FDSN recommendations for seismic network DOIs and related FDSN services

Adam Clark (1), Peter L. Evans (2), Angelo Strollo (2)

1. Incorporated Research Institutions for Seismology (IRIS), 1200 New York Avenue NW, Suite 400, Washington, DC 20005, 202-682-2220. adam@iris.washington.edu

2. Deutsches GeoForschungsZentrum GFZ, Telegrafenberg, 14473 Potsdam, Germany. {pevans,strollo}@gfz-potsdam.de

14 May 2014

0. Preamble and overview

There is a clear and growing need to ensure that permanent seismic networks and individual experiments undertaken by small groups or individuals are reliably identified in citations by researchers or monitoring networks that make use of their data. However a common standard to do this, approved by the FDSN, is not yet in place.

With this in mind the IRIS-DMC and EIDA data centers initiated a pragmatic discussion aiming at proposing methods for the generation, maintenance, and promotion of unique identifiers for seismic networks using the existing and widely used Digital Object Identifiers (DOIs) both within and external to the FDSN, for all networks using FDSN-assigned network codes. This document represents the result of this discussion in the form of a FDSN proposal to mint and manage DOIs for seismic networks.

In this view a seismic network is an entire collection of sensor data, but also the seismic metadata associated with it, such as station details, instrument types, response data. During the discussion several options were explored including using various persistent identifier (PID) types. These support better attribution by maintaining "bibliographic" metadata about objects. DOIs were found to be the most suitable type of identifier to properly acknowledge seismic networks, because existing resolution services and familiarity make them more easily accepted by publishers as standard procedure at least as a starting point to get the community involved. Nevertheless, the system as currently designed/proposed is also able to accommodate other PIDs; this will be necessary as the community begins to move towards reproducibility of the data sets. Note that this proposal is for PID for seismic networks only. Efforts to assign PID to actual seismic waveform data are on-going, but outside the scope of this document.

The remainder of this document is divided in three parts:
1. FDSN recommendations for metadata fields, citation format and landing page: starting from the DataCite model, recommendations are given in this section for mandatory and recommended fields. A standard citation format and guidelines for landing pages have been also provided.

2. Services provided by FDSN: in order to foster the usage of seismic network DOIs, FDSN will offer DOI mapping and management services. Within this section the newly implemented services related to DOI at FDSN, including minting and landing pages, will be described.

3. Extension to the StationXML Specification: within this section an extension to Station XML has been proposed to better interlink DOIs and other persistent identifiers into technical data describing the networks.

1. FDSN recommendations for metadata fields, citation format and landing page

1.1 FDSN Metadata Field Recommendations

FDSN will provide public recommendations regarding the content of DOI metadata fields, and will follow these recommendations for DOIs that it generates ("FDSN-minted DOIs" described in the following). FDSN member network operators self-minting DOIs are also requested to follow these recommendations while minting their DOIs.

DataCite is a leading organisation devoted to promoting better access to research data. The fields listed here come from the DataCite model for metadata describing citeable datasets (http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.0.pdf). At present, DataCite uses DOIs to describe datasets. This is one of a number of alternative metadata schemes such as Dublin Core. FDSN will strongly encourage all DOI minters to use the DataCite model. In this metadata scheme, four items are required and there are several optional items.

Mandatory (Citation) Fields

All DOI metadata schemes require a common set of fields, which are used to automatically generate citation text. In developing recommendations for the data to use in these fields, considering how they are used in a citation may be instructive.

The typical citation format is:
```
Creator (Year): Title. Publisher. ResourceType. DOIName
```
FDSN recommendations should result in citations that conform with audience expectations.

Creator

The principal investigator(s), or the organization(s) operating the network. This is the entity who should receive "primary" credit for producing the data set: the main researcher(s) involved in producing the data, in priority order. Assigning this value is the right of the network operator.

Personal names should be provided as family name, given name e.g. "Muller, Max". It might instead be an institution such as a university. For a permanent network, this is likely to be an organization, e.g. "USGS","ETH","GFZ".

Title

This should be about 5-10 words naming the seismic network, similar to the descriptions found in StationXML (http://www.data.scec.org/station/xml.html) and elsewhere. It is not necessary to include the FDSN network code, as the DOI-network code mapping is maintained by the FDSN DOI registration service, and the network code will be available on the landing page. Examples: "PUNA Project, Western Argentina 1997", "Serbian Seismic Network".

Publisher

The institution (or data centre) responsible for making the data, i.e. DOI and landing page, permanently available. The DataCite model allows for only one publisher. This is defined as the *entity that holds, archives, publishes prints, distributes, releases, issues or produces the resource.* As a general rule, this will usually be **the organization that mints the DOI.** FDSN-minted DOIs will default to identifying FDSN as the publisher, but allow the network operator to override this to identify a different organization.

