Header name="Height" format="%10.4f" />
    <Header name="HeightError" format="%10.4f" />
    <Header name="HeightComment" format="%20s" />
   </Array>
   <Array occurs="4">
    <Header name="_GravityCal" format="%20s" />
    <Header name="GravityCal" format="%10.4f" />
    <Header name="GravityCalError" format="%10.4f"↘
    → />
    <Header name="GravityCalComment" format="%20s"↘
    → />
   </Array>
   <Array occurs="4">
    <Header name="_PressureCal" format="%20s" />
    <Header name="PressureCal" format="%10.4f" />
    <Header name="PressureCalError" format="%10.4f↘
    →" />
    <Header name="PressureCalComment" format="%20s↘
    →" />
   </Array>
   <Array occurs="2">
    <Header name="_Author" format="%20s" />
    <Header name="Author" format="%40s" />
   </Array>
   <Header name="ColumnNames" format="%60s" />
   <Header name="EndOfHeaderSeparator" format="%60↘
   →s" />
   <Array occurs="FROM_PREPROCESSOR">
    <Field name="DataGapCode" value="66666666" ↘
    →minOccurs="0" maxOccurs="unbounded" />
    <Field name="StepCorrectionCode" value="↘
    →77777777" minOccurs="0" maxOccurs="unbounded↘
    →" />
    <Field name="EndOfFileCode" value="99999999" ↘
    →minOccurs="1" maxOccurs="1" />
    <Array occurs="4">
     <Field name="TimeYMD" format="%8d" />
     <Field name="TimeHMS" format="%6d" />
     <Field name="Gravity" format="%10.6f" />
     <Field name="Pressure" format="%10.6f" />
    </Array>
   </Array>
  </Structure>
 </Ascii>
</SyntacticMetaData>
</ESML>
```

An obvious, yet important conclusion, is that ESML is very effective in representing GGP data for gravity and pressure. Although the specifics have not been provided here, ESML is also very effective in representing the auxiliary and log data that complete the triad of GGP data (see §II). The process of representing these three files of GGP data via ESML, however, rapidly reveals a shortcoming of the GGP data model in particular, and semi-structured data in general: Data, and especially header data, is often repeated. (In fact, the GGP data model is quite vague on this point.) Even though the demands of the GGP are extremely modest in terms of storage requirements, the obvious concern is that this redundancy is inefficient. A more-subtle concern is that there exists metadata - i.e., data about data - that is being completely ignored.

ESML, in tandem with XSLT/XInclude, can be used to transform/merge these three separate files into a single monthly file for each station. This operation will eliminate redundancy through re-structuring, while preserving the desired degree of instance separation (formerly provided through separate files) through use of ESML's `Structure` element. An analogous operation via XSLT/XInclude can be used to transform/merge an identified collection of monthly records for a given station into single record over the course of a year or some other identified interval of time. Because this degree of automation via XSLT/XInclude is precisely what is required for multiple-station analyses (e.g., see §I and [3] for examples of geodynamic and seismic motivators), this is the subject of a separate investigation presently underway.

The more-subtle concern of ignored metadata is considered in the following section (V).

V. Latent Metadata

In the previous section (§IV), ESML proved itself effective as a per-file vehicle for introducing an XML-based data model into the GGP. When applied in tandem with XSLT/XInclude, ESML allows for a single monthly file on a per-station basis. Thus, use of an ESML-based data model exposes efficiencies that can be derived by the elimination of redundancy. However, of graver concern is the much-more-subtle exposure of metadata - 'data about data' that was ignored in the existing GGP data model. This comprises one example of 'latent metadata' - data about data that potentially exists, but is not presently evident or realized.

Another example of latent metadata is evident in the naming convention for files in the GGP data model. For understandable reasons, the GGP data model makes use of DOS-based standards for filenames ([9]). Since the three-character filename extension received consideration previously (see §IV), attention here focuses on the eight-character filename itself - `SCYYMMRC` where:

- `SC` is the Station Code
- `YY` is the year
- `MM` is the month
- `RC` is the Repair Code - an item which relates to treatment and/or decimation of the data

GGP data is identified using this convention, and this same filename is repeated internally in each of the three files - see, e.g., Listing 1 and [9]. There is a wealth of latent metadata

encapsulated in these eight characters! Incorporating a further decomposition of the `Filename` variable, introduced in the ESML data model illustrated in Listing 2, will allow this metadata to be extracted.

