EGU2010 SM1.3 Seismic Centers Data Acquisition session

QuakeML: status of the XML-based seismological data exchange format

Danijel Schorlemmer[1,2], Fabian Euchner[3,*], Philipp Kästli[3], Joachim Saul[2]
and the QuakeML Working Group[4]

[1] *Southern California Earthquake Center, University of Southern California, Los Angeles, CA, USA*
[2] *Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Potsdam, Germany*
[3] *Swiss Seismological Service, ETH Zurich, Zurich, Switzerland*
[4] *See Acknowledgments section*

Article history
Received May 5, 2010; accepted November 19, 2010.
Subject classification:
Seismology/General or miscellaneous, Computational geophysics/Data processing, Data dissemination/Seismological data.

ABSTRACT

QuakeML is an XML-based data exchange standard for seismology that is in its fourth year of active community-driven development. Its development was motivated by the need to consolidate existing data formats for applications in statistical seismology, as well as setting a cutting-edge, community-agreed standard to foster interoperability of distributed infrastructures. The current release (version 1.2) is based on a public Request for Comments process and accounts for suggestions and comments provided by a broad international user community. QuakeML is designed as an umbrella schema under which several sub-packages are collected. The present scope of QuakeML 1.2 covers a basic description of seismic events including picks, arrivals, amplitudes, magnitudes, origins, focal mechanisms, and moment tensors. Work on additional packages (macroseismic information, ground motion, seismic inventory, and resource metadata) has been started, but is at an early stage. Several applications based on the QuakeML data model have been created so far. Among these are earthquake catalog web services at the European Mediterranean Seismological Centre (EMSC), GNS Science, and the Southern California Earthquake Data Center (SCEDC), and QuakePy, an open-source Python-based seismicity analysis toolkit. Furthermore, QuakeML is being used in the SeisComP3 system from GFZ Potsdam, and in the Collaboratory for the Study of Earthquake Predictability (CSEP) testing center installations, developed by Southern California Earthquake Center (SCEC). QuakeML is still under active and dynamic development. Further contributions from the community are crucial to its success and are highly welcome.

1. Introduction

Seismological data cover a broad range of information and are stored and exchanged in many different formats. Often, these format definitions are tailored to fit the specific requirements for a narrow field of applications. Even though these formats share a good portion of the data fields, their definitions do not support any kind of compatibility, resulting in many formats essentially describing the same things. This is in particular the case when dealing with earthquake catalogs. In this domain, a plethora of formats is being used; almost each seismic network uses a custom format. This continues to be an impediment in the development of toolboxes for statistical seismology, because these codes need to provide import functionality for various formats. As an additional difficulty, documentation of these data formats, especially the more exotic ones, is often incomplete, and the semantics of data fields may be unclear and the lack of their description can lead to improper data usage. Some of the existing data formats stem from corresponding software products (e.g. Hypoinverse) [Klein 2007], others are standards that have been established by international organizations (e.g. the seismic format of the International Association of Seismology and Physics of the Earth's Interior) [Willemann et al. 2001], but often with no or little involvement of the community. Most of them use fixed-field width ASCII data files, which makes it virtually impossible to extend them. The motivation behind QuakeML was to provide comprehensive coverage of community-agreed content, and still have the possibility to add local extensions that are not part of the official standard but still retaining full standard compliance. To serve that purpose, XML (eXtensible Markup Language) was chosen as the base technology. XML is a standardized general-purpose markup language that allows the formal definition of descriptive languages for a broad range of applications [Bray et al. 2000]. One of its strengths is that it is plain-text based. Thus, it is platform-independent, readable by humans and machines, and probably reasonably future-proof regarding technological advancement.

XML-based data interchange formats have been developed for many fields of science and technology during

