ReCAP
Research for Community Access Partnership

UKaid
from the British people

Enhancing the sustainability of the ReCAP Rural Access Library

Phase 1 Final Report

We Are Potential Limited

KMN2113A

August 2018

we are potential

Preferred citation: Colmer, S., et al, We Are Potential Limited (2018). Enhancing the sustainability of the ReCAP Rural Access Library, Phase 1 Final Report, KMN2113A. London: ReCAP for DFID.

For further information, please contact: Simon Colmer, We Are Potential Limited, simon@wearepotential.org

ReCAP Project Management Unit
Cardno Emerging Market (UK) Ltd
Level 5, Clarendon Business Centre
42 Upper Berkeley Street, Marylebone
London W1H 5PW United Kingdom

Quality assurance and review table

Version	Author(s)	Reviewer(s)	Date
1.0	Simon Colmer Nason Bimbe Peter Mason	 Caroline Visser, ReCAP PMU Terry Willat	9 August 2018 23 August 2018 25 August 2018
2.0	Simon Colmer Nason Bimbe Peter Mason	 Caroline Visser, ReCAP PMU	11 September 2018 18 September 2018
		Annabel Bradbury, ReCAP PMU	20 September 2018

ReCAP Database Details: Enhancing the sustainability of the ReCAP Rural Access Library

Reference No:	KMN2113A	Location	Home based
Source of Proposal	AfCAP/AsCAP SC	Procurement Method	Open tender
Theme	Knowledge management	Sub-Theme	
Lead Implementation Organisation	We Are Potential	Partner Organisation	NA
Total Approved Budget	GBP 38'700	Total Used Budget	GBP 13'545
Start Date	28/5/2018	End Date	31/12/2018
Report Due Date	9/8/2018	Date Received	9/8/2018

Contents

Abstract

The Rural Access Library is a repository of rural roads and transport services evidence, containing outputs from the current ReCAP programme and previous programmes. The *Enhancing the sustainability of the ReCAP Rural Access Library* project aims to bring the repository up to international standards and to have all relevant information available for ReCAP to engage in negotiations with a potential future host of the Rural Access Library, hence paving the way for long-term quality, sustainability and transferability of the repository. This Phase 1 Final report outlines the activities undertaken by We Are Potential Limited during Phase 1. It provides a summary of the review that was conducted, the results that were found, and the recommendations that are made for possible implementation in Phase 2.

Key words

Rural Access Library, repository

Acknowledgements

The project team would like to acknowledge the help and support of the ReCAP team in the review activities, as well as Madelein Van Heerden and Adele Van der Merwe from CSIR, South Africa, for providing their perspective on repositories and the ReCAP project.

Acronyms, Units and Currencies

AfCAP	Africa Community Access Partnership
API	Application Programming Interface
The CORE	Connecting Repositories
CSV	Comma separated value (file)
DFID	Department for International Development, UK
DC	Dublin Core
DOI	Digital Object Identifier
ISO	International Organization for Standardization
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OpenDOAR	Directory of Open Repositories
ORCID	Open Researcher and Contributor ID
PaaS	Platform as a Service
QDC	Qualified Dublin Core
R4D	Research for Development (Department for International Development, UK)
RAL	Rural Access Library
RCUK	UK Research Council
ReCAP	Research for Community Access Partnership
ReCAP PMU	Research for Community Access Partnership Project Management Unit
SEACAP	South East Asia Community Access Programme
SaaS	Software as a Service
UK	United Kingdom (of Great Britain and Northern Ireland)
UKAid	United Kingdom Aid (Department for International Development, UK)
URI	Uniform Resource Identifier
WAP	We Are Potential Limited
XLS	Microsoft Excel file format
XML	eXtensible Mark-up Language

1 Introduction

1.1 Background

The Research for Community Access Partnership's (ReCAP) Capacity Building and Knowledge Management strategies[1] focus on strengthening research uptake by practitioners through supporting the documentation, storing, accessing, publication and dissemination of ReCAP research. One of its strategic directions specifically aims at improving access to, and dissemination of, rural road and transport services evidence. To this purpose, a repository has been established on the ReCAP website containing all outputs of the current ReCAP programme as well as its preceding UKAid funded research programmes, the Africa Community Access Programme (AFCAP) Phase I, and the South East Asia Community Access Programme (SEACAP).

The repository, called the Rural Access Library (RAL), currently contains close to 1,150 knowledge items, ranging from technical reports to research papers, blog posts and newsletters, milestone reports, policy briefs, conference papers and presentations.

Key to the success of the RAL is that it is easily accessible and searchable by rural transport practitioners and others, taking into account the limitations of internet access in partner countries.

1.2 Aims

The aim of this project is to bring the repository up to international standards and to have all relevant information available for ReCAP to engage in negotiations with a potential future host of the RAL, hence paving the way for long-term quality, sustainability and transferability of the repository.

It also aims to identify, where possible, a future host organisation that is willing and able to manage the repository beyond the tenure of ReCAP.

1.3 Phase 1 scope

Phase 1 of this project has been framed as an exploratory phase in which ideas, approaches, and existing content are researched and reviewed. The areas explored include the metadata, policies, cross-repository harvesting, possible transfer options and future hosts.

Following these reviews, We Are Potential Limited (WAP) has made recommendations based on their research, experience and expertise to allow the ReCAP PMU to decide on a set of activities to undertake in Phase 2 of the project.

2 Phase 1 tasks

As per the terms of reference, WAP has split the activities in Phase 1 into the following tasks:

1.1 Assessing the extent to which the RAL meets with international metadata standards for research repositories (Dublin Core) and recommending necessary adaptations to make it compliant. Also known as the **Metadata Review.**

1.2 **Identifying and defining relevant policies**, including relevant DFID policies and guidelines, related to metadata; content; submission and preservation; and access (including licensing) to the RAL.

1.3 **Exploring the options for cross-repository harvesting** according to the Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH) and drafting the specifications for adaptation of the RAL.

