Harmonization of Data Formats for Tsunami Simulation Products

Matthias Lendholt
German Research Centre for Geosciences
matthias.lendholt@gfz-potsdam.de

Martin Hammitzsch
German Research Centre for Geosciences
martin.hammitzsch@gfz-potsdam.de

Peter Löwe
German Research Centre for Geosciences
peter.loewe@gfz-potsdam.de

ABSTRACT

The development of sustainable tsunami early warning systems (TEWS) requires the adoption of proven standards for components on all system levels. This is crucial to ensure the successful operation of the overall system in the long term. Currently, components, data formats and models used to build an individual TEWS come from independent development efforts, using non-standardized proprietary interfaces. Integrating these components into a TEWS requires additional work effort due to the proprietary technologies and formats. This article discusses alternative cost-effective approaches. The successful integration of the TEWS system components depends critically on the adoption and application of industry standards and good practices. From this perspective, this article examines the role of tsunami simulation models, and the challenge to integrate the data products generated from independent tsunami models for a TEWS. The significance of tsunami simulation products, consisting of data and metadata, for the overall early warning workflow is described, including data exchange (among multiple TEWS) and information visualization in combination with additional spatial information. As an outcome, the use of standardized data formats for simulation products is recommended for future work. This approach is demonstrated on a simulation of the March 2011 Tohoku-Oki mega thrust earthquake.

Keywords

Open standard, tsunami simulation, data format, interoperability

INTRODUCTION

The devastating 2004 tsunami which killed more than 200,000 people in the countries bordering the Indian Ocean triggered worldwide efforts in tsunami research. This led to the implementation of tsunami early warning systems (TEWS) in the Indian Ocean and other world regions, to minimize future threats by tsunamis. TEWS development comprises challenges on multiple interconnected levels, ranging from the study of the underlying natural processes, the derivation of models for tsunami prediction, advances in technology to the improvement of international information exchange. The set-up of an early warning system requires a considerable – and often underestimated – integration effort to integrate the results from scientific models, concepts and technologies which stem from various sources. Owing to the spatial content of the tsunami-related data (Lendholt, 2011), this task is taken on by geoinformatics, being the interface between geosciences and computer science.

After the 2004 tsunami, research focused initially on a small range of research topics in order to improve tsunami early warning. Attention shifted only afterwards to the development of reference architectures with standardized interfaces and reusable components to facilitate and ensure a porting of the entire system or individual components for other geographical regions and to other natural disasters (Wächter et al., 2012). The adoption of existing standards is a significant step towards well defined interfaces between reusable system components. Standards of the Open Geospatial Consortium (OGC) and those of the Organization for the Advancement of Structured Information Standards (OASIS) are especially important for this. The OGC Sensor Web Enablement (SWE) is suitable for integrating sensors and their observations so significant events such as earthquakes and sea level changes can be quickly recognized. Based on these incoming sensor observations, a

wave propagation forecast must be derived, to identify areas, regions and infrastructures at risk. The required integration of a Tsunami simulation system could be accomplished via the OGC Web Processing Service (WPS) (Hammitzsch et al., 2012). For creating a situation overview including spatial information, such as thematic maps, the OGC Web Map Service (WMS) and the OGC Web Feature Service (WFS) can be used. In the case that Tsunami warning alerts must be disseminated to corresponding authorities, civil emergency installations, etc., the content of such Tsunami warning alerts can be encoded using the OASIS Common Alerting Protocol (CAP). CAP messages can in turn be wrapped and transferred with additional address information using the OASIS Emergency Data Exchange Language - Distribution Element (EDXL-DE). A high level summary of the complete workflow for tsunami early warning is provided in Figure 1.

Figure 1. Schematic Workflow of a tsunami early warning system (Lendholt and Hammitzsch, 2011)

PROBLEM STATEMENT

Tsunami simulations are based on incoming sensor measurements and both bathymetry and topography of the area of interest. The simulations predict formation, propagation and extent of hypothetical tsunamis. An essential part of such simulated tsunami forecasts is the determination of expected wave heights and arrival times at coastal regions (Behrens et al., 2010). Models, algorithms and software used for tsunami simulations are in most cases developed by scientific research groups with limited regard to existing best-practices in industrial software production. Until now, no common data formats were established to harmonize the exchange of simulation products while complying with established standards. While standard raster data formats were adopted for the storage of intermediate tsunami simulation products, a common description or reference of the actual information content contained in these formats is still lacking. Custom-made conversion software is still required to access and transform such data sets, which takes up extra time and effort for development and testing for the conversion tool. Worse, these tools are in the most cases not reusable. In the realm of internationally operated ocean-wide TEWS, this poses a significant obstacle, as simulation products of various research groups must to be harmonized.