the last decade, often setting new community-approved standards. Examples can be found in Geographic Information Systems, with the basic standard of the Geography Markup Language (GML) [Portele 2007], and domain-specific application schemas, like the GeoScience Markup Language (GeoSciML) [Laxton 2009]. In the astropyhsical Virtual Observatory community, many XML-based standard formats have been developed under the auspices of the International Virtual Observatory Alliance (IVOA, www.ivoa.net). A first basic outline of the general concept of QuakeML can be found in Schorlemmer et al. [2004]. The documentation of the current QuakeML version (1.2) can be found in the Documents section of the QuakeML web site. In the current step of its evolution, QuakeML has been given a modular design. There is an umbrella schema which defines the root XML element. First-level child elements are defined in separate packages which cover a specific thematic aspect. Their schema definitions are imported from the umbrella schema. For QuakeML version 1.2, a first package has been defined that provides a basic event description (BED) of seismic events and introduces a concept for unambiguous resource identification. The BED component covers all basic parameters as routinely reported by many networks, i.e. hypocentral parameters (location, time, magnitude), their uncertainties, moment tensor and focal mechanism description, and pick/amplitude and arrival information. The BED package will in the future be complemented with subsequent standards on seismic inventory, resource metadata, macroseismic information, and ground motion. Work on these is under way.

QuakeML (BED) describes properties of seismic events in a hierarchical manner, using *a posteriori* knowledge of the relations between elements (e.g. association of origins to events). When dealing with real-time processing of seismic data, this information may not be present. Therefore, an alternative version using a flatter hierarchy has been defined for real-time use (QuakeML-RT BED).

QuakeML also addresses one of the main challenges that arise in networked scientific infrastructures. There is a need to identify resources uniquely and to make them available for searches through registries, in order to enable reliable resource discovery. For that purpose, we introduce a specific format for resource identifiers that is intended to fit smoothly into a subsequent QuakeML package on resource metadata.

The development process of QuakeML is community-driven. Version 1.0 has been subjected to an extensive *Request for Comments* (RFC) process that was conducted from December 2007 until November 2008. The outcome of the RFC has been accounted for in version 1.1, which is now being superseded by current version 1.2, again based on user experience and comments from the community.

The QuakeML development shows the strenghts of community-driven processes that allow for early identification of possible problems as well as for easier adoption by the community that sees the user needs being accounted for.

2. Data model

The main rationale of QuakeML was to create a flexible format for data interchange to foster interoperability of distributed infrastructures. The definition of this XML-based exchange format has been created in XML Schema (XSD) [Biron and Malhotra 2004, Fallside and Walmsley 2004, Thomson et al. 2004]. From QuakeML version 1.2 on, we also provide a definition in a different schema language, Relax NG (RNG) [Clark and Murata 2001]. QuakeML, however, is more than just a data exchange format defined through a schema document. It is a data model that can be applied not only for data exchange, but also for data representation, manipulation, and persistent storage. We used the Unified Modeling Language (UML, a general-purpose modeling language that is developed under the auspices of the Object Management Group, www.uml.org) to create the definition of this data model. For that purpose, and also for maintenance, a graphical UML modelling tool is used (Enterprise Architect by Sparx Systems). From this tool, the model can be exported to the XML Metadata Interchange (XMI) format. XMI is an XML-based standard for the exchange of metadata information (www.omg.org/spec/XMI). The XSD and RNG schema documents are automatically created from the XMI representation of the data model by applying an XSLT transformation. XSLT, the XML Stylesheet Language, is a standard of the World Wide Web Consortium [Clark 1999]. The XMI document can also be used as the basis for other techniques of automated code generation. Some of the C++ classes and parts of the SQL database schema of SeisComP3 are automatically generated from the XMI representation of the QuakeML data model. SeisComP3 is a seismological software for data acquisition, processing, and distribution (www.seiscomp3.org) [Hanka et al. 2008, 2010]. Note, however, that SeisComP3 currently is not based on the most recent version of QuakeML.

In the course of QuakeML data model development it became clear that logical and hierarchical relations between the components depend strongly on the perspective of the modeller, and on the context in which the data model is intended to be used. Basically, there are two different perspectives that require slight differences in modelling. The first is an *a posteriori* view that is applicable when describing earthquake catalogs, or seismic bulletins. In this scenario, the hierarchy levels of the objects are known, e.g. picks are related to origins which then are related to events. Therefore, they can be arranged in parent-child like structures, yielding a multi-level tree of objects. We refer to this perspective as the *Bulletin* flavor of the data model. The other scenario is that of real-time processing of seismic data. In this case, associations of components like picks, amplitudes, magnitudes,

and origins to a specific event cannot be assumed as known. When using QuakeML-encoded messages for communication between modules of a real-time seismic processing unit or between seismic networks pooling their picks and arrivals for a single location, the format must allow for a stand-alone description of these components, without specifying to which event they will be assigned later. These constraints call for a model with much flatter hierarchies, which we call the *real-time* (RT) flavor of the data model. Except for the different hierarchy levels, the models are equivalent and the components carry the same data.