[1] http://www.research4cap.org/SitePages/Strategies.aspx
http://www.research4cap.org/SitePages/LeadershipDevelopment.aspx

1.4 **Exploring the possibility of, and issues with, transferring the RAL** repository system (MS SharePoint) to other (open-source or proprietary) repository systems, including an assessment of costs for maintaining the repository.

1.5 **Identifying and assessing possible future hosts** for the RAL and recommending the top three candidates. The assessment should be based on a comparison between the potential host organisations on key aspects for sustainable repository hosting and access, and include a list of points for negotiation between the ReCAP PMU and the potential new host.

For each of these tasks, we have listed the methodology used in our work, the results that we found, proposed changes based on our research, experience and expertise, and recommendations for tasks to be implemented in Phase 2. Where possible we have added costs, benefits and disbenefits so that options are comparable and decisions are made easier.

3 Metadata standards and adaptations to make the RAL compliant

As well as looking at the metadata in terms of standards and compliance, WAP also thought it would be very valuable to review the RAL from a content perspective. This additional review would be valuable for the ReCAP PMU as it would give an indication of the scope of the content, where its strengths and weaknesses lie. This would be very useful when having discussions with possible hosts. The other aspect of the content review is to look at how the quality of the content could be improved. Again making the RAL "offer" more attractive to a possible new host.

3.1 Metadata Review

3.1.1 Introduction

An important aspect of the review and development of the Rural Access Library is to audit and identify improvements in the current metadata structures. The aim is to bring them into a standard form so that the library is compliant with digital library and repository good practice.

Having a well-structured and standardised metadata format aids discovery of the content, encourages usage/reuse of the content and above all improves the integrity of the content. Without good quality metadata in a recognised format, the visibility and access to the content will be undermined, however good that content is.

This section aims to identify and review the metadata structure used by the RAL, to determine to the extent to which the metadata aligns to any existing international standards.

We have then mapped the current structures to recommended standards to determine any gaps or amendments, and list recommendations to enhance the data so that it is standards compliant.

3.1.2 What is Metadata?

WhatIs.com[2] defines Metadata as *data that describes other data*. Meta is a prefix that in most information technology usages means "an underlying definition or description" thus Metadata summarises basic information about data, which can make finding and working with particular instances of data easier. For example, **author**, **date created**, **date modified** and **file size** are examples of very basic document metadata. Having the ability to filter through that metadata makes it much easier for someone to locate a specific document, understand its attributes and judge its quality.

Within the scholarly communication space, metadata can be broken down into four categories[3]. We have analysed the RAL metadata in the categories below:

[2] https://whatis.techtarget.com/definition/metadata
[3] http://web.mit.edu/dspace-dev/www/Metadata-schema.htm

- **Descriptive:** The identification and a description of an object. It is used primarily for search and retrieval purposes, providing information about the contents of an object. It provides fields for table-driven searching, as well as a description of the intellectual contents and a physical description of the object. Typical descriptive metadata includes information about the object's source, creation, and content, as well as subject classification and identifying tags.
- **Structural:** Information about the relationships between different parts of an object. It binds together components of complex information objects.
- **Administrative:** Information used in managing and administering information resources within a system or a federation of systems. Typical examples of administrative data are information about rights and reproduction, legal requirements, version control, access restrictions, and statistical and audit trails.
- **Preservation:** Information about the physical specification of an object's creation, its format and condition, hardware and software requirements to render it, its transformation into other formats (change history or "provenance") and its authenticity ("fixity"). The purpose of preservation metadata is to help future generations interpret and recreate the information objects.

3.1.3 Metadata standards

The most commonly used, and basic, metadata standard for research documents and generic content is Dublin Core. More information about Dublin Core can be found in Annex 1.

Table 1 Other metadata standards considered

Schema	Responsible body	Scope and usage
RIOXX[4]	RCUK	UK research, including scientific
Eldis API[5]	Institute of Development Studies UK	The data schema for using data from the IDS OpenAPI and other IDS datasets such as the OKHub.
OpenAIRE[6]	European Union	European based research, now widely accepted and used in other western countries

Within the UK Research Council (RCUK), a new metadata profile, RIOXX, has been created to cater for those attributes that are pertinent to the UK research environment. The RIOXX Metadata Application Profile provides a mechanism to help institutional repositories comply with the RCUK policy on open access. RIOXX focuses on applying consistency to the metadata fields used to record research funder and project/grant identifiers and is designed to support the consistent tracking of open-access research publications across scholarly systems. It has to be noted that integration of the profile can be made to the repository software and modules or add-ons have been created through the support of JISC[7] for the most popular repository software platforms used in UK Higher Education such as EPrints and DSpace.

The OpenAIRE, a European Union initiative has also created standards to help integration of scholarly content with the EU's Research e-Infrastructure and have produced the OpenAIRE guidelines[8]. The metadata can also be expressed within the classic Dublin Core standard as described in Annex 1.

[4] http://rioxx.net
[5] http://api.ids.ac.uk
[6] https://www.openaire.eu/
[7] https://www.jisc.ac.uk/
[8] ttps://guidelines.openaire.eu/

Both RIOXX and OpenAIRE are enhancements that allow for easy integration of content within the larger national and multinational e-research infrastructure. They are essentially Dublin Core and still use OAI-PMH for exposure.

The IDS OpenAPI metadata structure was explicitly developed for managing and searching information about research and other materials useful in a development context. This metadata includes additional information about the metadata location (URL) and richer information about publishing organisations.

3.1.4 Methodology

In order to assess the activities required to make the RAL standards compliant, an export from the SharePoint system was taken on 26th July 2018 and the resulting XLS file was analysed by the WAP team along with the documentation about the purpose and content of each field (as described in the ReCAP Editorial Guidelines).

Having assessed that the RAL metadata structures were reasonably similar to existing metadata structures, we decided it was appropriate for the RAL to use metadata schema based on the Dublin Core metadata set.

