Data Source Harvest Configuration

The Research Data Australia (RDA) Registry is a software application that stores and manages collection descriptions (metadata records). The content of Research Data Australia (RDA) comes from the RDA Registry. This page describes how to configure the data source account settings to enable automatic harvesting of metadata records from institutional repositories.

RDA Registry harvesting capabilities

The RDA Registry software is capable of:

- Harvesting XML of any schema
- Harvesting JSON of any schema from the CKAN RESTful API
- Transforming harvested metadata that is not RIF-CS, to RIF-CS, by applying an XSLT (configured per data source)
- Harvesting via a selection of harvest methods, including:
  - HTTP GET (simple file retrieval from a URI)
  - OAI-PMH (query and response processing specific to this protocol)
  - OGC Catalogue Services for the Web (query and response processing specific to this interface)
  - CKAN Action API (query and response processing specific to this web service).

Harvest configuration

To proceed with the following configuration steps you will need the role of data source administrator for the data source you intend to configure. A harvest can be configured within the data source settings page.

Configuring for a source schema other than RIF-CS:

Applying an XSL Transformation to a data source harvest

To harvest content that is not RIF-CS XML, an XSLT that generates a RIF-CS XML representation of the retrieved content must be made available for the ARDC harvester.

The XSLT contains the extent of the logic that determines how many Registry records to construct, and their content, informed by the data retrieved. Additionally, standard vocabulary values can be applied where a value of similar meaning is detected, for example, a particular subject or license type. If any of the representation within the RDA Registry is to be altered, the XSLT can be modified to achieve this.

Send an email to services@ardc.edu.au to advise which metadata schema you intend to harvest, within which data source and via which interface implementation (HTTP GET, OAI-PMH, OGC CSW, or CKAN API). The ARDC Services team will advise whether an XSLT exists for generating RIF-CS from the metadata format you intend to harvest. If it does, it may require only a few customisations to suit your particular needs. Check the available crosswalks to know which schema already have an XSLT you can adopt or adapt.

When you have a new XSLT ready to use, log in to the RDA Registry and your Data Source Account. Select 'Edit Settings' and 'Harvester Settings' to access the harvest configuration section. Refer to the relevant section below for details on how to configure the particular harvest method that you require.

Harvest configuration is described on the ARDC documentation site within Data Source Account Settings.

Configuring the harvest method

HTTP GET harvest method

The HTTP GET harvest method retrieves a single file of source metadata XML from a publicly accessible URI.