TSUNAMI SIMULATIONS

The simulation of tsunami wave propagation relies on models and algorithms to determine the water displacement based on trigger events such as earthquake location and magnitude. Based on these, Tsunami wave propagation is calculated, which is affected bathymetry-related constraints. Advanced models refine the wave propagation by including additional sensor data like ocean level measurements. Depending on the required granularity, the calculation is done for a given grid of forecast points. Generally, a coarse grained grid is defined for the open ocean which is bordered by fine grained grids for the coastal areas. The outcome of these calculations is a three-dimensional result that depends on longitude, latitude, as well as time – even taking tides into account, if applicable.

Data Formats

The following data formats have become standards for the raster data obtained from the simulation calculation:

(1) The Network Common Data Format (NetCDF) is a binary format used mostly in scientific environments. It is a self-describing format for multidimensional data. Owing to the uniform programming interfaces, there are program libraries for many programming languages available. Many visualization programs support this data format.

(2) GoldenSurfer is a proprietary software by GoldenSoftware Co., which is used for terrain modeling, landscape visualization, etc. The GoldenSurfer data format is, like NetCDF, a binary format for storing multidimensional raster data.

(3) In addition, several proprietary ASCII formats are used. All of them apply their own data-structures and coding, without a common data format.

Although all the formats mentioned above meet the requirements of their respective scientific environment, they are not suitable in the context of a modular early warning system. Multidimensional, highly precise raster data for large areas (i.e. raster data that are closely separated by grid and time) have a large file size footprint by default. The amounts of data obtained in this field currently prevent the integration of the data into a service-oriented architecture. Both the transfer via a network and the data processing would result in unacceptable latencies, caused by the sheer size of the data sets. This would critically impair the functionality of the early warning system in an emergency. Therefore, lean simulation products are needed, containing only relevant data catering for specific application. Such effective formats drive efficient processing.

Required Simulation Products

Based on experiences gained while developing tsunami early warning systems in the projects German Indonesian Tsunami Early Warning System (GITEWS, www.gitews.de), Distant Early Warning System (DEWS, www.dews-online.org) and Collaborative, Complex and Critical Decision-Support in Evolving Crises (TRIDEC, www.tridec-online.eu), the following lean simulation products were identified as useful in the TEWS domain:

(i) Maximum water levels provide the maximum wave heights in different places in a defined observation period (Figure 2a).

(ii) Isochrones, to map the propagation of the tsunami by means of equidistant time intervals (Figure 2b). Based on the model content, the wave front can be marked accordingly. This requires information on the information quality, as an isochrone might represent a wave trough, a wave crest or a combination of several wave fronts .

(iii) Mareograms, to provide ocean level time series for selected positions, e.g. sensor locations (Figure 2c). Based on their incoming real data, a manual verification of the simulation values can be done.

(iv) Worst-case estimates provide coastal forecasting information for selected positions such as wave arrival time (estimated time of arrival, ETA) and maximum wave height (estimated wave height, EWH). A simple, yet useful hazard classification for coastal regions can be drawn up with these values (Lendholt, 2011).

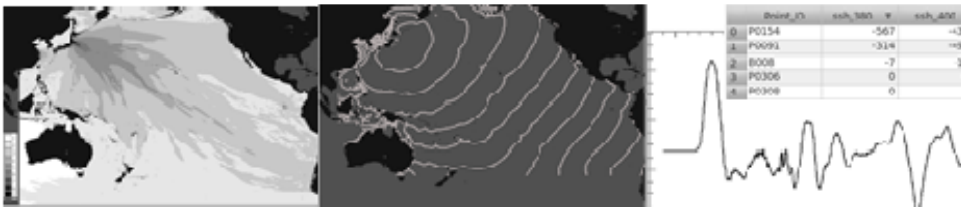

Figure 2. Tsunami simulation products. a) maximum water levels, b) isochrones, c) mareograms

All four products are encoded as vector data, contrary to the file formats for raster data. Therefore, services based on the simple feature specification (OGC 2011) can be applied for the standardized storage of features, their attributes and geometries. Apart from the four vector-based simulation products, another product (outlined below) was identified.