We then analysed the metadata in the following areas:

- Using our knowledge of existing standards and likely options for specific standards
- Comparing what others do in the sector
- Mapping to Dublin Core as the basic set (and industry fall-back position)
- Listing possible additional metadata fields to improve interoperability

3.1.5 Review findings

- RAL currently isn't using a standard but already has a good structure, which makes it easier to map to existing metadata schema.
- The table in Annex 1 shows the most commonly used Dublin Core elements that were used in the mapping of the RAL metadata. There was a good match between the available metadata in the RAL and Dublin Core.
- Results of Dublin Core mapping are provided in Annex 3.
- RAL metadata that does not map on to Dublin Core can be found in Annex 4.

Table 2 Summary of metadata review findings

Type	Fields	Findings
Descriptive	RAL has the following fields: – Author – Abstract – Focus Countries – Publisher – Theme and subtheme – Keywords – Title and subtitle – Creation date – Publication date (year) – Document type	RAL lacks the following: – Provenance[9] – Identifiers
Structural		RAL lacks the following: – Relationships

[9] A statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity, and interpretation [http://dublincore.org/documents/dcmi-terms/#terms-provenance]

Administrative		RAL lacks the following: – Legal and rights (licence) information – Version controls – Access restrictions
Preservation	RAL has the following fields: – Creation Date and responsible user – Modification Date and responsible user - only latest modification dates and responsible user is available	RAL lacks the following: – Provenance – Physical specifications
Full text	RAL has the following fields: – URL to full text document (usually hosted by ReCAP) – Filename – Type of file	RAL lack the following: – File descriptions - although it could be optional on implementation* – Physical size of the file (These are usually generated automatically by the implementing software.) * – Standard or full mime type[10] of the file* – Language * The fields are available in the SharePoint installation, but are not displayed in the view (back-end).

Depending on the implementing repository system, a number of additional metadata fields can be inferred by analysing the full text file itself, such as file type, file size etc. The need for such additional fields may not be necessary if the future host already uses software which provides this additional augmentation.

The naming convention of the file is described in the RAL guidelines[11]. This should be retained as any implementing repository system will not be able to automatically replicate this.

3.1.6 Summary

Overall, the RAL has a good quality metadata structure and excellent descriptive data, which has a close correlation with Dublin Core and similar repository standards. The structure and the values therein are well controlled.

The metadata is less comprehensive in the context-related information, which would ensure that the provenance, licensing and interoperability of the content was clear to users.

For the data covering the full text documents, the filename and description are directly managed by the implementing system, extracted from the uploaded file itself. The file naming scheme used is adequate, and the file type is available even if it is not linked to the standard mime type schema.

3.2 Content Review

3.2.1 Introduction

This section aims to determine the scope of the content contained in the RAL and assess it for quality and consistency. It allows ReCAP to understand the range of content that the Library contains, to highlight any possible gaps in the content and metadata, as well as ways in which the content can be augmented to improve the quality of the content and the offer of the Rural Access Library.

We have made recommendations based on our understanding of the needs of the project overall.

[10] A MIME type is a label used to identify a type of data and it is used so software can know how to handle the data. Please see Annex 5 for examples used with repositories.
[11] Metadata form for uploading documents in the Rural Access Library (Version: 2018.01-29)

3.2.2 Why is the scope and quality of the content important?

The purpose of this whole project is to explore suitable options for the transfer of the Rural Access Library to a new host. This will necessitate improving the attractiveness of the repository to a future, long-term host and make it easier to transfer. It is therefore important that the content that is being offered is of as high a quality as possible as well as in an easy to transfer format. The quality of the collection can be measured in a number of ways; in its ability to be understood, integrity and utility.

There are aspects of quality suggested by OpenDOAR guidance and other Open Access bodies[12], which include:

- Completeness
- Having good descriptors
- Having clear provenance statements
- Having clear licensing
- Having the full text of the document, or at least a link to where the full text resides

Unlike data standards, which need to be machine readable, the content does not have to be perfect, so the purpose of this part of the review is to discuss what level of quality is good enough.

3.2.3 Methodology

In order to achieve a collection of high quality content for the RAL, the WAP team have analysed the RAL content, in the following areas:

- Overview of the content
- Overview of main taxonomies; document types (or item types), countries of focus, themes and sub-themes
- Misspellings, input errors and other typographical errors
- Analysis of keywords, both in terms of distribution, and accuracy
- Duplicate record checking
- Document language
- Taxonomy analysis, to identify competitors and other taxonomies or ontologies for mapping.

An export from the SharePoint system was taken on 26th July 2018 and the resulting XLS file was analysed. The detailed findings from this review can be found in Annex 6.

WAP have provided recommendations of suggested changes required to improve the quality of the content, and how those changes could be implemented.

The WAP team also analysed the ReCAP outputs that have been published in Peer Review journals, supplied as a separate file. These contain knowledge items (journal articles) not all currently recorded in the RAL but provide an opportunity to include items which would offer a more complete corpus of the outputs generated by the project and its predecessors.

---

[12] http://v2.sherpa.ac.uk/opendoar/policytool/
https://www.ideals.illinois.edu/bitstream/handle/2142/13968/RepositoryMetadata_CCQ.pdf?sequence=2
https://www.records.nsw.gov.au/recordkeeping/advice/metadata-for-records-and-information/minimum-requirements
https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf
https://rdds.jiscinvolve.org/wp/2016/03/18/how-much-metadata-is-enough/

When assessing most and least popular or frequent entries, boundaries were allocated where appropriate based on the data itself. This approach was designed to give an overview of the content rather than provide any scientific analysis.

3.2.4 Assumptions

I. That, as the list of themes was generated by the steering group, the themes taxonomy is not to be amended – although this might be an opportunity to clarify the overlap between the "Transport Knowledge, Education and Dissemination" theme and the "Transport Research Uptake and Policy" theme as well as to record descriptions of the themes terms, to explain the scope and coverage of each theme.

II. Due to the regional categorisations being inconsistently applied (as reported by the ReCAP PMU), there is a lack of credibility in this metadata internally. WAP feel however that the information about regional focus could be useful for an end user so is included in the analysis.