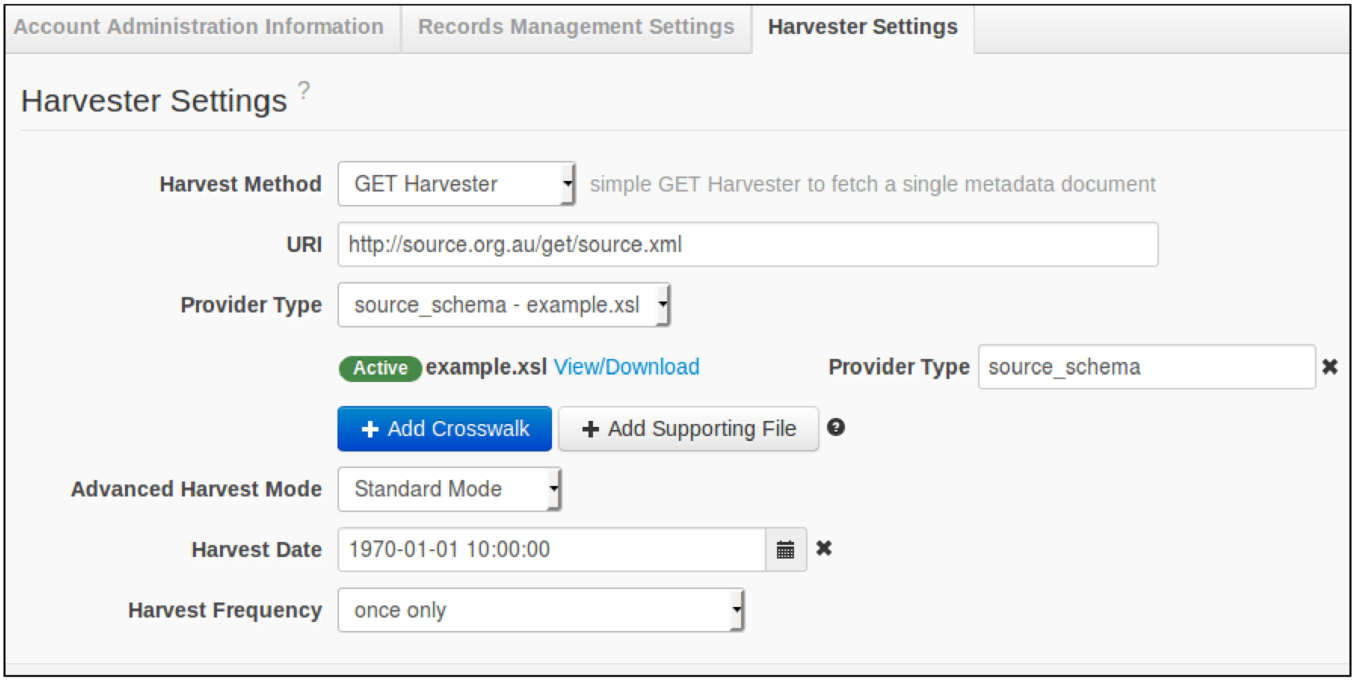
HTTP GET harvest configuration

An HTTP GET harvest requires the URI to the publicly accessible source metadata XML. If the source metadata is not in RIF-CS XML, a pre-configured XSLT must also be applied to generate a RIF-CS XML representation of the retrieved content

Within the Harvester Settings of the data source account, apply the following:

- Harvest Method: Get Harvester
- URI: {HTTP URI - eg. http://source.org.au/get/source.xml}}
- Provider Type: rif - if RIF-CS XML is being retrieved from the harvest end-point; otherwise, a crosswalk must be applied:
  - To apply a crosswalk, upload the XSL file using the “Add Crosswalk” button.
  - Set Provider Type (on the right of the XSL file name) to the source schema name and set selection within Provider Type (above the XSL file name) to “{Output Schema} - {XSL Filename}” - this will set the XSL Transform to “Active” for the harvest.
- Advanced Harvest Mode: Standard Mode or Full Refresh Mode
- Harvest Date: optional - set this to schedule a harvest
- Harvest Frequency: optional - set this to a value other than 'once only' if you've entered a Harvest Date and would like to schedule a repeating harvest

Fig 1. Example data source account configuration for HTTP GET harvest from a public URL



Account Administration Information
Records Management Settings
Harvester Settings

Harvester Settings ?

Harvest Method: GET Harvester — simple GET Harvester to fetch a single metadata document

URI: http://source.org.au/get/source.xml

Provider Type: source_schema - example.xsl

Active example.xsl View/Download — Provider Type: source_schema ✖

+ Add Crosswalk
+ Add Supporting File

Advanced Harvest Mode: Standard Mode

Harvest Date: 1970-01-01 10:00:00 ✖

Harvest Frequency: once only

OAI-PMH harvest method

The OAI-PMH harvest method retrieves XML of any schema from an OAI-PMH interface implementation.

OAI-PMH harvest configuration

An OAI-PMH harvest requires the URI to the OAI-PMH end-point. If the retrieved content does not conform to the RIF-CS XML schema, a pre-configured XSLT must also be applied.

