LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU, Geophysics • Robert Barsch

SeisHub: a web-based database

For storage, processing and simulation of multi-component data

What is SeisHub?

- Native, document-centric XML database
  - RESTful Web service (HTTP, HTTPS)
  - Standard relational database as back-end (PostgreSQL, SQLite)
  - both worlds: SQL for querying and manipulating data and any standard connected to XML, e.g. XSLT or XSD
  - Not restricted to seismology at all (collocated multi-disciplinary data)
- "Classical" seismic database
  - Index of local file-based MiniSEED waveform archive
  - Meta data about gaps, overlaps, quality and timing information
  - direct access via ObsPy to continuous seismic waveform data & meta data
  - remote waveform access via ArcLink

What is SeisHub?

- Python-based, standalone web service
- Platform independent, open source
- Implementation of various web protocols, like HTTP, SSH, SFTP
- Plug-in architecture: Dynamic discovering and loading of modules and support for Python .egg files
- Development remarks:
  - Test-driven development → proven software, so far ca. 250 test cases
  - Well-documented source code
  - Subversion
  - Trac: ticket system and project wiki

What is SeisHub?

- Data storage
  - Primary data → file based system
    - Waveform: local Mini-SEED archive
    - Other data via (GeoTIFF, GPS time series, etc.) file system
  - Meta Data → Web service on top of a XML/relational database hybrid
    - Data is packed into a XML document → Data structure is within the document, no need for a predefined database schema
    - XML resources are archived into a BLOB field
    - Only searchable values are indexed
    - Pointers to primary data

SeisHub

- Data access
  - HTTP/HTTPS: REST web service
    - XML documents have a fixed resource identifier (URL's)
    - Data transformation via XML Style Sheets on request (?output=...)
    - Data validation via Schema (XML Schema, RelaxNG, Schematron) on resource upload
    - Document properties like related meta data or indexes
  - SFTP: XML documents mapped into a virtual file system

SeisHub

- Indexing
  - Generated using a XPath expression, type and additional options
  - Simple creation + reindexing via web interface
  - Various build-in types (datetime, bool, numeric, double, float, etc..)
  - ProcessorIndex: custom processing
- Searching
  - XPath-like query on XML catalog object (restricted to indexes)
  - SQL on database object
- Mapper: predefined queries & output format bound to an fixed URL
- FileSystemResource: integrates a file system directory (read only)

Why using a web-based XML database?

Advantages:

- Data access via HTTP(S) protocol (no firewall problems)
- Most basic client to access the data is a browser
- Data provider:
  - May extend there XML based data at any time
  - Don't care about any SQL tables, because data structure is within the XML resource they provide
  - May add or delete indexes & reindex on the fly
  - Data validation on upload
  - Data transformation on request

Why using a web-based XML database?

Disadvantages:

- Slower than "common" solutions (direct access to a relational database)
  - XML parsing for validation and indexing
  - Data overhead (XML verbosity)
- Infrastructure
- Seismologist != IT nerds

www.seishub.org