Shared Service Infrastructure (SSI) Aggregator Guidelines for Harvesting

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

www.natlib.govt.nz

March 2012

Document Control

Revision history

Revision	Date	Author	Reason for Change
1.0	25 May 2010	Chelsea Hughes	Document created from previous edition as NRDS Metadata Guidelines version 1.0 NL_CIMS-#228693-NRDS Project: Metadata Guidelines
2.0	31 May 2010	Chelsea Hughes	Incorporating feedback from Gordon Paynter
3.0	28 June 2010	Janet Copsey & Leonie Hayes	Additional changes and further information re ANZSRC codes
3.1	7 July 2010	Chelsea Hughes & Gordon Paynter	Additional changes
3.2	9 August 2010	Chelsea Hughes & Janet Copsey	Finalised and circulated
4.0	5 April 2011	Leonie Hayes and Janet Copsey	First revision for new SSI environment
5.0	6 May 2011	IR Meeting Contributions compiled by Leonie Hayes	Contributions from the IR Meeting held 2/5/11
5.1	20/7/2011	Following review by Auckland Otago, AUT and Waikato Metadata Subgoup	Split document into relevant sections and simplify
5.2	15/3/2012	Emerson Vandy	Edits to reflect transition to production service delivery on Shared Search Infrastructure (SSI).

Table of Contents

1. Introduction

This document contains a set of guidelines for metadata aggregated for the Shared Search Infrastructure (SSI) service offered by the National Library of New Zealand. This service showcases NZ Research previously known as Kiwi Research Information Service (KRIS). It should be read in conjunction with:

- Metadata Guidelines for Contributors to NZResearch.org.nz
- Additional Metadata Guidelines for Contributors to NZResearch.org.nz

1.1. Background

Kiwi Research Information Service (KRIS) was developed in 2007 to build a national research discovery service; and promoting best practice, consistency and the use of standards in New Zealand's research repositories. Both of these goals were dependent on a network of institutional repositories with high-quality metadata.

Metadata Guidelines were developed to assist repositories in creating consistent metadata. Later editions of the Metadata Guidelines were created in 2010 and a major revision was drafted in April 2011 for discussion with stakeholders during May 2011.

The contributions of the stakeholders in May 2011 are reflected in this revision of the document into separate parts.

2. General principles and assumptions

- The harvester used by the Shared Search Infrastructure will harvest all metadata contributed to NZResearch partner repositories.
- The functionality and information presented on the new NZResearch site will be derived from an agreed set of metadata as described in the Contributor Metadata Guidelines document and the Additional Contributor Metadata Guidelines document.
- Any additional metadata beyond this will still be harvested, but will be unused unless requested.
- Some additional metadata from NZ Repositories, like subject, can be harvested by the SSI service and may be normalised for consistency.

3. Derived metadata used in the SSI service

SSI will transform parts of the supplied metadata to improve its findability and usability; however it will not augment the metadata with additional information.

The following metadata fields will be derived by KRIS from the metadata provided by the repositories.

NZIR Internal Metadata field	Importance in KRIS	Source metadata field	Format	Notes
Link	Mandatory	Identifier	HTTP URL.	The Primary Link is the HTTP URL of the metadata record in the repository, and is used to refer searchers back to the originating repository. If a single HTTP URL is supplied, it is used as the primary Link. If multiple HTTP URLs are used, then the primary is selected in this order: • The handle.net URL. • The URL at the source repository (shortest first). • The URL ending in .htm or .html (shortest first). • The shortest URL
Title	Mandatory (Single)	Title	Free text.	The title used to browse by Title. HTML will be stripped, and the words "the" "a" **and "an"** will be removed from the beginning.
Author	Mandatory (Repeated)	Creator	Surname, Firstnames/Initials.	The name used to browse by Author. The system will expect personal names to be in "surname, firstnames" format, and other names to be in **either jurisdictional or direct order.**

NZIR Internal Metadata field	Importance in KRIS	Source metadata field	Format	Notes
Institution	Mandatory (Single)		Institution. **or** **Institution-(name of repository)**	The Institution is used to provide a link to the source institution. **SSI** will assign an institution based on the originating repository**.**
Year	Mandatory (Single)	Date	YYYY	The date used to browse by Date, to limit searches, and (possibly) in citations.. It will be derived from the available Date metadata as follows: For unqualified DC, the earliest date will be chosen. For qualified DC, a date will be selected in this order: • Date issued • Date accepted • Date valid • The earliest date field • Metadata record creation date
EPrints Type Vocabulary	High (Single)	Type	Controlled vocabulary	The Type used to limit searches, limit notifications. The combined EPrints and DCMI terms be used. EPrints and DCMI Type terms will be added to this field.
EPrints AccessRights	Medium	Rights (AccessRights)	EPrints AccessRights controlled vocabulary as text or URI.	Defaults to 'OpenAccess' if not supplied.
NZIR Internal Metadata field	Importance in KRIS	Source metadata field	Format	Notes

NZIR Internal Metadata field	Importance in KRIS	Source metadata field	Format	Notes
ANZSRC Field of Research	High (Repeated)	Subject	ANZSRC code (as a 6 digit number)	The Subject used to browse by Subject, to limit searches, and to limit notifications. ANZSRC Subject codes will be verified and copied to this field. PBRF Subject terms will be cross-walked to ANZSRC codes then added to this field. Unqualified Subject metadata will be matched against ANZSRC terms, and matches added. Unmatched terms will be matched against PBRF codes and cross-walked to this field.
Language	Low (**Repeated**) **To allow for bilingual entries**	Language	RFC 3066 two-letter codes.	The Language used to limit searches. Not currently used in the KRIS interface, but may be added in the future.