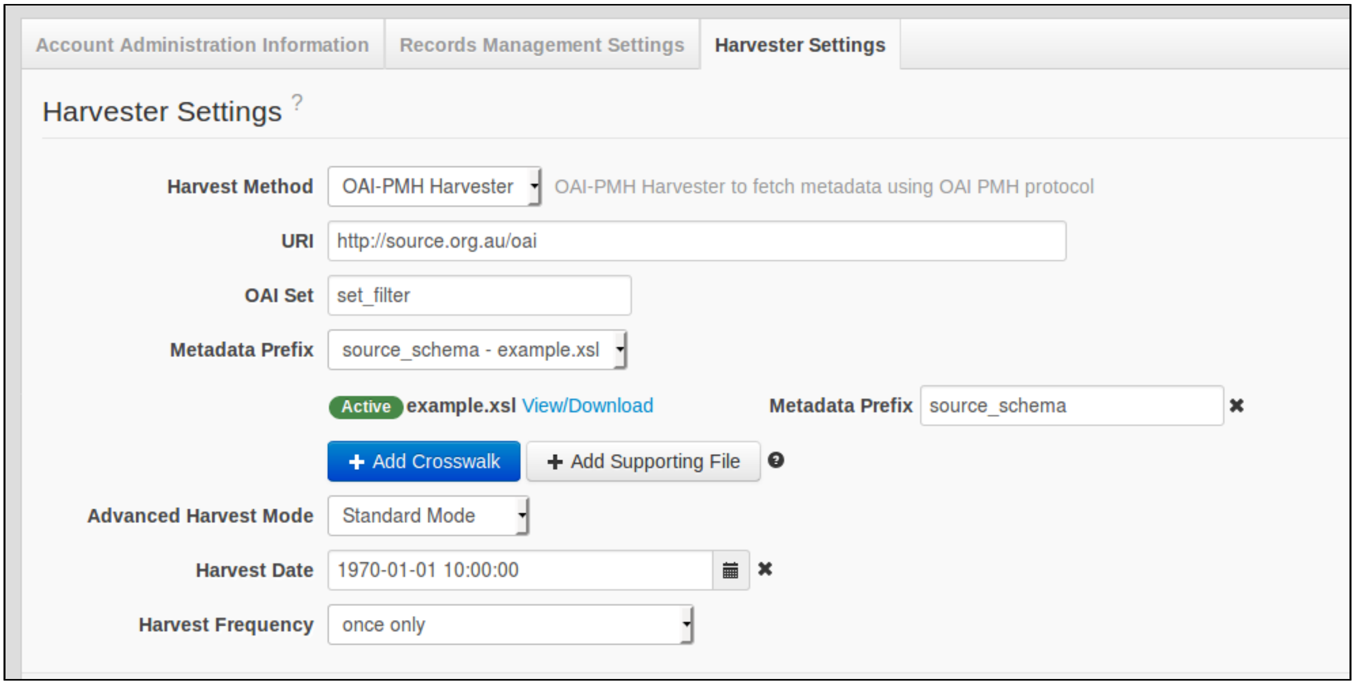
The currently available XSLTs for ISO19139, ISO19115-3, ISO19139.mcp and ISO19139.mcp-1.4 all apply the following rules per record:

- The Resource described by the source metadata is described within a new Registry record (i.e. a Collection, Activity or Service)
- per each Registry record:
  - per each Responsible Party
    - a Party record is constructed
    - the Party record is indicated as related to the Registry record
  - per each Online Resource URL
    - if the resource is determined to be a web service:
      - the resource is described within a new Service record
      - the Service record is indicated as related to this Registry object
    - otherwise
      - the resource URL is referenced within the Registry record as related information

Within the Harvest Settings of the data source account, apply the following:

- Harvest Method: OAI-PMH Harvester
- URI: {OAI-PMH interface URI - eg. http//source.org.au/oai}
- OAI Set: value required if it is necessary to filter the results by 'set'
- Metadata Prefix: rif - if RIF-CS XML is being retrieved from the harvest end-point; otherwise, a crosswalk must be applied:
  - To apply a crosswalk, upload the XSL file using the “Add Crosswalk” button
  - Set Metadata prefix (on the right of the XSL file name) to the metadataPrefix parameter value required by the OAI-PMH interface call
  - Set selection within Metadata Prefix (above the XSL file name) to “{Source Schema} - {XSL Filename}” - this will set the XSL Transform to “Active” for the harvest
- Advanced Harvest Mode: Standard Mode or Full Refresh Mode
- Harvest Date: optional - set this to schedule a harvest
- Harvest Frequency: optional - set this to a value other than 'once only' if you've entered a Harvest Date and would like to schedule a repeating harvest

Fig 2. Example data source account configuration for OAI-PMH harvest from source.org.au


Account Administration Information
Records Management Settings
Harvester Settings

Harvester Settings ?