III. Due to the nature of the repository being solely focussed on the outputs of the AFCAP/SEACAP and ReCAP programmes, the collection is primarily viewed as a repository of project outputs rather than a comprehensive collection of all resources about rural roads.

3.2.5 Review findings

I. Top level statistics

At the point of data export (26th July 2018), there were 1143 knowledge items in the RAL. On the whole the metadata was complete and there were very few missing metadata values. There were no duplicate records and all items had a top level theme. There were some inconsistencies noted, but these could probably be explained by the historical nature of the dataset and are relatively easy to rectify. Most of the knowledge items are saved as PDFs, there were no datasets or media files recorded.

II. Overview of main taxonomies; document types, countries of focus, themes and sub-themes

Document Types

In general, there is a good distribution of documents recorded in each document type. A large number (411 / 35% of total) are associated with procedural project documents (Activity Reports/General overview and Progress/Milestone Reports). This is reflective of the RAL's purpose as a documentation space for all ReCAP activities (and AFCAP Phase 1 and SEACAP activities) rather than just being a repository of information and research about rural roads.

As the programme progresses, and more commissioned research outputs are produced, the number of research papers, manuals, evaluations and policy briefs will increase.

There are a large number of training resources (15%) and Conference/workshop presentations (24%).

Countries of Focus

There is wide geographic coverage within the dataset, which is perhaps indicative of the longevity of various iterations of the programmes. Tanzania has the greatest coverage (13%), then Ethiopia (9%), then Vietnam (9%).

A number of current focal countries show very small numbers of knowledge items; Afghanistan (0.3%), Democratic Republic of Congo (0.9%), Liberia (0.9%), Myanmar (0.6%), Pakistan (0.1%), South Sudan (0.3%).

There are 10 knowledge items with more than 1 country/region and 8 knowledge items with more than 1 country, excluding terms containing a region.

Themes and Sub-themes

All knowledge items have a single top level theme, and only 8 do not have a sub theme.

There is a good spread of knowledge items across the top level themes with the majority (42%) categorised in the *Rural roads and infrastructure research. Transport knowledge management, education and*

dissemination (29%), *Transport services research* (21%) and *Transport research uptake and policy* (8%) make up the remaining top level theme terms.

There are 474 (41%) knowledge items that have only a single sub-theme - this is perhaps indicative of specialised content, therefore may not be problematic, though an opportunity to identify cross-cutting knowledge items may have been missed. As the taxonomic applicability of the knowledge items was outside of the scope of this review, it was not possible to ascertain whether this was the case.

The most common sub-themes assigned to over 20% (roughly 200 or more) of the items each were; *Asset Management and Road Condition, Capacity building, Seals and surfaces for low volume roads, Design*.

The least common sub-themes assigned to under 3% (roughly 30 or less) of the items were; *Needs assessment, Intermediate means of transport (IMTs), Information and communications technology (ICT) and mobile phones, Children, older persons and marginalised groups, Disability, access and universal design, Footpaths, trails and trail bridges, Integration of transport (including waterways).*

III. Misspellings, input errors and other user errors

Titles

There are some spaces at the beginning of titles.

Document Types

Duplication - There appears to be duplication (or at least some crossover) in two of the document types:

Conference presentation and *Conference/Workshop presentation*. Most items in this category are assigned to the *Conference/workshop* presentation term.

It is not clear from the raw data why this has occurred. Both terms have been added to different knowledge items at around the same time which discounts the likelihood of the taxonomy term evolving/expanding to incorporate new types of knowledge item. It has been confirmed that the use of *Conference presentation* is an input error.

Document Format

Unrecorded metadata - There are 5 knowledge items where the document format is not recorded, however on inspection of the filename of the stored file, these are known document formats.

Keywords

There are a number of data inconsistencies and input errors in the Keywords field. This is commonly found in free text fields of this sort.

Inconsistencies:

- Some keywords are in non-English languages
- “English” has been added as a keyword - this language attribute should be added separately - see Language section
- Too many separators (which give rise to low quality, non-specific keywords) e.g.

  Low; Traffic; Volume; Sealed; Roads; Design; Management; Sustainable

  Low Volume Sealed Roads; Design…. would provide more accurate and specific keyword metadata.
- Countries (such as Cambodia, Lao PDR and Vietnam) feature high in the list of keywords. This was probably added before a Country taxonomy was available and these should be mapped to the Country taxonomy and removed from the keywords.
- Inconsistency of similar terms - e.g. *Low volume sealed roads* and *LSVR*
- Inconsistency of plurals e.g. *low volume road* **and** *low volume roads*

Input errors:

- Unhelpful characters present, such as " and ,
- Unnecessary spacing in the keyword, such as some keywords are separated by "; " rather than just ;

  e.g.

  insecurity;scoping;remoteness;rural;market;access;ministry;public works;rehabilitation;transport

  And

  first mile; rural; roads; transport; agriculture; markets; poverty;food security;small-scale farming

  The set-up of SharePoint search delimiters will need to be clarified with the ReCAP PMU before a course of action is recommended. No matter which approach is taken, there are inconsistencies in the data which will need rectifying.

Publisher

There are a number of input errors and possible duplication in the recording of Publishers of knowledge items.

Input errors:

E.g. *Crown Agenta* and *Crown Agents*

Multiple Publishers:

Also, where items are jointly published, the publishers are recorded as free text. As well as the inconsistency this free text input introduces (e.g. the inconsistent use of & or and), it does not allow for the publishers to be separated in the data, and therefore filtered or separated out at a later stage.

E.g.
TRL
TRL and ILO
TRL Ltd
TRL Ltd & ILO
TRL Ltd & ILO Cambodia

Possible duplication:

There are a number of terms where it is not clear if they define a unique publisher or are inconsistencies.

E.g.

T2 Conference

T2 Conference 2017

We would need clarification from the ReCAP PMU to decide whether these types of terms were distinct or should be merged. A process of clarification with be developed at the beginning of Phase 2 to ensure that these types of queries are handled efficiently.