In addition to their primary role as instrumentation in support of operational and research activities into Earth tides, SGs have also proven themselves as effective seismometers (e.g., [1]). As such, the GGP also encourages the delivery of higher-frequency products aimed at capturing seismic activity. Although the primary difference is in the sampling interval at which data is collected, the content of these files is very similar to that presented previously (Listing 1 and [9]) for tidal purposes. In contrast to the tidal case, it's in the filename that the most-significant differences occur - i.e.:

- The filename itself is slightly modified ([9]) - `DD` replaces `RC` as the day of the month on which the earthquake focal time occurred (specified in terms of Universal Time).
- The filename extension can assume one of several values depending on the data - see Table I and [9].

It's eminently clear that additional-product requirements, such as those in support of seismic activity, create significant challenges for the existing GGP data model. Such challenges can be easily addressed with ESML. Moreover, these challenges amplify the need to properly incoporate metadata into the ESML data model.

TABLE I
GGP CONVENTION FOR FILENAME EXTENSIONS IN THE CASE OF SEISMIC DATA (FROM [9])

Extension	Content
`S1`	Gravity and pressure together, 1-second sampling
`S2`	Gravity and pressure together, 2-second sampling
`G1`	Gravity alone, 1-second sampling
`P1`	Pressure alone, 1-second sampling
`G2`	Gravity alone, 2-second sampling
`P2`	Pressure alone, 2-second sampling

GGP data is made available via the International Center for Earth Tides (ICET, [26]) according to a well-defined, but complex process ([9]). Inherent in this process are technical (e.g., specification of file locations, security, etc.) and non-technical (e.g., progression from priviledged to open access over an identified period of time) considerations. All of these considerations impact on metadata in the context of the GGP. For example, in the case of file location, metadata needs to be able to support use of XPath. As a whole, GGP metadata must support XQuery so that queries receive scientifically useful responses.

Even at the relatively modest scale of the GGP, metadata presents challenges. Legacy data models, such as the existing model in use by the GGP, encapsulate metadata - whether it is evident or latent. In the case of latent metadata, effort may be required to extract and apply this metadata - and this may involve the use of metadata tools. Of course, this isn't a challenge specific to tidal studies. In the marine sciences ([27]), peers are supporting each other in establishing and executing best practices for metadata, and calling for appropriate recognition in the context of peer review ([28]).

As a resource-centric World Wide Web Consortium (W3C) recommendation for representing metadata, the Resource Description Framework (RDF) underlines the importance of metadata via a data model, syntax and associated schema ([29]). Because one use of RDF is to create metadata *about* a document, as opposed to attaching metadata to *parts* of a document in the way XML does, RDF has the potential to address some of the challenges identified in this section ([10]) - especially in the case of latent metadata. Furthermore, RDF provides the logical underpinnings of the Semantic Web ([10]): A machine-processable web of smart data - i.e., data that is application-independent, composeable, classified, and part of a larger information ecosystem (ontology). Through the Semantic Web (e.g., [10], [30], [31]), and even more nascent Semantic Grid (e.g., [32], [33]), RDF has the potential to significantly amplify the importance of metadata. The semantic levels implied here are likely to be complimented by rules-based languages that execute on a specific instance of data using an ontology via standard, embeddable programs known as inference engines ([10]).

Because an effective metadata solution is required for multiple-station SG analyses, this is also the subject of a longer-term, future investigation.

VI. DISCUSSION

Now in its second phase, the GGP is a well-established, scientifically motivated effort supporting the operational and research activities of the Earth tides community ([2]). Through the introduction of a data model (§II), the GGP removed organizational and geographic boundaries by establishing a framework that supports collaboration on an international scale. Standards for instrumentation and data, together with bilateral agreements, were key to enabling this collaboration.