Harvest Method: OAI-PMH Harvester — OAI-PMH Harvester to fetch metadata using OAI PMH protocol

URI: http://source.org.au/oai

OAI Set: set_filter

Metadata Prefix: source_schema - example.xsl

Active example.xsl View/Download — Metadata Prefix: source_schema ✖

+ Add Crosswalk
+ Add Supporting File

Advanced Harvest Mode: Standard Mode

Harvest Date: 1970-01-01 10:00:00 ✖

Harvest Frequency: once only

OGC CSW harvest method

An OGC CSW harvest retrieves source metadata XML from an OGC Catalogue Service for the Web interface implementation.

OGC CSW harvest configuration

An OGC-CSW harvest requires the URI to the OGC CSW end-point. If the retrieved content does not conform to the RIF-CS XML schema, an OGC CSW harvest requires a pre-configured XSLT.

ARDC Services can provide example XSL Transformations for converting from Geographic MetaData extensible mark up language (GMD) schema to RIF-CS.

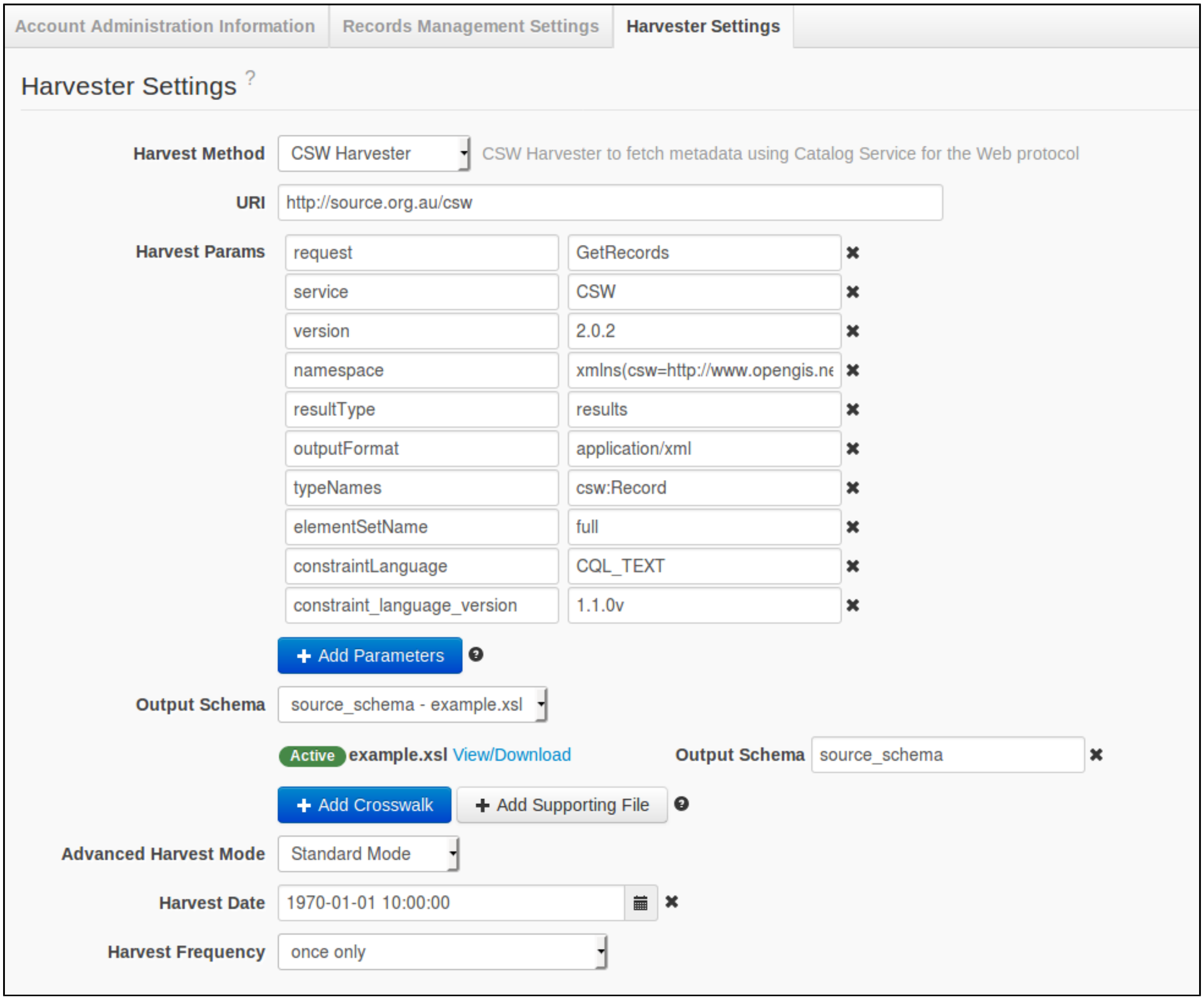
The example XSLT for GMD applies the following rules:

- The Resource described by the source metadata the Dataset is described within a new Registry Collection record (i.e. a Collection, Activity or Service)
- per each Registry record:
    - per each Responsible Party
        - a Party record is constructed
        - the Party record is indicated as related to the Registry record
    - per each Online Resource URL
        - if the resource is determined to be a web service:
            - the resource is described within a new Service record
            - the Service record is indicated as related to this Registry object
        - otherwise
            - the resource URL is referenced within the Registry record as related information.

Within the Harvester Settings of the data source account, apply the following:

- Harvest Method: CSW Harvester
- URI: {CSW interface URI - eg. http//source.org.au/csw}
- Output Schema - rif - if RIF-CS XML is being retrieved from the harvest end-point; otherwise, a crosswalk must be applied:
    - To apply a crosswalk, upload the XSL file using the “Add Crosswalk” button
    - Set Output Schema (on the right of the XSL file name) to the outputSchema parameter value required by the CSW interface call
    - Set selection within Output Schema (above the XSL file name) to “{Source Schema} - {XSL Filename}” - this will set the XSL Transform to “Active” for the harvest
- Advanced Harvest Mode: Standard Mode or Full Refresh Mode
- Harvest Date: optional - set this to schedule a harvest
- Harvest Frequency: optional - set this to a value other than 'once only' if you've entered a Harvest Date and would like to schedule a repeating harvest

Fig 3. Example data source account configuration for CSW harvest from source.org.au


Account Administration Information
Records Management Settings
Harvester Settings

Harvester Settings ?

Harvest Method: CSW Harvester — CSW Harvester to fetch metadata using Catalog Service for the Web protocol

URI: http://source.org.au/csw

Harvest Params

request	GetRecords
service	CSW
version	2.0.2
namespace	xmlns(csw=http://www.opengis.ne
resultType	results
outputFormat	application/xml
typeNames	csw:Record
elementSetName	full
constraintLanguage	CQL_TEXT
constraint_language_version	1.1.0v

Output Schema: source_schema - example.xsl

Active example.xsl View/Download — Output Schema: source_schema

Advanced Harvest Mode: Standard Mode

Harvest Date: 1970-01-01 10:00:00

Harvest Frequency: once only

CKAN API harvest method

The CKAN API harvest method requests data in JavaScript Object Notation (JSON) format from a CKAN Action API. An XML representation of the JSON data is then automatically constructed. This XML representation is subsequently provided as input to a pre-configured XSLT which generates a RIF-CS XML representation of the content.

The pre-configured XSLT itself contains the extent of the logic that determines how many registry records to construct, and their content, informed by the data retrieved. If any of the representation within the RDA Registry is to be altered, this is where to do it. Learn more about CKAN metadata and the transform to RIF-CS.

Currently, the XSLT applies the following rules:

- the Dataset is described within a new Collection record
- per each Dataset:
- per each Organisation
    - a Party record is constructed
    - the Party record is indicated as related to the Collection record