IV. Keyword analysis

We reviewed the keywords in the RAL both in terms of distribution, and accuracy. We used a number of different methods in our review to analyse the content in the keyword field, to produce an overview where ReCAP can understand the scope of their keywording and to test the accuracy of the keywording applied.

In total, the RAL contains 1654 unique keywords (this uniqueness does not count plurals) out of a total of 7140 keywords applied to all knowledge items.

The top 50 keywords are represented below in a word or tag cloud[13] using the tool Tagcrowd[14] to provide a visual representation of the frequency and significance of the keywords contained in the data set.

Figure 1: Tag cloud of 50 most popular keywords

The larger the font size, the more frequently the word is used.

Issues

As well as the inconsistencies and input errors noted above, due to the nature of the system currently used, the keyword data include non-English language keywords. This does not skew the keyword analysis significantly, as the percentage of knowledge items in non-English languages is relatively small.

Another consideration is that the overall view of the keyword scope might be skewed by the historical nature of the content. Once the data is tidied, it will be much easier to create a clearer picture of the keyword scope.

Accuracy

A random sample of 10 knowledge items were taken from the dataset and the keywords recorded in the RAL record for the item were compared to the keywords listed in the PDF full text document. The sample of 10 included items from the date range (2015-2018) and were mainly published by ReCAP for DFID. Due to the fact that keywords are not present in most non-ReCAP published documents, the sample was relatively self-selecting.

Table 3 Keyword comparison between the RAL keyword field and the keywords noted in the full text document

Keywords in RAL	Keywords from PDF
rural roads; community; development; MoU; Memorandum; inception; conference; program support	Afghanistan, rural roads, community development, MOU, annual conference, project support, program inception
baseline survey; road sector research; electric document management system; EDMS	Baseline survey, road sector research, electronic document management system
Boda-boda; Household; Motorcycle; Pedestrian; Piki-piki; Road; Traffic; Injury; Rural	Boda-boda, household, motorcycle, pedestrian, piki-piki, road traffic injury, rural road

[13] https://en.wikipedia.org/wiki/Tag_cloud
[14] https://tagcrowd.com/

community development officers; district engineers; local government; motorcycle crashes; road traffic injury; rural road safety	community development officers, district engineers, local government, motorcycle crashes, road traffic injury, rural road safety.
erosion; control; rural roads; community; participation; technologies; climate change; bio-engineering	Erosion Control, Rural Roads, Community-Participation, Appropriate Technologies Climate Change, Bio-engineering
Gender norms; disparities; Transport Rehabilitation Programme; Gender mainstreaming; Northern Region Pilot Infrastructure Scheme; NRPIS; Policy	Gender norms, Gender disparities, Transport Rehabilitation Programme, Gender mainstreaming, Northern Region Pilot Infrastructure Scheme, NRPIS, Gender Policy
Gender; mainstreaming; rural transport; quantitative research; qualitative research; Inclusivity	Gender, mainstreaming, Kenya, rural transport, quantitative and qualitative research, inclusivity
low volume roads; maintenance; design manual	Low Volume Roads, Manuals, Maintenance, Ethiopia
Motorcycle taxis; unions; track construction; access; social amenities; training; maintenance; safety; empowerment	Motorcycle taxis, unions, track construction, access to social amenities, training and maintenance, safety, empowerment
rural transport; advocacy; Sustainable Development Goals; indicators; Rural Access Indicator; sustainable transport; financing	Rural transport, advocacy, Sustainable Development Goals, financing, rural infrastructure, indicators, sustainable transport, Rural Access Indicator

An additional 2 documents, which didn't have keywords listed in the PDF document, were selected and an editorial assessment on the abstract or executive summary was performed to determine the likely keywords. This assessment was not performed by a subject specialist, but a common sense approach was taken.

Table 4 Suggested keywords where no keywords were noted in the full text document

Keywords	Suggested keywords
Modular; Steel; Bridges	Modular steel bridges; steel bridges; bridges
Vietnam; Rice; Husk; Fired; Clay; Brick; Road; Paving	Rice husk; low maintenance; environmentally optimised; fired clay; brick; low-cost; road; paving

On the whole we found the keywords to be relatively accurate. This increased when keywords were already present in the PDF document.

It is not necessary to implement both singular and plurals of terms as modern repository systems implement search technologies which have the capability to match on either. Many are underpinned by search engines such as Solr[15] or Elastic search[16].

Open Calais[17]

The existing editorial guidelines restrict keywords to 10 terms. The following exercise was undertaken to see if value could be added (and searchability augmented) by extending the keywords beyond the 10 term limit.

[15] http://lucene.apache.org/solr/
[16] https://www.elastic.co/
[17] http://www.opencalais.com/

For this we used Open Calais, an online service, run by Thompson Reuters. It allows you to submit text (by pasting text, or uploading a document), which it then processes and returns suggested semantic metadata based on its analysis.

Four knowledge items were passed through Open Calais and the suggested Open Calais Topic tags, Social tags, and Industry keywords are included below. An option to optimise tagging for research report input was selected.

More documents were attempted, but the service timed out before returning results, possibly because the documents selected were too large to process on the free tier of the service.

Table 5 Open Calais suggested term sample results

ReCAP Keywords in dataset	Open Calais Suggested terms
Erosion; control; rural roads; community; participation; technologies; climate change; bio-engineering	alternative solutions;appropriate technologies;energy;erosion site;long-term solution;road drainage systems;road infrastructure;road maintenance camp infrastructure;road network;road networks;road organisation management;rural networks;rural road network;satellite pull;serviceable rural road network;social services;transportation
Community development officers; district engineers; local government; motorcycle crashes; road traffic injury; rural road safety	finance;Government / Politics;Ground Accidents / Collisions;Ground Freight & Logistics;Health / Medicine;Highways & Rail Tracks;Labour / Personnel;Land transport;Motorcycle safety;Motorcycling;Passenger Transportation, Ground & Sea (TRBC);Performance / Results / Earnings;Road;Road Freight;Road safety;Road traffic safety;Safety;Telecommunications Services (TRBC);Traffic collision;Transport;transport network;Worker road safety
Gender; mainstreaming; rural transport; quantitative research; qualitative research; Inclusivity	Gender studies;Feminism and society;Gender;Gender mainstreaming;Mainstreaming;Orwa;Public policy;Women
Rural roads; community; development; MoU; Memorandum; inception; conference; program support	roads infrastructure;satellite mapping;security contractors;technology innovations;transportation; knowledge management

The results from this data augmentation process were varied and possibly too general for the specialist nature of the RAL. It was, however, an interesting exercise and could be used as a helpful suggestion tool if the keyword field/process were to be extended in the future.