Figure 1 shows simplified UML class diagrams for the Bulletin (Figure 1a) and RT (Figure 1b) flavors of the data model, respectively. In these diagrams, only the class names are shown for clarity, and the attributes and complex types have been omitted. The full class diagrams can be found in the Documents section of the QuakeML web site, and in the standard documentation that can be downloaded from there.

3. Resource identifiers and metadata

In a global network of seismological resources there is a need for a mechanism which allows to unambiguously identify resources. In this context, resources can be of vastly different character, e.g. institutions, working groups, seismic stations, technical equipment, but also algorithms, computer codes, or published papers. We propose a naming scheme for resource identifiers which adopts the format of *Uniform Resource Identifiers* (URIs) [Berners-Lee et al. 1998]. In the following, we propose a syntax for resource identifiers that is designed along the lines of the *Identifiers* specification of the IVOA [Plante et al. 2007]. Identifiers take the generic form of

`smi:⟨authority-id⟩/⟨resource-id⟩[#⟨local-id⟩]`

They consist of an authority identifier (`authority-id`), a unique resource identifier (`resource-id`), and an optional local identifier (`local-id`). The URI schema name `smi` stands for *seismological meta-information*, thus indicating a connection to a set of metadata associated with the resource. The URI schema name prefix is not strictly a part of the resource identifier. Other URI schema names can be used with the identifier in order to retrieve other kinds of information associated with the resource, e.g. `quakeml` for resources that have a QuakeML representation. For the description of resources which are not officially controlled by an authority, local identifiers can be assigned using the keyword «local» as `authority-id`.

Resource identifiers are intended to be resolved by registries, i.e. institutions acting as registries will provide web services that will return a metadata description of a resource if queried with a resource identifier as a parameter. The metadata description will be largely based on the Dublin Core vocabulary [Dublin Core Metadata Initiative 2003] and will provide information on the resource's identity, curation, general content, collection and service content, and data quality. Again, existing standards from the IVOA will be incorporated as much as possible [Hanisch 2007]. In particular, the metadata contain information on how to retrieve the resource, e.g. a URL pointer to an electronic document, or a Web Service description. This mechanism is particularly useful for resources that have a QuakeML representation. In that case, a resource identifier that is used in an extensive QuakeML file can be interpreted as a short cut for a QuakeML chunk that has been left out for conciseness.

The main purpose of the registry mechanism will be the resolution of identifiers which allows subsequent data retrieval in the way outlined above. Beyond this basic functionality, we envision future application of registries as the key infrastructure components for resource discovery. High-quality metadata collections are essential for advanced search services which can be used both by humans and intelligent information retrieval agents.

4. Community aspects

One of the basic development strategies for QuakeML was to make it a community-based standard from the very early stages on, and to pursue an open style of communication. The first steps towards a conceptual description of such a data format was developed by Schorlemmer et al. [2004]. In January 2007, a meeting on XML data formats was organized by the European-Mediterranean Seismological Centre (EMSC), on which the initial develoments done by the teams from ETH Zurich and GFZ Potsdam were presented to a larger group of researchers from Europe, Northern America, and Japan. After the meeting, the team that authored the standard documentation was formed. We subsequently set up web-based collaboration tools. The QuakeML web site is a wiki that is open for contributions from everyone. We also set up a public mailing list (see www.quakeml.org for subscribing), which has more than 100 subscribers from all over the world.

The key element for getting a community-approved version of the standard was the installation of a process similar to the Technical Report Development Process of the World Wide Web Consortium (see www.w3.org/Consortium/Process/tr). The most important part is a public *Request for Comments* (RFC) process to which the *working draft* version 1.0 was subjected. The RFC was conducted from December 2007 until November 2008. For that purpose, we set up several wiki pages to collect reviewer's comments. A documentation of the process can be found under the RFC section of the QuakeML web site.