Ongoing success, in combination with the overall utility of SGs as 'general-purpose' instruments, has resulted in interest in GGP data from communities external to those interested in Earth tides. Of particular note is interest from the seismological community. Although seismic activity requires a higher frequency of sampling than that normally used in the GGP for tidal purposes, even more compelling is the requirement for near real time data. Extension of the GGP data model to incorporate the needs of the seismic community amplifies existing and creates new challenges.

Based on scientific and technical motivators (§I), Lumb & Aldridge ([6]) assessed GGP as it exists today, and identified opportunities for use of Grid Computing in this context. GGP's challenges and opportunities included the need for an improved data model. Although the choice of an XML-based representation was clear, it was necessary to find a solution that:

1) Makes use of XSD
2) Provides support for ASCII-format files
3) Has Earth Sciences affinities

4) Has industry standard affinities

Even though BFD and BinX deliver considerable value, only ESML leverages the most-desireable technologies, provides the required support and has the identified affinities (§III). Use of state-of-the-art XML technologies like XSD also means that ESML can make use of XPath (for location paths), XQuery (for cross-repository queries), XSLT (for transformations) and XInclude (for merging). Clearly there is a lot to be gained by leveraging this XML-based foundation.

ESML's native support for ASCII-formatted documents allowed GGP data to be represented effectively and efficiently (§IV). Repeated use of multidimensional arrays and rules-based constructs proved especially valuable. Use of ESML rapidly identified two shortcomings of the existing GGP data model. First, use of an XML-based data model illustrated that three GGP files could be easily represented as a single file. XSLT and/or XInclude were suggested as vehicles to enact this representation. Analogous and approapriate applications of XSLT/XInclude could also produce aggregated records for identified periods of time. In both of these transformations, use of ESML's `Structure` element is expected to be of value.

Application of ESML to GGP data also revealed that metadata is often ignored. The term 'latent metadata' was introduced (§V) to describe this data about data that potentially exists, but is not presently evident or realized. Although the examples presented suggest that there may exist tools to assist in extracting this metadata, such approaches are unable to overshadow extant challenges with metadata. As use of XML-based data models increases, and especially as use is made in large-scale projects (e.g., [27]), metadata challenges are receiving attention. In the long term, the most-attractive solution will necessarily have a semantic basis. This makes proactive investment in RDF, plus the emerging Semantic Web and Semantic Grid technologies, appear sensible.

The GGP transformed isolated scientists with isolated instruments and a regional focus on Earth tides into internationally collaborative scientists with a global network of instruments and a global focus on Earth tides. The GGP also permitted less-formalized collaboration with respect to the seismicity community. When metadata is introduced effectively, isolated disciplines like global geodynamics are easily incorporated into an even broader interdisciplinary collaborative fabric. This evolution is highly consistent with the trend towards systems science that is evident from weathering science (e.g., [34]) to systems biology ([35]) to computer science ([36]). With the promise of knowledge, delivered via semantic technologies, this systems science has the potential to be even more compelling.

A number of topics offer interesting prospects for further research. XInclude/XSLT-based structural manipulation of ESML-represented GGP data is a pressing requirement for compressing monthly, per-station GGP data from three down to a single file. Aggregating monthly data into longer period records is another priority. Also of interest in the short term is the need to gain expertise with XQuery and XPath. Although initial investigations can be based on local access, experimentation in Web and eventually ICET settings will become necessary. Longer-term investigations will be required to better understand the role of metadata in the context of the ESML-enabled GGP. An RDF-based approach also shows promise, but requires considerable investment to be realized.

In addition to the uses mentioned above, GGP metadata includes data relevant in the negotiation of consumer-provider relationships within the context of virtual organizations ([37]). GGP's non-technical bilateral agreements provide another example of data required in the negotiation of such relationships. Although considerable effort will be required, it is noteworthy that the technical underpinnings for such negotiations are effectively in place. On the standards front, Web Services Agreement (WS-Agreement, [38]) is already in public-comment phase under the auspices of the GGF. In parallel, but in terms of implementation, the Community Scheduler Framework (CSF, [39]) makes use of WS-Agreement and makes available a triad of core services to enable such negotiations. As a representative small-to-medium-scale science, GGP has the potential to provide interesting challenges and opportunities for CSF in particular, and Grid Computing in general.