V. Duplicate record checking

Duplicate checking was performed across the RAL dataset, looking at both the link to the full text document as well as title. As the title and subtitle are recorded separately, these two fields were concatenated to create a single field to check duplicates against.

Neither the full text document filename nor the knowledge item title showed any instances of duplication.

VI. Document language

It has been noted by the ReCAP PMU that although there are knowledge items in languages other than English, there are no language attributes recorded in the data. The review looked at both the language of the full text document and the language of the metadata itself (i.e. what language is the title and keywords in).

Full Text

As anticipated, the majority of the content is in English (92%). There are 6 other languages featured in the RAL; French, Lao, Vietnamese, Khmer (Cambodian), Portuguese and Spanish (in order of count).

There are three bi-lingual knowledge items. If language attributes are added, then special care would be needed to describe these latter items accurately in the metadata.

Metadata

All of the French, Portuguese and Spanish knowledge items identified above, also have titles and keywords in the language they are in. Other metadata attributes do not have non-English translations e.g. Themes and Sub-Themes, so do not have non-English versions.

Connecting language versions

At present, there is no way of identifying whether a knowledge item has a version in another language. We would recommend creating a method by which you can connect items to each other. To achieve this, a unique identifier would need to be created. Currently the only unique identifier is the URL of the full text document, and each language version has a different URL.

VII. Taxonomy analysis - competitors/other taxonomies/ontologies we can map to

As a sector-specialist organisation, ReCAP has developed a robust and comprehensive thematic taxonomy, supported by experts in the field. In our review, we have not found any obvious candidates for alternative taxonomies that feature relevant terms across the whole of the ReCAP programme.

We have identified a number of comparable document type taxonomies. The ReCAP taxonomy has grown organically depending on the project outputs. If different document types were added to the collection, then these comparable taxonomies might be useful for suggesting terms.

Comparable taxonomies are listed in Annex 7.

VIII. ReCAP Published Journal Articles

During discussions with ReCAP PMU, it was noted that not all ReCAP items that were published in third party journals were recorded in the RAL (although some have been added). ReCAP PMU provided a document that was used for internal management purposes and noted some bibliographic details for each published article (and also included articles not yet published).

WAP suggest that this is a good place to start to try to integrate the remaining knowledge items into the RAL. At present, only knowledge items that have the full text held on the RAL are recorded in the dataset. WAP has suggested that the RAL should also include items not held on RAL servers and where this occurs the URL or Digital Object Identifier (DOI) of the knowledge item should be recorded.

On review of this spreadsheet, it is clear there is a small amount of editorial work needed to incorporate the knowledge items into the RAL. For example, as the spreadsheet is used for internal purposes, several of the RAL taxonomies have not been applied. This would entail a short editorial task, categorising the documents by Theme/Sub-theme, Country of Focus, and adding keywords.

There is also a small amount of data tidying including using delimiters to separate out authors and two items do not appear to have an external URL for the full text document.

It is suggested that a new document type of Journal Article be created for these knowledge items and existing knowledge items in the RAL amended to this document type.

Table 6 SWOT Analysis of metadata

Strengths	Weaknesses
– Sensibly organised and marked up metadata – Good spread of knowledge items across the thematic areas – Good sized repository of specialist knowledge – Roughly in line with existing and comparable taxonomies	– Lack of metadata standard – Lack of provenance, licensing and language attributes – Some inconsistencies in metadata (due to free text input availability and changing systems and capabilities over time)
Opportunities	**Threats**
– Open Access could be the publishing default by 2020 (when the project comes to an end) – Extend keywording to include more terms (which would require reassessment of all knowledge items, as well as reassessing the 10 keyword per item limit) – Provide more clarity of provenance, licence and language to make the RAL more attractive and usable to other low-volume road resources	– By the time host is found, technology and practice might have moved on

4 Identify and define relevant policies

4.1 Introduction

This section aims to determine the most appropriate policies for ReCAP to implement in the delivery of the RAL. It allows ReCAP to understand what different policies can offer, what is needed to develop the policies and prioritise which policies to implement.

We have made recommendations based on our understanding of the needs of the project overall.

4.2 Why are policies important?

As in any everyday life activity, there is need to provide clear rules and protocols that will govern the behaviour of all interested stakeholders of the activity. Policies are the best way of encouraging such a good and consistent behaviour and the RAL is no exception. It is therefore important that the RAL and the service provider of the library are very clear about the rules and protocols that users and content providers must adhere to as well as a clear articulation of the obligation that the RAL and its provider will give to its stakeholders.

The main aims of every digital library or repository is to collect, curate and then distribute knowledge to the users that are best placed to use it in creating interventions and innovative solutions that will make people's lives better.

Funders of research have been at the forefront in encouraging wider use of the research they are funding, aiming to eliminate access barriers to the outputs, as well as encouraging use of standards that encourage linking of data with other services. This encouragement has been reinforced with open access policies provided by the funder as a requirement of the grant.

Policies for the RAL should therefore be clear as to how the content is collected, the extent of the collection and how it is managed and maintained. This is crucial in providing a sense of quality of the content and its integrity.