Publication Year

DataCite recommends using the "year when the data was or will be publicly made available." However in seismology, changes in access policy plus uncertain embargo end dates introduce ambiguity, so the publication year should be the first year of data collection by the network rather than the first year of unrestricted access. Embargo dates may be shown using the Dates fields described below.

Recommended Fields

These fields are not strictly required, but should be included in all DOIs.

Resource Type

Primary type should be "Other" and subtype should be "Seismic network". The full type would thus be "Other/Seismic network".

Description

A short summary of the network, no more than 200-300 words in length. This should include the number, location, and types of sensors, and the type of data collected. Permanent networks frequently gain or lose stations, so there is little sense in being too specific concerning these. Suggested components: Description, aim and scope -- Geodynamic setting (for temporary networks) -- Geographical coverage -- Instrument types (number of stations for temporary networks) -- Data policy (e.g. fully open/restricted/embargoed) -- Any peculiarities of the network.

The DataCite descriptionType should be "Abstract".

Without any better ideas, we suggest the text "SEED data" when that is appropriate.

Optional Fields

Contributor

For DataCite a contributor is an "institution or person responsible for collecting, managing, distributing or otherwise contributing to the development of the data set". We suggest that data centres be included as contributors, with a contributorType of "DataCollector", "DataManager", "Distributor", "Sponsor", "ContactPerson" etc. as appropriate.

Examples:

`GE network`: GFZ physically holds the archive data for the GE network, storing it offline, and makes it available on the Internet, so appears as "HostingInstitution". GFZ is responsible for maintaining the finished data set, so also appears as "DataManager":

```
<contributor contributorType="HostingInstitution">
     <contributorName>Deutsches GeoForschungsZentrum
GFZ</contributorName>
</contributor>
<contributor contributorType="DataManager">
     <contributorName>GEOFON Data Centre</contributorName>
</contributor>
<contributor contributorType="ContactPerson">
     <contributorName>geofon@gfz-potsdam.de</contributorName>
</contributor>
```

`5E network (2011-2013)`: GFZ's GIPP owns, maintains, and loans the instruments used for this temporary deployment, so appears as a "Sponsor" in this case:

```
<contributor contributorType="Sponsor">
     <contributorName>Geophysical Instrument Pool Potsdam
(GIPP)</contributorName>
</contributor>
```

The integral metadata records in XML format for these networks are available at the following links:
GE: http://data.datacite.org/application/x-datacite+xml/10.14470/TR560404
5E: http://data.datacite.org/application/x-datacite+xml/10.14470/ab466166

Location

This can be specified in any of three forms:

1. A bounding box (pair of latitude/longitude coordinates) containing all stations. This is the recommended format in most cases.
2. A list of latitude/longitude coordinates, one for each station. This is recommended only for networks with just a few stations.
3. A free-form geographic location name (e.g. "Chile" or "global"). This is recommended only for global networks, or networks whose bounds cannot easily be represented in any other form.

Size

For a completed deployment, such as a temporary network, the total amount of data in GB. For an ongoing or permanent network, this is difficult, but a rough measure of the amount of new data available per day may be suitable e.g. "500 MB/day".

Date

There are several types of dates available. If possible, metadata should clearly show the date range covered by the data.

- DateCollected:
    - For a permanent network, use this for the first day for which data was available, with a trailing slash to indicate an open date range, e.g. "1993-04-01/".
    - For a temporary deployment, DateCollected should be a range as YYYY-MM-DD/YYYY-MM-DD in the [RKMS-ISO8601](#) standard e.g. "2011-10-01/2013-05-31".
- DateAvailable: If there is an embargo on the data, this should be used to indicate when the embargo period ends.

Related Identifier

Where other resources (published publications, scientific/technical reports, data, etc.) related to the seismic network have globally unique identifiers (DOI, Handle, PURL etc.) , these can be included in the DOI metadata for the network.

1.2 Citation Format

DOI metadata may be used to generate citations, which may be rendered in a number of formats. The preferred format is:

```
Creator (PublicationYear): Title. Publisher. ResourceType. DOIName
```

The DOI system currently provides a limited form of publicly-accessible citation formatting via the CrossCite content negotiation system (see http://crosscite.org/cn/#sec-4-1). This system allows the global DOI resolution service to return citation data in various formats based on a requested HTTP content type. This standard is not yet fully mature -- currently the only widely supported content type is "x-bibliography" which returns a plain text citation, and none of the available citation styles exactly matches the preferred format shown above (see https://github.com/crosscite/citeproc-doi-server/issues/3). For now, the "apa" style is suggested as a close alternative:

```
$ curl -LH "Accept: text/x-bibliography; style=apa" http://dx.doi.org/10.7909/C3RN35SP
MASE. (2007). Meso America Subduction Experiment. Caltech. doi:10.7909/C3RN35SP
```

FDSN will provide a citation tool (http://www.fdsn.org/networks/citation/) allowing users to easily retrieve citations for one or more networks. This tool is simply a front-end to the CrossCite system, and will depend on the network's DOI registrar to return the proper data and formatting.

