SAGE Seismological Facility for the Advancement of Geoscience

IRIS Data Services

Status and challenges of FAIR data principles for a long-term repository

Chad Trabant, Rick Benson, Rob Casey, Gillian Sharer, and Jerry Carter

IRIS Data Services
A division of NSF’s SAGE facility operated by IRIS

Repository overview

The data center of the National Science Foundation’s Seismological Facility for the Advancement of Geoscience (SAGE), operated by IRIS Data Services, has evolved over the past 30 years to address the data accessibility needs of the scientific research community.

The FAIR data principles are well aligned with the needs of our data users, with many of the FAIR principles are already supported and actively promoted by IRIS.

Currently:

- 600+ terabytes of data, primarily seismological
- 300,000+ data channels from 40,000+ recording sites around the world
- Vast majority of data are open available without restriction

Findable

Global identifiers, rich metadata, shared identifiers, standardized and registered search interfaces

- Identifiers: a long-running, broadly used data identification standard (FDSN SEED), which hierarchically identifies data owners (networks), sites (stations), to unique data series (channels) is used for both data and metadata formats.
- DOIs are allocated to identify networks, aka owners of the data
- FDSN StationXML and SEED standards are very rich descriptions of the recorded data
- Standardized web service interfaces are supported by all large (and many small) centers and offer a consistent mechanism to search for data. Furthermore, these centers and their services are registered and findaable
- JSON Linked Data support is being implemented for enhanced discoverability

Challenges

- Persistence, data are renamed (new identifiers), and there is no standard mechanism to record changed identifiers
- Facilitating the discovery of seismological data beyond traditional audiences requires more work, from federated searches via OGS-compliant and linked systems to discovery via commercial web searches

Accessible

Retrieval via open, standardized web services

- (Meta)data are retrievable using standardized, REST-based web services
- Protocol and interface definition is free, open and designed for simple use
- The standard allows authentication for access to restricted data
- All metadata are considered open and there is no restriction to access even when the data they describe is embargoed

Challenges

- Metadata for a network (e.g. an experiment) is maintained in perpetuity by the FDSN (standards body), but lower-hierarchy metadata is removed if the data are removed (A2)
- Need for user identification may add a new barrier to data access

Interoperable

Standards for rich (meta)data, derivative simplified versions for broader audiences

- (Meta)data standards are open and documented
- Open source libraries and many tools exist for reading the formats
- Simple to read, derivative CSV text formats
- Newest metadata format includes capability to embed DOIs and URLs to other metadata

Challenges

- Standards are seismology-specific: while some derivatives are easy to read (e.g. CSV text), there is not broad support outside of seismology
- No formal vocabulary for these standards

Reusable

(Meta)data standards are rich, open data is the norm

- The metadata standards fully describe the transformations from measured quantity to recorded data
- (meta)data formats, norms and software are very well known in seismology, they are community standards

Challenges

- (Meta)data versioning and persistence is not currently the norm, metadata are replaced
- Data provenance beyond recording process transforms, e.g. research (pre)processing, are not supported in a standardized manner
- The community of data centers and users recognize the need for clear data licensing, how to proceed and what license to suggest/apply to data from wide variety of sources (governments, agencies, states, universities, investigators, etc.)

Closing remarks

IRIS Data Services is committed to adhering to and promoting FAIR data principles. Many of the principles are already well supported and are community norms.

The main areas where more work is needed are

1) related to use in and integration with data of other domains, through data brokers, direct access, etc.

2) data licensing (in a domain where data sharing and attribution are relatively good)

3) improved reusability/identification through (meta)data versioning

4) systematic data provenance

Thanks!