(v) Tsunami animations serve as overlays in map representations for allowing intuitive and fast understanding about the unfolding event. However, the data volume also prevents here a dynamic embedding in distributed and yet time-critical systems. A compression by reducing the chronological and spatial resolution must therefore be done to make the information useful. Contrary to the first four vector products, storage in a multidimensional raster format or in appropriate video formats is done to minimize the amount of data through additional compression.

Evaluation of Data Formats

Different standard file formats allow geodata storage in vector format. Currently, it cannot be predicted which standard will become commonplace for all geoinformatics application fields. Whereas the Shapefile format is still the most widely used among professional GIS users, the Keyhole Markup Language (KML) and JavaScript Object Notation for Geometry and Feature Description (GeoJSON) are used in web applications. Well-known text (WKT) and Well-known binary (WKB) are exchange formats in the backend area, but not for GUI applications. The Geography Markup Language (GML) has not prevailed yet in spite of the enormous

potentials. More important than the external container format is the metadata specification for the identified products. Only standardization at this level will allow uniform access to the data – regardless of the format in which they were coded.

For animations, no widespread container format can be identified. Whereas there are various GIS-supported formats for 2D raster data (e.g. Geo-TIFF, Geo-JPG), the geo-referenced embedding of video data in maps (as overlay) is still rarely applied and is supported by few GIS. The integration of video formats has not taken place so far, but incorporation of Geo-GIF has in isolated cases. Although the Graphics Interchange Format (GIF) can store several individual images for allowing the creation of an animation, the GIF format does not support partial transparency (alpha transparency). An alternative for use in world browsers are 'animated ground overlays', in which geo-referenced individual images with chronological stamp are referenced in a KML file.

Generation of the Simulation Products

The needed simulation products can be derived from the calculation results from the different simulation sources. Since there is currently no common standard for data exchange formats of tsunami simulations, the observation results from the simulation providers are delivered in proprietary formats and converted to a unified format in the subsequent generation steps. For example, this process is carried out in the TRIDEC project with different solution approaches. Tsunami simulation provided in ASCII and GoldenSurfer formats are converted via Python scripts to ESRI Shapefiles using a Shapefile library and the Geospatial Data Abstraction Library (GDAL). Simulation results provided in the NetCDF format are transferred into ESRI Shapefiles with a second script using an additional NetCDF library. This generation and conversion process is not very satisfactory because it is time consuming and can only be used for the generation of isochrones. For these reasons, an additional GI-supported process was developed (Löwe et al. 2011) that is implemented with GRASS GIS on the High-Performance Computing (HPC) cluster of the German Research Center for Geosciences (GFZ) (Löwe et al. 2012). This also led to to the definition of quality checks for simulation results, focusing on the inherently spatial consistency of the simulation data and the validity interval of the simulation model. The results delivered from simulation models can become spatially inconsistent and therefore invalid if the modeled time interval is exceeded. The processing of such data leads to excessive run time increases, which can be prevented by appropriate quality checking.,

SPECIFICATION

For the specification both the simulation providers with pre-calculated simulations and on-demand generators must be considered. Whereas in the first group hundreds to thousands of pre-calculated simulations must be provided in order to select the best match based on the incoming data, in the second group the calculation of the propagation model is done based on actual observation. The latter can be achieved in accurate, high-resolution models very efficiently with Graphics Processing Unit (GPU) computing. Nevertheless, the specification of service interfaces, e.g. a WPS application profile group and other APIs, has to be done to integrate simulation products into a service-oriented architecture (SOA). Furthermore, the data formats should be defined for other mark-up languages to support the exchange of simulation data in an event-driven architecture (EDA). An attempt of the necessary data format specification for the identified simulation products is given below.

(i) Maximum water levels: Vector data with line or polygon geometry. Line segments span the area in which a certain wave height is not exceeded. Related segments result from the same wave height. Necessary attribute: Name = "EWH", type = float, value = ocean level in meters over normal. ETA not specified.

(ii) Isochrones: Vector data with line geometry. Line segments span wave fronts at time t. Related segments result from the same values. Necessary attribute: Name = "ETA", type = integer, value = seconds since the tsunamigenic earthquake. Wave height is not stored because it can differ during the course of the wave crest. Different wave fronts (first or second wave and wave trough or wave crest) are stored in separate files.