4. Type metadata definitions.

A list of common type metadata schema's

Scheme	Maintainer	Size	Examples	Notes
DCMI Type Vocabulary	Dublin Core Metadata Initiative	12	Text Dataset Image	12 terms: Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text. http://dublincore.org/documents/dcmi-type-vocabulary/ (Accessed 2007-02-12)
EPrints Type Vocabulary Encoding Scheme	JISC Digital Repositories Programme	15	Book Conference Item Conference Paper	15 terms: Scholarly Text, Book, Book Item, Book Review, Conference Item, Conference Paper, Conference Poster, Journal Item, Journal Article, News Item, Patent, Report, Submitted Journal Article, Thesis or Dissertation, Working or Discussion Paper. Formal definitions are supplied. The list is a refinement of the DCMI term "Text", and is used with the DCMI Types list to produce a combined list of 27 descriptors. The Eprints authors are also members of the DCMI, so there is the potential parts of it may be promoted for adoption by DCMI. Also, the Eprints project is aiming to embed it in future releases of IR applications (DSpace, Eprints, Fedora, Digital Commons) in the same way that DC is native to these applications currently. http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Type_Vocabulary_Encoding_Scheme (Accessed 2007-02-13)
PBRF List of Research Output Types	TEC	25	Authored Book Journal Article	25 top-level terms, plus refinements of Conference Contribution: Artefact/Object/Craftwork; Authored Book; Awarded Doctoral Thesis; Awarded Research Masters Thesis; Chapter in Book; Commissioned Report for External Body; Composition; Conference Contribution (abstract - full conference paper - conference paper in published proceedings - poster presentation - oral presentation – other); Confidential Report for External Body; Discussion Paper; Design Output; Edited Book; Exhibition; Film/Video; Intellectual Property (eg. patent, trademark); Journal Article; Literary translations, where these contain significant editorial work in the nature of research; Monograph; Oral Presentation; Performance; Scholarly Edition; Software; Technical Report; Working Paper; Other Form of Assessable Output. Likely to change every few years when PBRF is revised. Lacks some important categories that are included in DSpace, such as Dataset, Image, Learning Object, Map, and Recording. See mapping to DCMI and Eprints terms below. http://www.tec.govt.nz/Documents/Publications/PBRF-Quality-Evaluation-Guidelines-2012.pdf (page 49).

Scheme	Maintainer	Size	Examples	Notes
DSpace Default Types	DSpace Project	22	Book chapter Dataset Learning Object	22 Terms: Animation; Article; Book; Book chapter; Dataset; Learning Object; Image; Image, 3-D; Map; Musical Score; Plan or blueprint; Preprint; Presentation; Recording, acoustical; Recording, musical; Recording, oral; Software; Technical Report; Thesis; Video; Working Paper; Other. Not limited to research output types. See mapping to DCMI and EPrints terms below.

4.1. Comparison of EPrints, DSpace and PBRF Terms

The following table illustrates the relationship between the DCMI, EPrints, PBRF and DSpace types and will be used to cross-walk from DSpace/PBRF to ePrints type. It should be read as follows:

- *High-level DCMI Type*: This column contains the DCMI Type fields.
- *Refinements from EPrints Type Vocabulary*: This column contains the EPrints Type Vocabulary terms that refine the DCMI Type "Text" descriptor. The ► symbol is used to show subclass relationships within the EPrints Type vocabulary.
- *Equivalent PBRF Terms*: This column arranges the PBRF Terms against the single matching term from DCMI terms and EPrints. This column can be used to crosswalk from PBRF to DCMI/EPrints terms (but not vice-versa).
- *Equivalent DSpace Terms*: This column arranges the DSpace Default Terms against the single matching term from DCMI and EPrints. This column can be used to crosswalk from DSpace terms to DCMI/EPrints terms (but not vice-versa).

High-level DCMI Type	Refinements from EPrints Type Vocabulary	Equivalent PBRF Terms	Equivalent DSpace Terms
Text	Scholarly Text		
	►Book	Authored Book Edited Book Monograph Scholarly Edition Translation	Book
	►Book Item	Chapter in Book	Book chapter
	►Book Review		

High-level DCMI Type	Refinements from EPrints Type Vocabulary	Equivalent PBRF Terms	Equivalent DSpace Terms
	►Conference Item	Conference Contribution Conference Contribution – abstract Conference Contribution – oral presentation Conference Contribution – other	Presentation
	►►Conference Paper	Conference Contribution – full conference paper Conference Contribution – conference paper in published proceedings	
	►►Conference Poster	Conference Contribution – poster presentation	
	►Journal Item		
	►►Journal Article	Journal Article	Article
	►News Item		
	►Patent	Intellectual Property (eg. patent, trademark)	
	►Report	Commissioned Report for External Body Confidential Report for External Body Technical Report	Technical Report
	►Submitted Journal Article		Preprint
	►Thesis or Dissertation	Awarded Doctoral Thesis Awarded Research Masters Thesis	Thesis
	►Working or Discussion Paper	Discussion Paper Working Paper	Working Paper
Collection			
Dataset			Dataset
Event		Exhibition Oral Presentation Performance	
Image			
InteractiveResource			
MovingImage		Film/Video	Animation Video
PhysicalObject		Artefact/Object/Craftwork	

High-level DCMI Type	Refinements from EPrints Type Vocabulary	Equivalent PBRF Terms	Equivalent DSpace Terms
Service			
Software		Software	Software
Sound			Recording, acoustical Recording, musical Recording, oral
StillImage			Image Image, 3-D Map Musical Score Plan or blueprint
Unclassified / Unclassifiable		Composition Design Output Other Form of Assessable Output	Learning Object Other