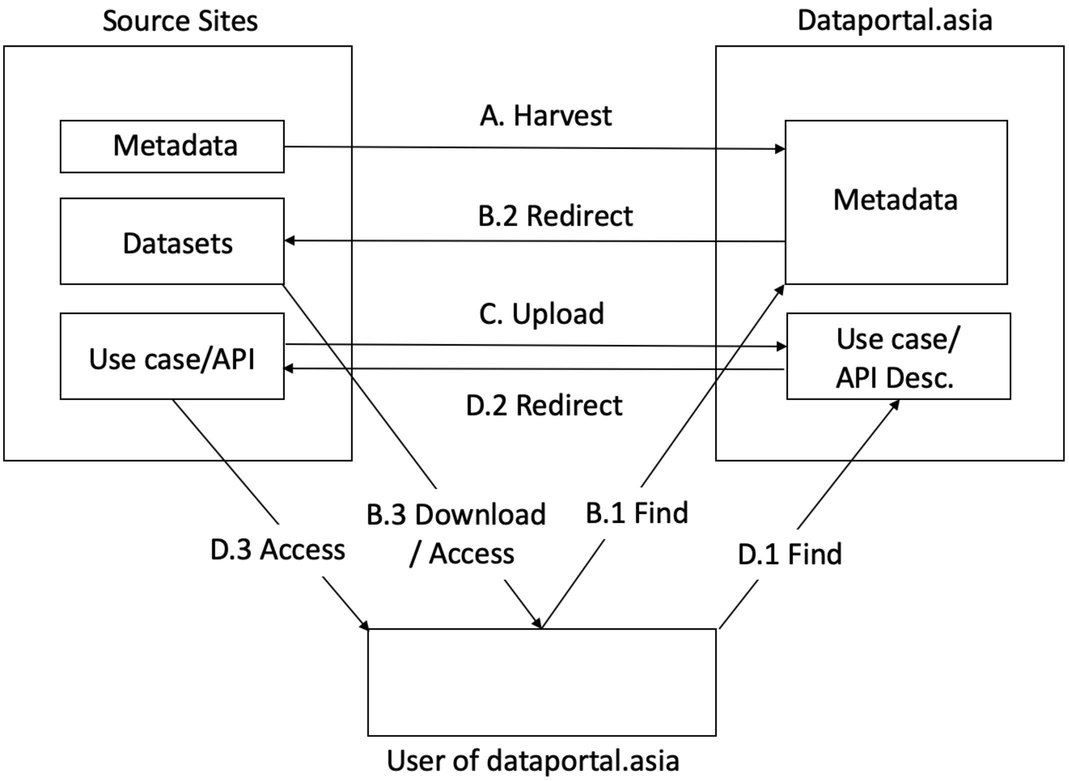
Specifications

1. Overview of the harvesting service

The following figure provides an overview of the harvesting process between Dataportal.asia and source sites.



A. Harvest: Dataportal.asia regularly visits source sites and harvests the metadata of their open datasets. Then, the harvested metadata are stored in dataportal.asia and presented to users in a bilingual way (i.e., the original language and English).

B. Processes associated with accessing Open Datasets on dataportal.asia

B.1 Find: Users look up the harvested metadata of datasets.

B.2 Explore or Redirect: When the needed datasets are found, users can explore the datasets on dataportal.asia (e.g., visualization of the dataset) or be redirected to the source site of the dataset.

B.3 Access: Users download the file of the dataset or retrieve the content of the dataset through API.

C. Users can submit their application forms of Open Data Use Cases or Web API catalogues for being published on dataportal.asia.

D. Processes associated with accessing Open Data Use Cases or Web API catalogues on dataportal.asia

D.1 Find: Users look up the Use Cases Albums or Domain Experts Albums.

D.2 Redirect: When the needed Use Cases or Web APIs are found, users can click the item for being redirected to its source site.

D.3 Access: Users access the Website of a specific use case or get the results returned by a Web API.

2. The Harvesting Program

Dataportal.asia has developed a python program (hereafter, referred to as HARVESTING program) to undertake the metadata harvesting task. The program uses the API interfaces to harvest metadata from source sites. The main supported API interfaces are CKAN APIs. HARVESTING program can be more easily modified to harvest metadata from CKAN-based source sites as long as source sites authenticate access. If a source site provides API interfaces other than CKAN APIs, the site needs to provide the API specifications and authenticates the access to API in order to be harvested by dataportal.asia.

Developers can go to https://docs.ckan.org/en/2.9/api/ for complete information about CKAN APIs. Basically, HARVESTING program uses the "package_search" API endpoint to harvest metadata from source sites. The API accepts query parameters in a request and returns a dictionary with datasets as a result. For example, HARVESTING program may make a request like the following to a CKAN-based source site.

GET https://URL of source site/api/action/package_search

The source site then answers the request with a response whose data structure contains a "results" field. The field contains a list of datasets with corresponding metadata. HARVESTING program will parse the data structure, retrieve the required metadata, create a English-description field of the metadata, and store both the original metadata as well as the created bilingual field into the database of dataportal.asia.

Notably, the metadata model of the "results" field may be different among source sites. In other word, the metadata of datasets may be described with different data structure in different source sites. Accordingly, there is a need to stipulate a metadata standard for ensuring that each source site provides metadata with minimal levels of data quality. Moreover, the metadata standard also can benefit dataportal.asia by facilitating HARVESTING program to integrate metadata harvested from different source sites.

3. The Standard of Metadata Model

Despite the importance of metadata model standard, dataportal.asia does not stipulate an official standard for source sites to describe the metadata of their datasets. Currently, HARVESTING program accommodates the differences among various source sites in a case-by-case manner. In order to increase the efficiency and quality of the harvesting process, however, dataportal.asia recommends source sites adopt DCAT (Data Catalog Vocabulary) Version 2 to describe their metadata in the future. By so doing, the harvested metadata in dataportal.asia will gradually converge in their metadata descriptions.

DCAT, recommended by W3C, is an RDF vocabulary designed to describe datasets and data services in a catalog using a standard model and vocabulary. It achieves so by stipulating six main classes of RDF vocabulary for representing data catalogs, such as dcat:Catalog, dcat:Resource, dcat:Dataset, dcat:Distribution, dcat:DataService, dcat:CatalogRecord. Using DCAT can facilitate the consumption and aggregation of metadata from multiple catalogs based on a standard set of vocabulary, thus, achieving interoperability and improving data quality. Please refer to https://www.w3.org/TR/vocab-dcat-2/ for the official documentation of DCAT 2.

4. API Interfaces

As a CKAN-based platform, users can access the harvested metadata in dataportal.asia through CKAN APIs. Users can obtain the following information through making the corresponding CKAN API calls. Please go to https://docs.ckan.org/en/2.9/api/ for complete specifications of CKAN APIs.

- List all datasets: https://dataportal.asia/api/action/package_list
- Search a datasets matching a query: https://dataportal.asia/api/action/package_search?q=*[your search condition]*
- Retrieve a specific dataset: https://dataportal.asia/api/action/package_show?id=[id of the dataset to be retrieved]