Table 7 Policy requirements for stakeholders

Stakeholder group	Questions and concerns If I was going to use the data, what would I want to know?
End-users Those who wish to download and use content from the library	– What is the scope of the RAL? – What does it cover? – Can I use the content? And in what context? – Can I use the content in an automated way and create derivatives i.e. when doing text and data mining?
Data users Those who wish to harvest, host and re-share the content	The above list, plus – Will the URIs (or access points) be persistent, so that I don't have broken links in my data? – How long will the set be available?
Researchers Those who wish to use, and cite content from, and contribute content to the library	– Who is allowed to deposit research content in RAL? – How can I submit my research into RAL? – How is Intellectual Property going to be handled? – How long will they store my research? – Can RAL provide me with insights on how my research is being used?
Funders Those who have invested in the project and want to see results/impact or invest in new areas	– Which outputs are coming from the research I am funding? – How are you complying with my OA policy? – How can I understand the reach of the research I am funding? – How can I ensure what I'm funding advances or builds on existing work without duplication?
RAL contributors/providers Those who have contributed to the body of work in the RAL	– Am I being properly represented and attributed for my work? – Can I take down anything which isn't correct? – Can I signpost others to my work? – Can I continue to see how much my work is being used?
ReCAP	– Have we been clear on our obligation to the various stakeholders of this library? – How do we mitigate any liabilities on the content we are hosting? – Do I know how to respond when requests for data (full text items, dataset and metadata) are made? – Can I feel confident I am meeting best practice requirements?

4.2.1 Research

I. Methodology

In order to provide clear and robust set of policies for the RAL, the WAP team have:

- Analysed the relevant ReCAP policies and strategies relating to communication, archiving, editorial overall direction
- Analysed the digital library and digital repository policies that constitute good practice
- Analysed funder mandates and policies that will have an influence on the policies that will be required for the RAL
- Selected the policies that will be required for the RAL in order to conform to best practice
- Provided recommendations on the policies required and suggested how they will be created and operationalised for the RAL.

II. An assessment of current policies

The OpenDOAR entry for the RAL as shown in Annex 8 does not list any policies for the library and there are no formal policy documents held. Two documents were provided by ReCAP for review which contain some aspects of policy.

- *Knowledge Management and Communications Strategy prepared for DFID in September 2015*
  The document discusses the aims of the repository (RAL), identifies its stakeholders and also some of the content that will be hosted. It does not precisely define any clear policies regarding RAL content, its management and use.
- *DFID Research Open and Enhanced Access Policy: Implementation guide (V1.1: January 2013)*[18]
  A guide produced by DFID to help grantees understand the policy and how they should apply it to the work they are undertaking that is funded by DFID. The guide discusses how the 'outputs' should be made Open Access and this includes the metadata standard requirements, licensing regimes to be used and the repository that can be used to host the outputs. The guidelines also discuss how datasets can be handled as well as book chapters and how to publish in open access journals. Please note that all Research Councils UK funding bodies including DFID requires that outputs[19] from the research they fund is Open Access.

In this context they are considered guidelines and not policies per se, but provide a direction as to which policies could be created and applied to the RAL.

There may be more organisational level policy documents, for example from Cardno Emerging Markets (UK) Ltd., but these were not available for review during Phase 1.

It is our conclusion that no standard policies exist for RAL that explicitly conform to standard repository or digital library good practice.

III. Good practice

One of the key mitigation factors in making the accessibility, use and reuse of content frictionless is having good policies. Policies also help in providing the credibility to the content in terms of perceived quality and utility of the research.

As a minimum there is a need for policies that describe who can use RAL, what type of content is held, the terms and conditions of access, use/re-use and preservation actions.

Many of the good practice policies and the sector-wide initiatives working shared standards have come from the academic library sector. This collaborative approach has culminated in OpenDOAR, which provides a useful best practice framework for publishers.

Examples of good practice guidelines/frameworks

OpenDOAR[20] is the quality-assured global directory of academic open access repositories which enables the identification, browsing and search for repositories, based on a range of features, such as location, software or type of material held. The major criteria for Inclusion and Exclusion is that OpenDOAR collect and provide information solely on sites that wholly embrace the concept of open access to full text resources that are of use to academic researchers. Thus sites where any form of access control prevents immediate access are not included: likewise, sites that consist of metadata records only are also declined. Clear policies and licensing regimes are critical as to metadata, accessibility and interoperability, therefore aligning the RAL to the OpenDOAR ethos will be important.

[18] DfID Research Open and Enhanced Access Policy https://www.gov.uk/government/publications/dfid-research-open-and-enhanced-access-policy
[19] Outputs here includes publications, research data, software, impact narratives.
[20] http://v2.sherpa.ac.uk/opendoar/information.html

- For online publishing, the UK government has useful Government Digital Service (GDS) guidelines[21] in a number of areas, including web publishing. Some DFID contracts stipulate that grantees follow these guidelines.
- The Open Data Institute, based in the UK has extensive background of support of governments in developing countries in developing and maintaining Open Data initiatives. They have provided guidance on writing an Open Data policy.[22]
- The UK's Economic Social Research Council has an excellent impact toolkit[23], and more recently through the Impact Initiative[24] has published an Impact Lab of useful case studies.

IV. Stakeholder policies and the impact on RAL

There will be stakeholder policies which impact on what is required of ReCAP, or provide useful examples of policies which ReCAP wish to adopt.

The content of many repositories is restricted by policies which relate to their scope. Examples include:

- Geographical coverage
- Publishing organisation
- Licence or cost (including Open Access)
- Document type
- Thematic coverage

For example, many university libraries and repositories have policies to explicitly only host content created or published by themselves.[25] The UK High Education Institutes (HEI) will only host research content generated by their own researchers. This ensures compliance with the Research Excellence Framework.[26] The Institute of Development Studies (IDS) on the other hand does not prescribe such restrictions as most of the research comes from research projects that may include researchers that do not work for IDS.

The Eldis[27] dataset contains only materials which are freely available and free to download and subject repositories e.g. for those used for preprints such as arXiv[28], BioRxiv[29] will only accept content that fall within the subject domain. Some of these subject repositories may also enforce that the content is of scholarly nature in that they should be peer-reviewed and/or must include references.