CKAN API harvest configuration

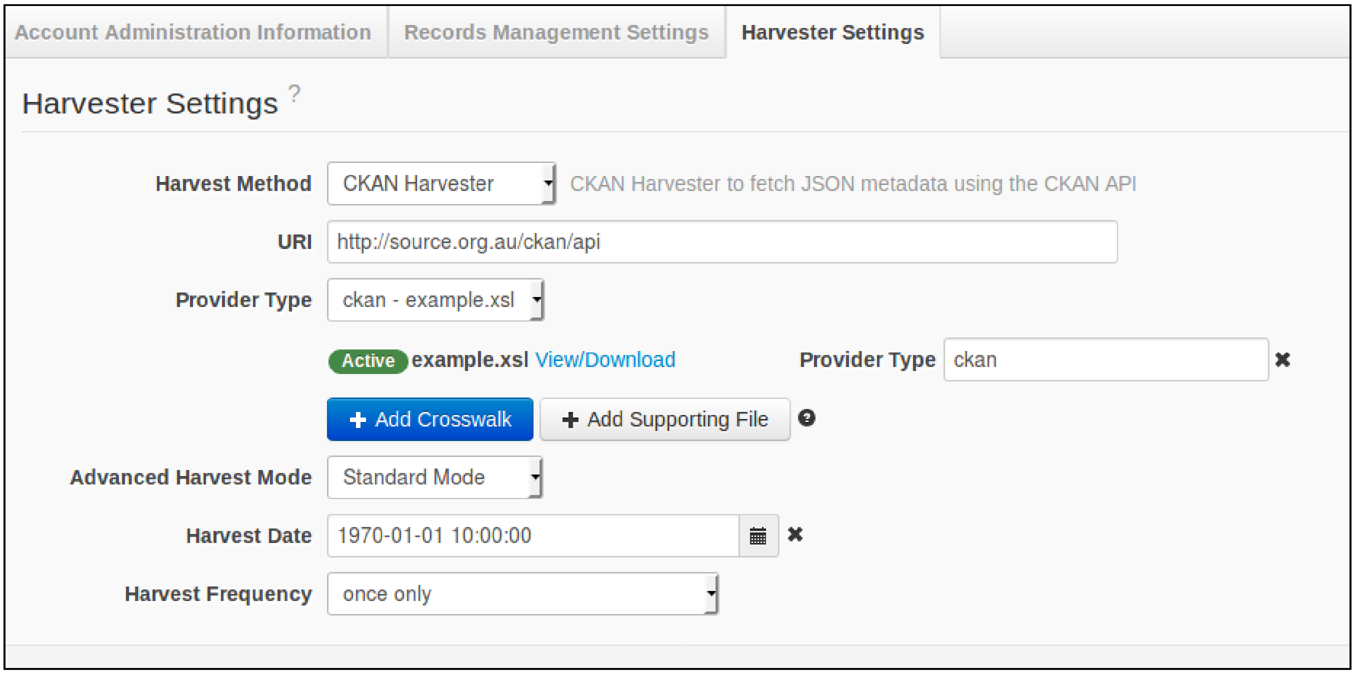
A CKAN API harvest requires a pre-configured XSLT.

Within the Harvester Settings of the data source account, apply the following:

- Harvest Method: CKAN Harvester
- URI: {CKAN API base URI - eg. http//source.org.au/ckan/api}.
- JSON from CKAN Action API will be automatically converted to XML in the same structure as the JSON. A crosswalk will be required to convert this XML to RIF-CS XML:
    - Upload the XSL file using the “Add Crosswalk” button
    - Set Provider Type (on the right of the XSL file name) to ‘ckan’

  - Set selection within Provider Type (above the XSL file name) to “ckan - {XSL Filename}” - this will set the XSL Transform to “Active” for the harvest
- Advanced Harvest Mode: Standard Mode or Full Refresh Mode
- Harvest Date: optional - set this to schedule a harvest
- Harvest Frequency: optional - set this to a value other than 'once only' if you've entered a Harvest Date and would like to schedule a repeating harvest

Fig 4. Example data source account configuration for CKAN API harvest from source.org.au


Account Administration Information
Records Management Settings
Harvester Settings
Add Crosswalk
Add Supporting File

Harvester Settings ?

Harvest Method: CKAN Harvester — CKAN Harvester to fetch JSON metadata using the CKAN API

URI: http://source.org.au/ckan/api

Provider Type: ckan - example.xsl

Active example.xsl View/Download — Provider Type: ckan

Advanced Harvest Mode: Standard Mode

Harvest Date: 1970-01-01 10:00:00

Harvest Frequency: once only