The outcome of the RFC led to the first *proposed recommendation*, QuakeML version 1.1, which was the predecessor of the current version 1.2. Since changes

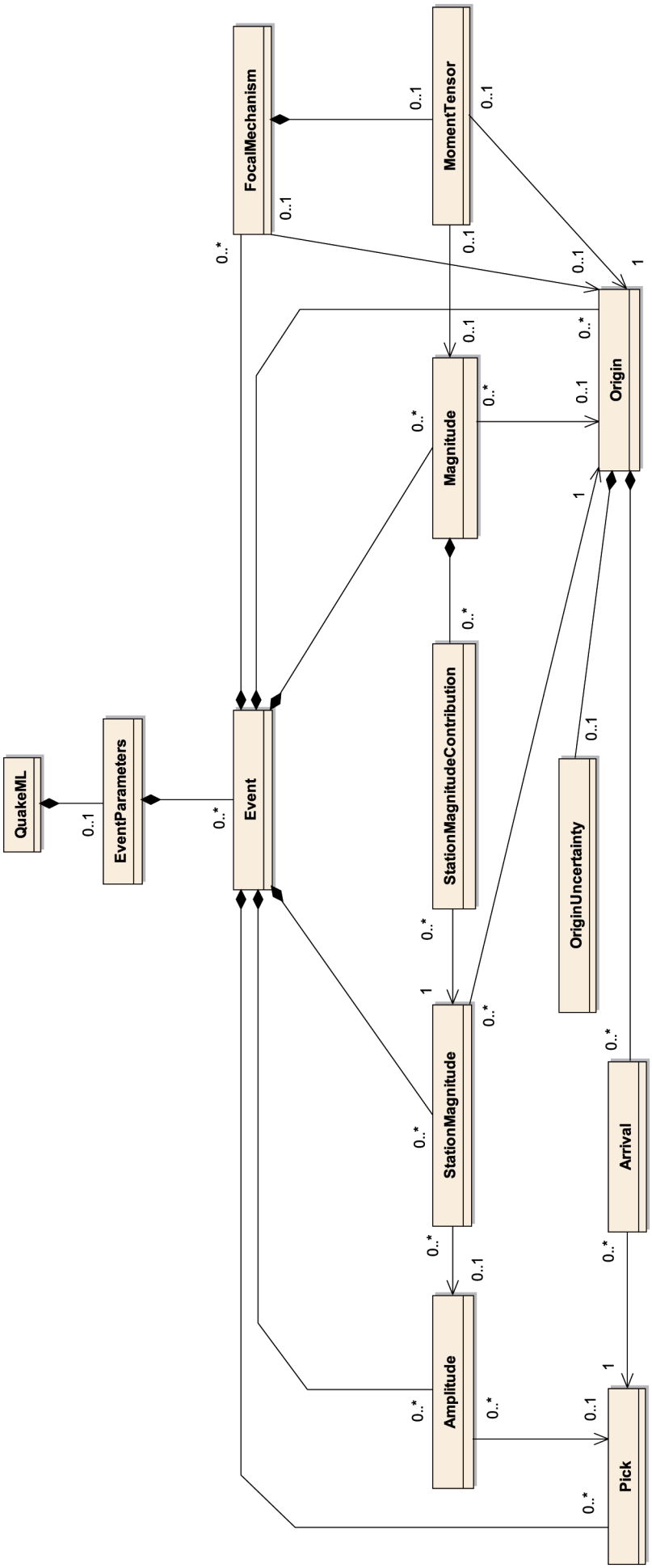
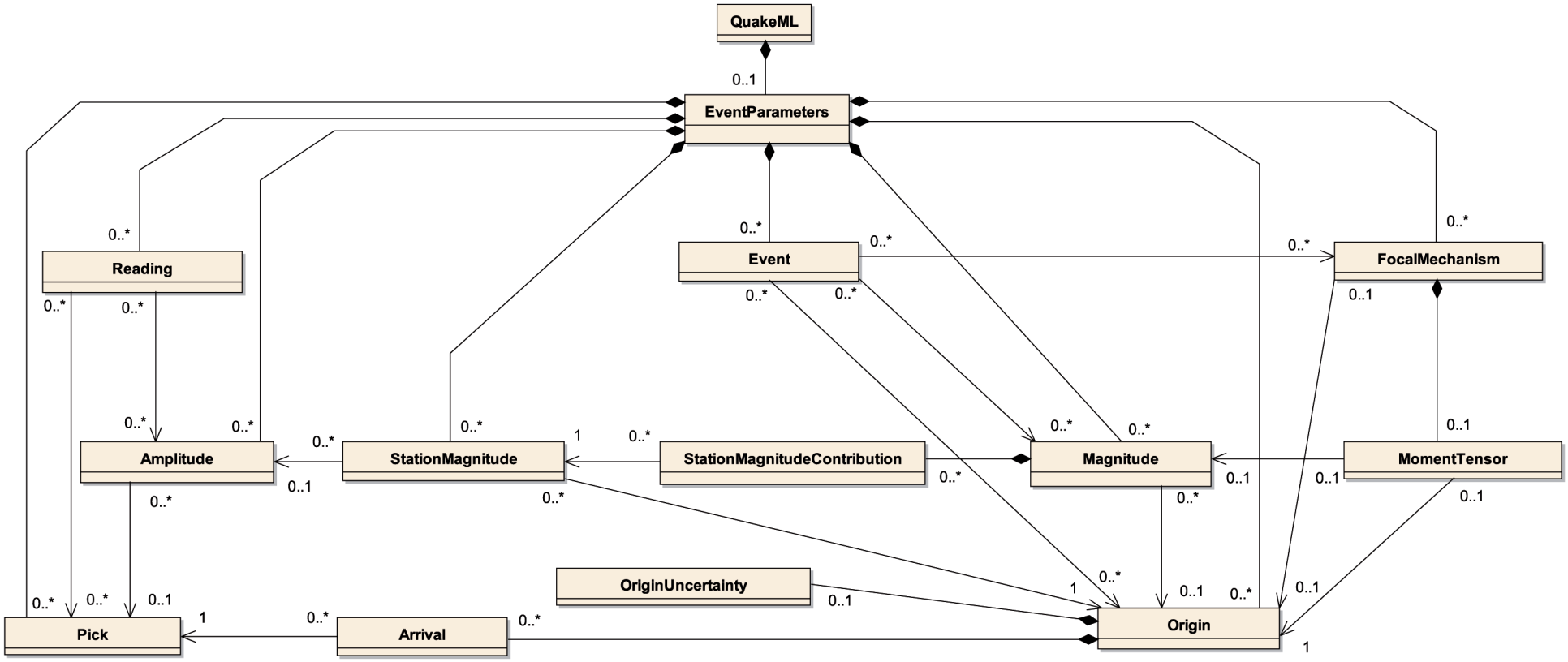


