Best Practice Poster: The NDL Great East Japan Earthquake Archive: Features of Metadata Schema

Akiko Hashizume
National Diet Library, Japan
hasizume@ndl.go.jp

Julie Fukuyama
National Diet Library, Japan
ju-fukuy@ndl.go.jp

Keywords: The Great East Japan Earthquake; archive; dcndl; ndlkn

1. Background

The Great East Japan Earthquake, which struck Japan on March 11, 2011, caused extensive damage in several parts of Japan and has affected Japanese society, culture and economy. Since immediately after the earthquake, the importance of passing on this historical experience to future generations has been pointed out in Japan and overseas. The Japanese government announced its basic policy towards the recovery from the earthquake. This policy pointed out the need to develop a system to collect, preserve and provide access to records of and lessons learned from the earthquake, tsunami and nuclear disaster.

Based on this policy, the National Diet Library (NDL), in conjunction with numerous other organizations throughout Japan, has developed the Great East Japan Earthquake Archive Project for the collection, preservation, and provision of information related to the earthquake.

2. The NDL Great East Japan Earthquake Archive

A portal site for this project was developed by the NDL and opened to the public on March 2013. Features available at the portal site include integrated searches of resources and reports on the earthquake and subsequent disasters produced by public institutions, private organizations, and mass media companies as well as research publications by universities, academic societies, and research institutions. The portal site has been named HINAGIKU, which means daisy in English.[1] This name is intended to convey an image of hope for the future and mutual concern in support recovery from the earthquake.

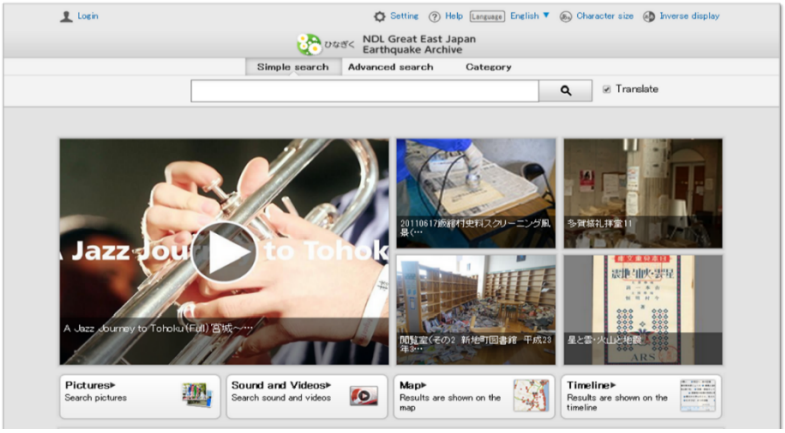

FIG. 1. Top page of the Great East Japan Earthquake Archive (HINAGIKU) (English version).

[1] HINAGIKU is an acronym of Hybrid Infrastructure for National Archive of the Great East Japan Earthquake and Innovative Knowledge Utilization.

HINAGIKU allows you to search the following resources. By the end of April 2014, the number of searchable records in HINAGIKU had reached 2,642,788.

TABLE 1: Resources collected in HINAGIKU.

Subject	Records of the Great East Japan Earthquake and the damage it caused, records of the affected areas before the earthquake, records of the restoration and reconstruction after the earthquake
	Records of aid activities by the national government, local municipalities, and other public organizations as well as records of aid activities by volunteer groups, non-profit organizations, and other private initiatives.
	Records of disaster prevention planning and academic research before and after the Earthquake as well as records of disaster prevention planning for the future
	Records of nuclear hazards resulting from the earthquake
	Records of earthquakes, tsunami, and other natural disasters from the past
	Records of the impact of past earthquakes on politics, economics, and society in Japan and around the world
	Records of the Great East Japan Earthquake and the damage it caused, records of the affected areas before the earthquake, and records of restoration and reconstruction after the earthquake
Format	Books, journals, newspapers, and other publications and digitized data
	Reports, research papers, news
	Websites of public and private organizations
	Images
	Video
	Audio (interviews, etc.)
	Fact sheets (observed data, geodetic data, etc.)

The user-friendly HINAGIKU interface includes a map display and a timeline display. Users interested in searching documents, images, video, and other digital material from a particular region can browse via the map display. Users interested in searching digital material chronologically search via the timeline. The time base can be changed to facility tracking the passage of time and reviewing the progress of reconstruction initiatives.