(iii) Mareograms: Vector data with point geometry. Every feature corresponds to a selected place, e.g. a sensor position. Ocean levels are coded in columns/attributes. For n different time steps, n attributes are needed. The names of these attributes correspond to the time elapsed (in seconds) since the initial earthquake trigger with the prefix "ewh_". The values, being EWHs, then are provided as float values in meters.

(iv) Worst-case estimates: Vector data with point geometry. Every feature corresponds to a Point of Interest (POI), e.g. harbor, beach, bay, etc. Two attributes are needed for the worst-case classification: earliest

ETA and largest EWH (in each case, the same coding as in isochrones and mareograms).

(v) Tsunami animations: Raster data in Geo-GIF format. The geological reference data are stored in a world file. The GIF contains an animation stored as individual images with the wave propagation. The wave crests are represented in red and the wave troughs in blue. The maximum level (wave crest maximum) is identified by the RGB hexadecimal color value #FF0000, the lowest (wave crest minimum) by #0000FF. No statement about the absolute height can be made, only in combination with other data (e.g. maximum water levels or worst-case estimates).

Since different parameters, algorithms, data basis etc. may lead to different simulation results it is worth to mention that relevant metadata for each simulation would be beneficial. The metadata might be provided additionally within the simulation file names or with separate files in an agreed format, e.g. to allow automatic processing based on this metadata.

CONCLUSION AND OUTLOOK

The data format suggestions given in this article for the discussed simulation products are based on experiences and results gained from previous research projects dealing with tsunami early warning. For example, tsunami early warning systems are being developed for Turkey and Portugal as part of the TRIDEC Natural Crisis Management (NCM) demonstrators. Simulation results from several research groups have been considered so an accurate forecast can be given and a full coverage of the respective region can be achieved. Apart from scientific issues that address the comparability of different models, various simulation results with different simulation products and different formats must be integrated at the technological level. The lack of a uniform format for tsunami simulation products significantly hampers the integration of existing research results, thereby limiting their reusability as well. This identified gap should be closed with the help of the data formats defined here and allow a better integration. The data formats presented consider information contents and associated metadata in which correct data use with uniform semantics is much more important than the data formats employed. Shapefiles, for example, can be easily transferred into other encodings by means of standard software. The data formats discussed should serve as stimulus so different research groups can reach preliminary agreements and possibly initiate a future standardization process. In this context, already existing standards from OGC and OASIS should be considered, in order to ensure a seamless integration of the newly defined data formats into the area of standards that should also be supported with the addressed WPS profile. As part of the TRIDEC project, a continuous evaluation of the harmonized data formats will also take place in the future. Beyond that, a prototypical implementation of a WPS application profile group supporting the discussed formats will take place to facilitate a seamless and standardized integration of the simulation products into the system architecture.

REFERENCES

1. Behrens, J., Androsov, A., Babeyko, A. Y., Harig, S., Klaschka, F., and Mentrup, L. (2010) A new multi-sensor approach to simulation assisted tsunami early warning. *Natural Hazards and Earth System Sciences*, 10, 1085ı 1100, Copernicus Publications
2. Hammitzsch, M., Reißland, S., and Lendholt, M. (2012) A Walk through TRIDEC's intermediate Tsunami Early Warning System, *Geophysical Research Abstracts*, Vol. 14, EGU2012-12250
3. Lendholt, M. (2011) Tailoring spatial reference in early warning systems to administrative units, *Earth Science Informatics*, 4 (1), 7-16, Springer
4. Lendholt, M. and Hammitzsch, M. (2011) Addressing administrative units in international tsunami early warning systems: shortcomings in international geocode standards, *International Journal of Digital Earth*
5. Löwe, P., Hammitzsch., M., and Wächter., J. (2011) The TRIDEC Project: Future-Saving FOSS GIS Applications for Tsunami Early Warning, *AGU 2011*, San Francisco
6. Löwe, P., Klump, J., and Thaler, J. (2012) Die FOSSGIS Workbench des GFZ Compute Clusters, *Angewandte Geoinformatik 2012*, Wichmann
7. OGC (Open Geospatial Consortium) (2011) OpenGIS Implementation Standard for Geographic Information - Simple Feature Access - Part 1: Common Architecture, v1.2.1
8. Wächter, J., Babeyko, A., Fleischer, J., Häner, R., Hammitzsch, M., Kloth, A. und Lendholt, M. (2012) Development of Tsunami Early Warning Systems and Future Challenges. *Natural Hazards and Earth System Sciences*