Figure 1a. Schematic view of the UML class diagram of the QuakeML data model: Bulletin style. Two different types of relations between classes are used: (i) compositions, marked by lines with filled diamonds; and (ii) associations, marked by arrows. Compositions indicate a stronger coupling between two classes than associations. The integer numbers separated by two dots shown next to the connectors indicate the multiplicity of the respective relation.

Figure 1b. Schematic view of the UML class diagram of the QuakeML data model: Real-time processing (RT) style. Two different types of relations between classes are used: (i) compositions, marked by lines with filled diamonds; and (ii) associations, marked by arrows. Compositions indicate a stronger coupling between two classes than associations. The integer numbers separated by two dots shown next to the connectors indicate the multiplicity of the respective relation.



between versions 1.1 and 1.2 are only mild, no further RFC has been called.

At the 32nd General Assembly of the European Seismological Commission in September 2010, the working group on Internet Macroseismology established a core team for the development of an XML-based data format that will be integrated as a new package into QuakeML.

5. Applications

QuakeML is being used in several applications around the globe. A couple of institutions host web services or catalog retrieval tools that allow downloading earthquake catalogs in QuakeML format. Among these are GNS Science (www.geonet.org.nz/resources/earthquake/quake-web-services.html) in New Zealand and the SCEC Data Center (www.data.scec.org/catalog search). The Earthquake Data Portal, which was created by the Observatories and Research Facilities for European Seismology (ORFEUS) foundation and EMSC in the course of the NERIES project, also provides a QuakeML catalog data export (www.seismicportal.eu). NERIES was an Integrated Infrastructure Initiative project in the Sixth Framework Programme (FP6) of the European Commission (www.neries-eu.org). The portal also has a metadata administration component which makes use of some of the concepts outlined in Section 3 [Spinuso et al. 2008]. QuakeML was endorsed as the preferred format for parametric earthquake data exchange in the NERIES project, and its follow-up (FP7) project NERA. As a consequence of the QuakeML-related developments within NERIES, parametric earthquake data exchange in QuakeML format has been established between the major European data centers for seismology, EMSC and ORFEUS.