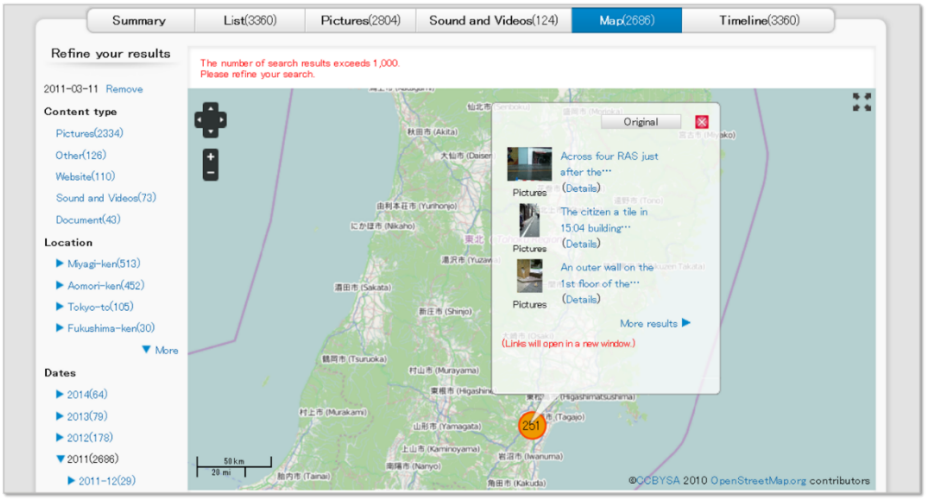

FIG. 2. Map page of HINAGIKU (English version).

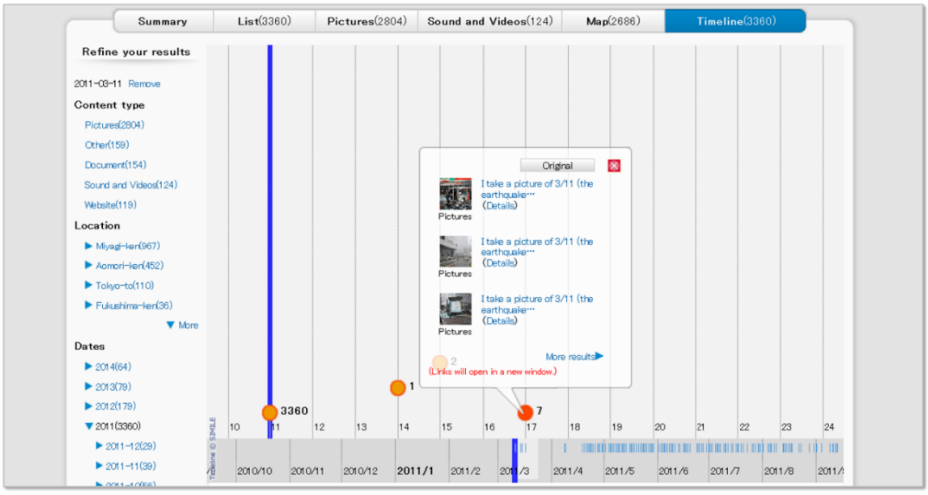

FIG. 3. Timeline page of HINAGIKU (English version).

To enable integrated searches, HINAGIKU collects three types of metadata:

1. metadata on digital materials stored in HINAGIKU
2. metadata from the NDL's other databases, including the online catalog (NDL-OPAC)
3. metadata collected from other databases created by other organizations,[2] including those of local municipalities, universities, and mass media

To handle this metadata in HINAGIKU, we developed the Great East Japan Earthquake Archive Metadata Schema (NDLKN).[3] This schema is based on the National Diet Library Dublin Core Metadata Description (DC-NDL), which is our own metadata schema, based on the DCMES and DCMI Metadata Terms, for facilitating interoperation of metadata between libraries and

[2] The Examples of the cooperating organization with HINAGIKU is as follows:
CiNii Article by National Institute of Informatics
JAEA OPAC by Japan Atomic Energy Agency's Library
Digital Archive of Japan's 2011 Disasters by Edwin O. Reischauer Institute of Japanese Studies at Harvard University
East Japan Earthquake Picture Project by Yahoo!JAPAN etc.

[3] "NDLKN" is from 'NDL Knowledge infrastructure system metadata schema'.

related institutions in Japan. DC-NDL comprises NDL Metadata Terms as well as Application Profile and RDF Schema for NDL Metadata Terms.

Mechanical searches and harvesting of metadata are supported on HINAGIKU through Web API with SRU, OpenSearch, and OAI-PMH. API/SRU returns search results in RDF/XML.