The 26 December 2004 Sumatra-Andaman earthquake has been recorded by seismometers ([4]), the resulting tsunami has been recorded by tide gauges ([5]) and satellite altimeters ([40]), while the associated crustal deformation has been estimated by the Global Positioning System (GPS, [41]). GGP data is currently being analyzed and will provide a gravimetric record of this event. In addition to the data from these sensors, there are countless anecdotal reports available. Despite the availability of data, tsunami physics is a discipline requiring considerable research - especially in the high-impact area of early warning ([5]). In parallel with traditional modes of investigation, Sensor Data Fusion (SDF) appears promising in its ability to provide a formal framework for the alliance of data originating from different sources ([42], [43]). XML-based data models, such as that introduced here for the GGP, provide a solid foundation for incorporation into the broader-based framework of SDF.

ACKNOWLEDGMENT

The authors acknowledge the substantial efforts of those involved in the Global Geodynamics Project. They have established a solid foundation that helped to motivate the current investigation, and continues to inspire further research into the intersection of Grid Computing with Earth Sciences. The authors also acknowledge Christopher Smith and an anonymous reviewer for constructive feedback, plus Jayson Durham for introducing them to Sensor Data Fusion.

REFERENCES

[1] D. Crossley, J. Hinderer, G. Casula, O. Francis, H.-T. Hsu, Y. Imanishi, G. Jentzsch, J. Kaarianen, J. Merriam, B. Meurers, J. Neumeyer, B. Richter, K. Shibuya, T. Sato, and T. van Dam, "Network of superconducting gravimeters benefits a number of disciplines," *Eos Trans. Am. Geophys. U.*, vol. 80, pp. 121–126, 1999.

[2] "Global Geodynamics Project," http://www.eas.slu.edu/GGP.

[3] L. I. Lumb and K. D. Aldridge, "Evidence for a generalized core resonance phenomena in tidal gravimetry," *Eos Trans. AGU*, vol. 85, no. 47, 2004, Fall Meet. Suppl., Abstract MR43A-0878.

[4] J. Park, K. Anderson, R. Aster, R. Butler, T. Lay, and D. Simpson, "Global Seismographic Network records the great Sumatra-Andaman earthquake," *Eos Trans. Am. Geophys. U.*, vol. 86, pp. 57, 60–61, 2005.

[5] C. Lomnitz and S. Nilsen-Hofseth, "The Indian Ocean disaster: Tsunamic physics and early warning dilemmas," *Eos Trans. Am. Geophys. U.*, vol. 86, pp. 65, 70, 2005.

[6] L. I. Lumb and K. D. Aldridge, "Towards grid-enabling the Global Geodynamics Project," *Eos Trans. AGU*, vol. 85, no. 17, 2004, Joint Assembly Suppl., Abstract G34A-03.

[7] J. J. Bunn and H. B. Newman, "Data-intensive grids for High-Energy Physics," in *Grid Computing: Making the Global Infrastructure a Reality*, F. Berman, A. Hey, and G. Fox, Eds. John Wiley & Sons, Ltd., 2003, ch. 39, pp. 859–905.

[8] "GWR Instruments, Inc." http://www.gwrinstruments.com.

[9] "Global Geodynamics Project Agreements and Standards," http://www.eas.slu.edu/GGP/ggpas.html.

[10] M. C. Daconta, L. J. Obrst, and K. T. Smith, *The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management*. Wiley Publishing Inc., 2003.

[11] T. Erl, *Service-Oriented Architecture: A Field Guide to Integrating XML and Web Services*, ser. Professional Technical Reference. Prentice Hall, 2004.

[12] "XML Inclusions (XInclude) Version 1.0," http://www.w3.org/TR/xinclude.