There are several implementations that bind QuakeML representations to objects in code: i) A C++ implementation is part of SeisComP3 (see Section 2). ii) QuakePy has a Python implementation of the data model (see www.quakepy.org). One of its applications is in the CSEP testing center code for downloading and manipulating earthquake catalog data. CSEP is a project by SCEC that provides methods and tools for conducting earthquake forecasting experiments in a controlled environment (www.cseptesting.org) [Schorlemmer and Gerstenberger 2007, Zechar et al. 2010, Schorlemmer et al. 2010]. iii) GNS Science have developed `jquakeml` (codegeo.org/confluence/display/jquakeml), a Java implementation of the data model. Note, however, that at the moment none of these implementations supports the latest version of QuakeML.

Another NERIES project that makes use of QuakeML is MIDOP, the Macroseismic Intensity Data Online Publisher (www.emidius.eu/MIDOP) [Locati et al. 2006, Locati and Cassera 2010]. It allows to retrieve earthquake catalog data in QuakeML format, including a prototype version of the macroseismic extension. Furthermore, the industry is adding QuakeML support to their products to an increasing degree, as it can be seen in current products from ISTI and Nanometrics.

6. Conclusions

QuakeML, an XML representation and data model for seismological data, is intended to standardize seismological data exchange, and to be applicable for a wide range of scientific and technical problems in seismology. The current version 1.2 describes parametric earthquake data. It is community-approved in a sense that in the course of its development, a public *Request for Comments* process was conducted, and the results were incorporated in the subsequent versions. QuakeML finds an increasing level of acceptance within the community, as it is evidenced by the emergence of a significant number of tools and online services that are based on it. Efforts to extend the scope of QuakeML to other fields of seismology are underway.

Acknowledgments. We acknowledge the efforts of our co-authors of the QuakeML standard documentation: Jan Becker, Ray Buland, Andres Heinloo, Linus Kamb, Alessandro Spinuso, and Bernd Weber. We thank Rémy Bossu, John Clinton, John Douglas, Torild van Eck, Göran Ekström, Paul Friberg, Stéphanie Godey, Paul Grimwood, Winfried Hanka, Maria Liukis, Philip Maechling, Silvio Maraini, Gilles Mazet-Roux, Andy Michael, Rob Newman, Johannes Schweitzer, Stefan Wiemer, Jochen Wössner, Adrian Wyss, and John Yu for their contributions to QuakeML. QuakeML development at ETH Zurich has been funded as part of the NERIES project (EC contract no. 026130). The contact e-mail address for QuakeML is quakeml@sed.ethz.ch, and the project web site is at www.quakeml.org.

References

Berners-Lee, T., R. Fielding and L. Masinter (1998). Uniform Resource Identifier (URI): Generic Syntax (IETF RFC 2396); www.ietf.org/rfc/rfc2396.txt.

Biron, P.V. and A. Malhotra (2004). XML Schema Part 2: Datatypes, Second Edition, W3C Recommendation, 28 October 2004; www.w3.org/TR/2004/REC-xmlschema-2-20041028/.

Bray, T., J. Paoli, C.M. Sperberg-McQueen and E. Maler (2000). Extensible Markup Language (XML) 1.0, Second Edition, W3C Recommendation, 6 October 2000; www.w3.org/TR/2000/REC-xml-20001006.

Clark, J. (1999). XSL Transformations (XSLT), Version 1.0, W3C Recommendation, 16 November 1999; www.w3.org/TR/1999/REC-xslt-19991116.

Clark, J. and M. Murata (2001). RELAX NG Specification, Committee Specification, 3 December 2001; www.relaxng.org/spec-20011203.html.

Dublin Core Metadata Initiative (2003). Dublin Core Metadata Element Set, Version 1.1: Reference Description, 02 June 2003; dublincore.org/documents/2003/06/02/dces/.

Fallside, D.C. and P. Walmsley (2004). XML Schema Part 0: Primer, Second Edition, W3C Recommendation, 28 October 2004; www.w3.org/TR/2004/REC-xmlschema-0-20041028/.

Hanisch, R. (2007). Resource Metadata for the Virtual Observatory, Version 1.12, IVOA Recommendation, 2 March 2007; www.ivoa.net/Documents/REC/ResMetadata/RM-20070302.html.