3. The Great East Japan Earthquake Archive Metadata Schema (NDLKN)

NDLKN was created as an extension of DC-NDL, so that HINAGIKU could search the metadata not only of other institutions but also of NDL search systems, such as a discovery tool "NDL Search", which implement DC-NDL for metadata schema.

NDLKN comprises

1. 87 terms described in DC-NDL (dcndl:),
2. 33 terms described by W3C and adopted internationally (exif: etc.), and
3. 5 terms described originally in NDLKN (ndlkn:).

There were two major issues to solve in development of NDLKN. The first was coordination of metadata in various systems over multiple domains. The second was to satisfy requirements for archiving disaster records. NDLKN was developed to be a solution to these issues.

3.1. Coordinate with metadata of variable systems over domains

It was not possible to create metadata in the new NDLKN schema for existing domestic and foreign disaster record archive systems, because they held metadata in original schema. Therefore, we decided to harvest and keep metadata in the original schema in one storage and map this data to the NDLKN schema for storage for searching. We ask newly building archives to adopt NDLKN and to extend the terms according to the needs of each institution.

As mentioned above, NDLKN was extended from DC-NDL. The main differences between these schema are changes of the classes from [dcndl:Item] to [ndlkn:Resource] and from [dcndl:BibAdminResource] to [ndlkn:MetaResource]. The NDL Search, which implements DC-NDL based on FRBR model, holds terms for individual items in the class [dcndl:Item]. However, we felt that it would be difficult for organizations other than libraries to understand the concept of FRBR item, especially since HINAGIKU was intended to utilize digital materials such as images and videos more than books and journals found in traditional libraries. Also, we set [ndlkn:Resource] and changed the class [dcndl:BibAdminResource] to [ndlkn:MetaResource].

We also decided to store the URI of metadata providers in [dcterms:creator] of [ndlkn:MetaResource] class and the URI of the NDL in [dcterms:publisher]. We did this because we consider metadata providers to be primarily responsible for the metadata, which the NDL accepts and makes available.

We also assumed that the number of cooperating archives would continue to increase, and therefore it would be preferable to use identifiers for HINAGIKU metadata that would not require adjustment or reduction and would never be exhausted or overlap. As a result of these considerations, we adopted the UUID (Universally Unique Identifier)-RFC4122 and decided to add UUID to one file as a minimum unit.[4]

It is necessary to specify a license or terms of use for each resource that will be reused. Therefore, we decided to use [dcterms:license] for the information of the license and to adopt [cc:attributionURL] from Creative Commons Rights Expression Language to describe the name of the rights holders. Both of these are used in the form of URI.

Ex. 1: Creative Commons license

```
<dcterms:license rdf:resource="http://creativecommons.org/licenses/by/3.0/us/"/>
```

[4] The UUID Version 4 is a string of random 32 hexadecimal digits, so it is impossible to overlap the identifiers.

Ex. 2: Yahoo! JAPAN East Japan Earthquake Picture Project

<dcterms:license rdf:resource="http://archive.shinsai.yahoo.co.jp/contents/guide/"/>

Ex. 3: The NDL

<cc:attributionURL rdf:resource="http://www.ndl.go.jp/"/>

3.2. Meeting the needs of archiving disaster records

HINAGIKU functions not only as a web portal that enables integrated search for either digital or analogue resources but also as an archive that stores and preserves resources themselves with metadata. HINAGIKU archives digital materials such as images and videos at the moment. We considered the terms of NDLKN for each material types of objective resources.

NDLKN adopted [premis:formatName] and [premis:formatVersion] from PREMIS as terms for preservation technology. For images and videos recorded on digital cameras, we selected from Ontology for Media Resources by W3C, for example, concerning the recording location, [ma:createdIn] for the URI, [ma:locationLatitude] for latitude and [ma:locationLongitude] for longitude, and regarding the sound and video, [ma:samplingRate] for sound, [ma:frameRate] for video and [ma:duration] for playing time. We adopted the terms minimum amount necessary for images only [exif:width] for the width and [exif:height] for the height of the image from Exif data description vocabulary.

It is important for post-disaster surveys that resources such as images and videos have geospatial information. For this reason, we set terms not only for describing address, longitude, or latitude but also for distinguishing the objective space from the recording location of the resource. As for recording location, we adopted [v:street-address] and [v:postal-code] from Ontology for vCard. To describe the objective space of the resource, we described the value structure using [dcterms:spatial] and adopted [rdfs:label] for the name of the objective space, [v:region] for the prefecture, [v:locality] for the city, town and village, [v:street-address] for the street address, [v:postal-code] for the postal code, additionally [geo:lat] for the latitude, [geo:long] for the longitude from the terms of the Basic Geo (WGS84 lat/long) Vocabulary.

