SECOND EUROPEAN CONFERENCE ON EARTHQUAKE ENGINEERING
AND SEISMOLOGY, ISTANBUL AUG. 25-29, 2014

A PROOF-OF-CONCEPT METADATA REGISTRY BASED ON QUAKEML IDENTIFIERS

Philipp KÄSTLI[1] and Fabian EUCHNER[2]

Identifying datasets with unique and persistent identifiers and augmenting them with adequate metadata has been a key topic in recent discussions on improving infrastructures for data-based research in geosciences. The development started from classical citation of publications towards identifying digital documents (later also large, well consolidated research datasets) and augmenting them with standardized metadata.

Standardization is often introduced implicitly by relying on large, centralized service providers with well-defined infrastructure and requirements, such as the DOI system of the International DOI Foundation, or the PID system of EUDAT.

While these systems have proven success in identifying and describing consolidated datasets, identification is usually done a posteriori. Thus, they cannot help scientists in assembling their datasets or tracking real-time analysis processes. In an example from seismology, it is a long way from recording waveforms in the field, preprocessing data, detecting anomalies, collecting those and re-interpreting them as a located seismic event, calculating derived characterizations (such as magnitudes, moment tensors, or rupture models), and repeating this process for many events in many variants in order to create an earthquake catalog identified by a DOI. The process involves many instruments, algorithms, and lab operator decisions, which all require identification and description in order to document the final product transparently.

In order to achieve this, three modifications to the standard identity and metadata curation process are required:

a) Identifiers must be introduced early in the process, directly at the creation of the data, even if this is within the context of a (possibly off-line) code execution. This requires a multi-step identification of (i) the agency defining this context, (ii) the context, and (iii) the information unit within the context.

b) Metadata must be allowed to be added late, as it may depend on interpretation being added at subsequent steps, and with great flexibility in the metadata schema (as the types of information to be described varies widely, also the structure of adequate metadata does).

c) Metadata resolver services need to be provided in a de-centralized way, on each level of context where data becomes available.

With SMI (seismological metadata identifiers), an adequate identifier system has been described as a part of QuakeML. However, there was no proof-of-concept implementation of a metadata registry available, leading to misunderstandings in the application of these identifiers.

[1] Swiss Seismological Service, Federal Institute of Technology, Zurich, Switzerland, kaestli@sed.ethz.ch
[2] Institute of Geophysics, Federal Institute of Technology, Zurich, Switzerland, fabian@sed.ethz.ch

As a proof-of-concept and technology demonstration, we have implemented a local metadata registry for a selection of seismological data products of SED (based on QuakeML Basic Event Description), as well as information on staff, organisational structures, and publications. The service is built upon the Apache Jena framework, using the Fuseki server with the SDB storage backend operating on a PostgreSQL database. Searches based on metadata criteria can be performed through a SPARQL endpoint.. The infrastructure is agnostic of the vocabularies used in the metadata, allowing for a wide range of pre-existing vocabularies to be used, as adequate for the data entities to be described. The vocabularies used in our implementation are, among others, Dublin Core elements and terms, DOI, Friend-of-a-Friend, Publishing Role Onthology, Geo-Positioning Vocabulary, and QuakeML.
Returned metadata in RDF format is available in TTL, XML, and JSON representations.

Currently, three services are exposed publicly:

http://quake.ethz.ch/metadata?r=<MY_SMI>&f=rdf-ttl|rdf-xml|rdf-json
-> basic service to retrieve all available metadata for an identifier, in RDF

http://quake.ethz.ch/rdfdemo?r=<MY_SMI>&f=rdf-ttl|rdf-xml|rdf-json
-> demo application to show all available metadata for an identifier, in human-readable formatting, and RDF

http://quake.ethz.ch:3030/sparql.html
-> generic SPARQL endpoint for any type of queries.

While this resolver service is operational and used on the SED web site, there are several ingredients missing for a full-featured and scalable distributed metadata infrastructure:

- The query performance in case of large result sets needs to be improved.

- There is currently no meta-registry (registry of registries) available.

- The referenced identifier systems (FOAF, DC, etc.) do not necessarily provide public SPARQL endpoints, requiring either mirroring of data, or non-uniform data retrieval.

Independently of these shortcomings of a typical playground implementation, we hope the service may be useful as a source of inspiration for the development of future production-level information registry and metadata services.