Hanka, W., J. Saul, B. Weber, J. Becker and GITEWS Team (2008). Timely Regional Tsunami Warning and Rapid Global Earthquake Monitoring, ORFEUS Newsletter, 8 (1); www.orfeus-eu.org/Organization/Newsletter/vol8no1/onl_seiscomp/onl_seiscomp.htm.

Hanka, W., J. Saul, B. Weber, J. Becker, P. Harjadi, Fauzi and GITEWS Seismology Group (2010). Real-time earthquake monitoring for tsunami warning in the Indian Ocean and beyond, Nat. Hazards Earth Syst. Sci., 10, 2611-2622.

Klein, F.W. (2007). User's Guide to HYPOINVERSE-2000, a Fortran Program to Solve for Earthquake Locations and Magnitudes; ftp://ehzftp.wr.usgs.gov/klein/hyp2000-docs/hyp2000-1.1.pdf.

Laxton, J. (2009). GeoSciML v2: an interchange and mark-up language for geologic information. EGU General Assembly 2009, held 19–24 April, 2009 in Vienna, Austria, abstract EGU2009-5609.

Locati, M., C. Meletti, A. Rovida, G. Rubbia, E. Ercolani and F. Meroni (2006). A WebGIS tool for the dissemination of earthquake data. EGU General Assembly 2006, held 2–7 April, 2006 in Vienna, Austria, abstract EGU06-A-09097.

Locati, M. and A. Cassera (2010). MIDOP – Macroseismic Intensity Data Online Publisher, Rapporti Tecnici INGV, 123, 87 pp.

Plante, R., T. Linde, R. Williams and K. Noddle (2007). IVOA Identifiers, Version 1.12, IVOA Recommendation, 14 March 2007; www.ivoa.net/Documents/REC/Identifiers/Identifiers-20070302.html.

Portele, C. (2007). OpenGIS® Geography Markup Language (GML) Encoding Standard, Version 3.2.1, OGC 07-036; www.opengeospatial.org/standards/gml.

Schorlemmer, D., A. Wyss, S. Maraini, S. Wiemer and M. Baer (2004). QuakeML—An XML schema for seismology, ORFEUS Newsletter, 6 (2), 9; www.orfeus-eu.org/Organization/Newsletter/vol6no2/quakeml.shtml.

Schorlemmer, D. and M.C. Gerstenberger (2007). RELM Testing Center, Seismol. Res. Lett., 78 (1), 30-36.

Schorlemmer, D., J.D. Zechar, M.J. Werner, E.H. Field, D.D. Jackson, T.H. Jordan and the RELM Working Group (2010). First Results of the Regional Earthquake Likelihood Models Experiment, Pure Appl. Geophys., 167 (8-9), 859-876; doi: 10.1007/s00024-010-0081-5.

Spinuso, A., L. Trani, S. Rives, P. Thomy, F. Euchner, D. Schorlemmer, J. Saul, A. Heinloo, R. Bossu and T. van Eck (2008). Network of Research Infrastructures for European Seismology (NERIES)—Web Portal Developments for Interactive Access to Earthquake Data on a European Scale, In: S.R. Brady, A.K. Sinha and L.C. Gundersen, eds., Geoinformatics 2008—Data to Knowledge (Potsdam, Germany, June 11–13, 2008), Proceedings: U.S. Geological Survey Scientific Investigations Report 2008-5172, 46-47; http://pubs.usgs.gov/sir/2008/5172.

Thomson, H.S., D. Beech, M. Maloney and N. Mendelsohn (2004). XML Schema Part 1: Structures, Second Edition, W3C Recommendation, 28 October 2004; www.w3.org/TR/2004/REC-xmlschema-1-20041028/.

Willemann, R.J., R. Luckett, B. Presgrave and J. Havskov (2001). IASPEI Seismic Format; www.isc.ac.uk/doc/analysis/2001p07.

Zechar, J.D., D. Schorlemmer, M. Liukis, J. Yu, F. Euchner, P. J. Maechling and T.H. Jordan (2010). The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science, Concurr. Comp. Pract. E., 22 (12), 1836-1847; doi: 10.1002/cpe.1519.

*Corresponding author: Fabian Euchner,
Swiss Seismological Service, ETH Zurich, Zurich, Switzerland;
e-mail: fabian.euchner@sed.ethz.ch.