Here are some citation examples:

Self-minted DOIs

Permanent network GE: *only one organisation creates and publishes data, they mint the DOI and host landing page.*
GEOFON Data Centre (1993): GEOFON Seismic Network. Deutsches GeoForschungsZentrum GFZ. Other/Seismic network. doi:10.14470/TR560404

Temporary network 5E: *PI of a temporary network is the creator, chooses their own institution to be publisher.*
G. Asch et al. (2011): MINAS Project 2011/2013. Deutsches GeoForschungsZentrum GFZ. Other/Seismic network. doi:10.14470/ab466166

FDSN-minted DOIs

Permanent network II: *operator chooses to use the FDSN DOI minting service.*
IRIS GSN / University of California San Diego (1998): IRIS/IDA Seismic Network. International Federation of Digital Seismograph Networks (FDSN). Other/Seismic Network. doi:10.7914/SN/II

Temporary Network XQ: *PI chooses to use the FDSN DOI minting service.*
University of Oregon (2007): Mendocino Experiment (FAME) - EarthScope Flex Array. International Federation of Digital Seismograph Networks (FDSN). Other/Seismic Network. doi:10.7914/SN/XQ_2007

1.3 Landing Page

The network DOI will point to a "landing page" URL. DOIs are typically "linked" via a global resolver (e.g. http://dx.doi.org/10.7909/C3RN35SP) which redirects to this page.

For FDSN-minted DOIs this URL will be on the http://www.fdsn.org/ site (e.g. http://www.fdsn.org/networks/detail/II/) and FDSN will assume responsibility for maintenance of the page.

For DOIs minted through other systems, the network operator assumes responsibility for creating and maintaining this page (e.g. http://geofon.gfz-potsdam.de/doi/network/GE). **FDSN will not maintain landing pages for networks with DOIs that are registered by other groups but will only store the self-minted DOI and provide the mapping to the network** (see DOI mapping service).

All landing pages should include a minimal set of information including:
- The FDSN network code for the seismic network.
- A description of the network's temporal and spatial coverage (if available), such as a list of stations.
- Links to access data from the network (if available).
- A sample citation text as in "Citation Format" above.

Other suggested information includes:
- Terms of use.
- State of data processing (e.g. gap filling).
- Links to related scientific technical reports.

Landing pages maintained at FDSN (for FDSN-minted DOIs) will be designed according to these recommendations. FDSN member network operators maintaining their own landing page (self-minted DOIs) are requested to follow these recommendations while preparing their landing page.

2. DOI services to be offered by FDSN

2.1 Identifier Mapping Service

DOI Mapping is a basic service that FDSN will provide for every member network, which stores a DOI name (which may be created and managed by any organization) for each network and

provides a look-up service for this mapping. At first, this will not support persistent identifiers other than DOIs.

In order to maximize flexibility, the basic service that FDSN will offer is nothing beyond a "look-up service", mapping network codes (together with a year for temporary network codes) to DOIs.

1. Single Value Lookup

The primary use case is the retrieval of a single DOI using a known network identifier:

Request: GET http://www.fdsn.org/networks/doi/II
Response: II,doi:10.7914/SN/II

Request: GET http://www.fdsn.org/networks/doi/GE
Response: GE,doi:10.14470/TR560404

For permanent networks, the network code alone is sufficient to identify the network, but network codes are reused for temporary networks, so in this case the identifier will also include the start year (as appears in the "startDate" attribute in the network's StationXML record), separated from the network code by an underscore, e.g.:

Request: GET http://www.fdsn.org/networks/doi/ZU_2009
Response: ZU_2009,doi:10.1029/2012GC004201

2. Temporary Network Listing

Because the inclusion of the year is not currently standard practice (there is no accepted way to uniquely identify a temporary network) the start year associated with a particular temporary network may not be universally known. For example, suppose an experiment was set up in 2008 but did not begin collecting data until 2009. In order to support this case, temporary networks can also be accessed using just the network code. In this case, a list of all networks using that code will be returned.

Request: GET http://www.fdsn.org/networks/doi/ZU
Response: ZU_2009,doi:10.1029/2012GC004201
             ZU_2008,doi:10.7914/SN/ZU_2008

Note that the year is still considered part of the unique identifier of the network -- even if the ZU_2009 network was in operation through 2010, it could not be addressed as ZU_2010; a request for http://www.fdsn.org/networks/doi/ZU_2010 would return no results.