The temporal information is also important for disaster records. Therefore we described the date the image or video was recorded in [dcterms:created] and the date it was started to collect from a website in [dcndl:dateCaptured]. We recommend that values be stored in W3CDTF format, specifying by [rdf:datatype]. Furthermore, in HINAGIKU, metadata is mapped to W3CDTF format uniformly if possible, even if the provided metadata is not in W3CDTF format.

At the beginning of the development of the NDLKN, we assumed that it would be necessary to group the data by region, kind of disaster, or other characteristic useful to searching the data and displaying the search results. For this, we discussed to use the terms collection and item to represent a parent/child relationships in resources. However, after consideration, it became clear that it is almost impossible to describe collection uniquely. Therefore, we described both collection and item by [ndlkn:Resource] and chose to represent parent/child relationships in resources by connecting them with [dcterms:isPartOf] or [dcterms:hasPart].

HINAGIKU was initially intended to be an archive of the Great East Japan Earthquake. The target of the collection, however, includes records of earthquakes, tsunamis, and other past disasters, too, and other new archives might also be developed for future disasters. Based on these assumptions, we described [dcterms:coverage] to store the name and URI of disasters in order to describe the objective disaster of the resource.

4. Characteristic utilization examples of NDLKN in HINAGIKU system

We introduce several utilization examples of implementation of NDLKN in HINAGIKU system.

As HINAGIKU coordinates with domestic and foreign archive systems of disaster records, we assume that it would be necessary to confirm metadata schema definitions of its acquired time if

cooperative organizations change their schema in the future. Therefore, HINAGIKU system stores its own URI in [dcterms:conformsTo] of [ndlkn:Resource] class and information of original metadata schema of providers in [ndlkn:sourceConformsTo] as internal term.

We also utilized the NDLKN terms [geo:lat], [geo:long], [ma:locationLatitude], [ma:locationLongitude] for the longitude and latitude of resources such as images and videos. HINAGIKU stores the latitude and longitude data automatically from either the name of the objective space or the recording location through the Yahoo! Geocoder API when the provided metadata does not have the value of latitude or longitude.

Web sites of the local governments of stricken areas and the Japanese government are also important as disaster records. The NDL has archived web sites for a long time by the WARP system and we have started to archive disaster related web sites with higher frequency after the Great East Japan Earthquake. As for the web sites, the titles (for example 'Sendai city') are not changed even if the content changes. Therefore it is necessary for searching and distinguishing the search results to add temporal information such as year, month, and date collected to the collected web sites. For this reason, HINAGIKU stores not only the value of title but also related information in [dcterms:title] in regard to the web sites collected by the WARP system. More specifically, we described to store the date started to collect too in [dcndl:dateCaptured] with [ ] after the title.

References

Basic Geo (WGS84 lat/long) Vocabulary. Retrieved April 28, 2014, from http://www.w3.org/2003/01/geo/wgs84_pos#.

Creative Commons. Creative Commons Rights Expression Language. Retrieved April 28, 2014, from http://creativecommons.org/ns#.

IFLA (2009). Functional Requirements for Bibliographic Records: FRBR. Retrieved April 28, 2014, from http://www.ifla.org/publications/functional-requirements-for-bibliographic-records.

NDL. (2011). National Diet Library Dublin Core Metadata Description (DC-NDL), version Dec. 2011. Retrieved April 28, 2014, from http://www.ndl.go.jp/en/aboutus/standards/index.html. The DC-NDL is DCMI Metadata Terms 2010-10-11 based metadata schema and compliant with the recommendations of the DCMI.

NDL Search. Retrieved April 28, 2014, from http://iss.ndl.go.jp/?locale=en.

PREMIS (2011). PREMIS Data Dictionary for Preservation Metadata, version 0.99. Retrieved April 28, 2014, from http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl.

The Great East Japan Earthquake Archive (HINAGIKU). Retrieved April 28, 2014, from http://kn.ndl.go.jp/.

W3C (2003). Exif data description vocabulary. Retrieved April 28, 2014, from http://www.w3.org/2003/12/exif/ns.

W3C. (2006). Ontology for vCard. Retrieved April 28, 2014, from http://www.w3.org/2006/vcard/ns.

W3C (2012). Ontology for Media Resources 1.0. Retrieved April 28, 2014, from http://www.w3.org/ns/ma-ont.

WARP. Retrieved April 28, 2014, from http://warp.da.ndl.go.jp/info/WARP_en.html.

Yahoo! Geocoder API (Japanese only). Retrieved April 28, 2014, from http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/geocoder.html.