[13] M. Westhead, T. Wen, and R. Carroll, "Describing data on the grid," in *4th International Workshop on Grid Computing*, IEEE Computer Society. IEEE Computer Society, 2003, pp. 134–140.

[14] "Binary Format Description Language (BFD)," http://collaboratory.emsl.pnl.gov/sam/bfd.

[15] "eXtensible Scientific Interchange Language," http://www.cacr.caltech.edu/SDA/xsil/.

[16] "Pacific Northwest National Laboratory," http://www.pnl.gov.

[17] "BinX," http://www.edikt.org/binx.

[18] "e-Science Data, Information and Knowledge Transformation," http://www.edikt.org.

[19] "National e-Science Centre," http://www.nesc.ac.uk.

[20] "Earth Sciences Markup Lanaguage (ESML)," http://esml.itsc.uah.edu.

[21] "The Information Technology and Systems Center (ITSC), The University of Alabama in Huntsville," http://www.itsc.uah.edu.

[22] R. Ramachandran, "Earth Science Markup Language Tutorial," Jul. 28 - Aug. 1 2003, Earth Science Federation Meeting, Boulder, Colorado.

[23] "Data Format Description Language Working Group," http://forge.gridforum.org/projects/dfdl-wg.

[24] "The Global Grid Forum," http://www.ggf.org.

[25] R. Ramachandran, A. McDowell, X. Li, S. Movva, and M. He, *Earth Science Markup Language: Schema Documentation for v3.0*, ESML Team, November 2003.

[26] "International Center for Earth Tides (ICET)," http://www.astro.oma.be/ICET.

[27] "Marine Metadata Interoperability," http://www.marinemetadata.org.

[28] S. P. Miller, D. Clark, J. Helly, D. Sutton, and T. Houghton, "SIOExplorer: Advances across disciplinary and institutional boundaries," *Eos Trans. AGU*, vol. 85, no. 47, 2004, Fall Meet. Suppl., Abstract SF42A-08.

[29] S. Abiteboul, P. Buneman, and D. Suciu, *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann Publishers, 2000.

[30] "Semantic Web," http://www.semanticweb.org.

[31] "World Wide Web Consortium's Semantic Web," http://www.w3.org/2001/sw.

[32] D. D. Roure, N. R. Jennings, and N. R. Shadbolt, "The Semantic Grid: A Future e-Science Infrastructure," in *Grid Computing: Making the Global Infrastructure a Reality*, F. Berman, A. Hey, and G. Fox, Eds. John Wiley & Sons, Ltd., 2003, ch. 17, pp. 437–470.

[33] "Semantic Grid," http://www.semanticgrid.org.

[34] "Weathering System Science," http://www.wssc.psu.edu.

[35] "Systems Biology," http://www.systems-biology.org.

[36] I. Lumb, "High productivity computing systems valuations," *Scientific Computing*, vol. 21, no. 11, pp. 27–28, October 2004 2004.

[37] I. Foster, C. Kesselman, and S. Tuecke, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," in *Grid Computing: Making the Global Infrastructure a Reality*, F. Berman, A. Hey, and G. Fox, Eds. John Wiley & Sons, Ltd., 2003, ch. 6, pp. 171–197.

[38] "WS-Agreement," https://forge.gridforum.org/projects/graap-wg#gotoDocuments.

[39] "The Community Scheduler Framework," http://sourceforge.net/projects/gcsf.

[40] J. Gower, "Jason 1 detects the 26 December 2004 tsunami," *Eos Trans. Am. Geophys. U.*, vol. 86, pp. 37–38, 2005.

[41] S. A. Khan and O. Gudmundsson, "GPS analyses of the Sumatra-Andaman earthquake," *Eos Trans. Am. Geophys. U.*, vol. 86, pp. 89, 94, 2005.

[42] L. Wald, "A European proposal for terms of reference in data fusion," *International Archives of Photogrammetry and Remote Sensing*, vol. XXXII, pp. 651–654, 1998.

[43] ——, "Some terms of reference in data fusion," *IEEE Trans. Geosci. Rem. Sens.*, vol. 37, pp. 1190–1193, 1999.