3. Full Data Retrieval

If no identifier is given, the service will return the complete set of mapping data. Example:

```
Request:     GET http://www.fdsn.org/networks/doi/
Response:    XQ_2007,doi:10.7914/SN/XQ_2007
             TO,doi:10.7909/C3RN35SP
             GE,doi:10.14470/TR560404
             II,doi:10.7914/SN/II
             5E_2011,doi:10.14470/ab466166
             …
```

Virtual Networks

Virtual networks will not initially be supported.

Data Storage

To implement this service, FDSN will create and manage an internal database storing this mapping information.

FDSN will not provide (or commit to storing) any data about the network beyond the DOI name itself; the DOI name can be resolved to a landing page (target URL) using a resolution service (e.g. http://dx.doi.org/10.7909/C3RN35SP), so the landing page and the DOI metadata are responsible for providing any additional network information.

Usage

Network operators will submit a web-based form supplying their network identifier and a DOI name. This request will be reviewed manually by an FDSN representative and then added to the database.

2.2 DOI Generation and Management Service ("FDSN-minted DOIs")

Through this *optional* service, FDSN will offer to mint and manage the DOI for a network. Such a DOI will then be owned by FDSN and the network operator will not have full or direct control over the DOI metadata record.

This service is provided entirely on top of the core mapping service described above -- once FDSN mints a DOI it is added to the mapping database, and within the web service described above there is no functional distinction between an FDSN-minted DOI and any other DOI.

DOI Naming Convention

The DOI name will be generated within a "namespace" owned by FDSN, thus FDSN is responsible for ensuring that the DOI name is unique. Because FDSN is already committed to ensuring a unique identifier for each network (used for the mapping service), it will incorporate this identifier into a semantic DOI name:

- Permanent networks: based on network code (`10.1234/SN/CO`)
- Temporary networks: based on network code and start year (`10.1234/SN/XE_2007`)

Landing Page

FDSN will host the DOI landing page (as described in Landing page) on the http://www.fdsn.org/ site. Because FDSN is responsible for ensuring a valid landing page, network operators will not have the option to change this.

Data Provenance From fdsnws-station

Much of the information shown on the landing page (for example, a list of the network's stations and their locations) is expected to be harvested dynamically using the `fdsnws-station` web service. Since some networks maintain more complete or canonical metadata at a particular data center, the FDSN DOI-minting service will allow a network operator to select a preferred `fdsnws-station` service URL. If a network's data is not available via any `fdsnws-station` service, the network may still use an FDSN-minted DOI, but the landing page will omit the relevant components, such as the station map.

Self-Service Data Management

Currently, network operators submit a form at http://www.fdsn.org/ giving network information when requesting a new network code, and can subsequently submit updates to this information. The FDSN DOI-minting service will be integrated into this self-service data management process:

- When requesting a new network code, or updating information for an existing network that has no registered DOI, the network operator will be invited to either provide a self-minted DOI or elect to have FDSN mint and manage a DOI for the network.
- If the network operator chooses to use an FDSN-minted DOI, the DOI metadata will be generated and updated based on the internal FDSN record of the network; thus the information submitted in these forms applies to both the FDSN record and the DOI. For example, if a network uses an FDSN-minted DOI, and the operator updates the network name through the http://www.fdsn.org/ form, both the FDSN network data and the DOI "title" field will be updated.

Submitted data is always manually reviewed before FDSN data is updated, but once the FDSN data is updated the related DOI record will be updated automatically.

Because DOIs are immutable, a network operator should decide at the outset whether to use this system or to mint a DOI through some other mechanism. Network operators should avoid creating duplicate DOIs for their network.

3. Extension to StationXML Specification

In order to better integrate DOIs into network data, support should be added to the StationXML schema (http://www.fdsn.org/xml/station/fdsn-station-1.0.xsd). Operators should then ensure that DOI are included in StationXML produced by data centers.

To support future use cases in which alternative identifier systems are used in addition to DOI, StationXML should be able to support multiple identifiers of different types. For example, here a network has a DOI identifier as well as an ARK (Archival Resource Key) identifier:

```
<Network code="CO">
 <Identifier type="DOI">doi:10.1341/abcde</Identifier>
 <Identifier type="ARK">ark:12341515</Identifier>
 …
</Network>
```

The proposed additional schema definition is:

```
<xs:complexType name="IdentifierElement">
  <xs:simpleContent>
    <xs:extension base="xs:string">
      <xs:attribute name="type" type="xs:string">
      </xs:attribute>
    </xs:extension>
  </xs:simpleContent>
</xs:complexType>
```

This should be added to the BaseNodeType in the StationXML schema, making it available for Network, Station and Channel objects:

```
<xs:complexType name="BaseNodeType">
 <xs:sequence>
   <xs:element name="Identifier" type="fsx:IdentifierType"
minOccurs="0"/>
 </xs:sequence>
</xs:complexType>
```