The relationship between ReCAP and future hosts will be in part an alignment between the content policy of the RAL and that of the hosting organisation as much as cost or other considerations.

Relevant policies

The following list of policies, in Table 8, have been identified as relating to the needs of the RAL.

[21] https://www.gov.uk/government/organisations/government-digital-service
[22] https://theodi.org/article/how-to-write-a-good-open-data-policy/
[23] https://esrc.ukri.org/research/impact-toolkit/
[24] http://www.theimpactinitiative.net/
[25] https://www.ed.ac.uk/information-services/research-support/research-data-service/sharing-preserving-data/data-repository/service-policies/submission-policy
[26] http://www.hefce.ac.uk/rsrch/ref2021/
[27] https://www.eldis.org/
[28] https://arxiv.org/
[29] https://www.biorxiv.org/

Table 8 Policies relevant to the RAL

Name	Description
Collection Policy	The policy covers all activities involving what content is and will be collected, curated and preserved. The policy will cover the following areas: – Scope of the collection – Thematic Focus – Geographic Focus – Sources – Languages – Management of the collection
Metadata Policy	The policy contains information regarding how the metadata content should be used/reused by users. The policy articulates the licensing regime used for the metadata, how it will respect licences of the content collected, and how enforced for providers. Information of what a user can and cannot do with the metadata will also be made clear in this policy. Essentially the policy will articulate clearly the following: – Access to metadata – Re-use of metadata
Data Policy	The policy contains information regarding how full text and/or other full data items should be used by users. The policy articulates the licensing regime; how other licences are respected, especially for full text which is still under copyright of the publishers. Information of what a user can and cannot do with the full text or data will also be made clear in this policy. Essentially the policy will articulate clearly the following: – Access to full items – Re-use of full items
Submission Policy	This policy concerns depositors, quality and copyright. The policy will articulate clearly the following: – Eligible depositors – Deposition rules – Moderation – Content quality control – Publishers' and funders' embargos – Copyright policy
Content Policy	This policy describes the types of content collected and how it is treated once collected into the digital library. The policy compliments the Collection Policy. This policy will also articulate issues such as: – Repository type – Type of material held – Principal languages
Preservation Policy	This policy guarantees continued access to the content including the integrity of the content for a 'mandated' period of time. This policy will articulate issues such as: – Retention period – Functional preservation – File preservation – Withdrawal policy – Withdrawn items – Version control – Closure policy
Deposit Licence Agreement	This acts as a contract between the depositor and a library. It will stipulate what is expected from the depositor in terms of the content and licensing, and also spells out clear obligation of the host on how the depositor's content will be managed throughout its life in the library. This helps to safeguard the library's integrity.
Copyright and liability statement	Provides information regarding the general handling of copyright of the material and a disclaimer on the use of content in the library. It also provides mechanisms to absolve responsibility for the validity of the content - this would include commonly seen text such as the 'opinions expressed in the content is that of the author/s and not the RAL'.
Data protection policy	Provides information on obligations on holding of personal data whether about employees, partners, or as subjects of research. This will include ethical considerations and may cover the secure storage, transfer and disposal of data.
ICT policy	Provides information on how the host ensures the resilience of the computer systems used and also ensures proper use of the ICT systems.
Social Media Policy	Provides guidelines of how to use social media effectively and in line with the organisation ethos.

Open Access and Open data policies	Although primarily used for guidance to enable transparency and the widest distribution of datasets, open access and open data policies are particularly useful for organisations in articulating their principles and commitment to openness. It may also be helpful in auditing activities and reporting to funders on commitments to open publishing.

5 Exploring options for cross-repository harvesting and drafting a specification

ReCAP has recognised that it is imperative that the metadata of the RAL should be able to be harvested automatically by other repositories, third-party discovery services and content aggregators. This is an important conduit to sharing the RAL content across many different platforms and ultimately getting more impact from the ReCAP work. The most common way of enabling cross-repository harvesting in the sector is to use the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

In Annex 9 we discuss further the benefits of cross-repository harvesting and outline why OAI-PMH is the market leader.

We have considered the options for enabling OAI-PMH capability as part of our review on the transfer options of the RAL. We feel that the approach and specification is entirely dependent on the technical configuration of the host that ReCAP chooses to take on for the repository, and any chosen method of transfer. The selection of OAI-PMH and its specification are the same whether to facilitate the transfer process from RAL to another host, or if the transfer is facilitated by another means and the new host makes the data available via OAI-PMH through their technical infrastructure.

We have presented a number of options in the transfer section, all of which take into account the cross-repository harvesting component.

6 Routemap for technical transition of management, publishing and sharing of RAL library content

6.1 Introduction

ReCAP wishes to transition from the current situation, where the organisation is holding, managing and publishing RAL content and full text documents, to a situation where the content is wholly hosted by a third party, with responsibility for managing and publishing the content, both as a public-facing interface and a machine-readable dataset.

This section outlines the potential routes to achieving the end goal and the steps required to migrate the data in each case. It is worth noting that this section considers the technical options for data transfer and publishing, but does not cover the policy, governance and ownership aspects of library migration.

6.2 Current position

- The metadata and full-text documents are for the most part held on the ReCAP servers as a SharePoint library and filestore.
- This content is being published via the ReCAP website, with the exception of a very limited number of research outputs published in external open access journals.
- There is no external programmatic access to the dataset as a whole.

6.3 Aims

The aims of the transition work are that:

1. The data is held and published by a responsible 3rd party host

2. As a minimum, the data will be held, as is, as an archive, where no additional materials are added. An extended option would be for the data to be actively managed and added to with new content. In both cases, all programme outputs would need to be added to the dataset, even if published after the programme end.
3. Content is available through at least one appropriate 3rd party website
4. Metadata is available via an OAI-PMH interface to allow computer-to-computer access to the dataset
5. Full text documents are accessible via a persistent identifier (URI)
6. The quality and scope of the metadata is maintained as far as practical in the migration, and enhanced where possible.

Table 9 Overview of the RAL transition process

	Starting point	End point
Data management	Data held, managed and published by ReCAP	Data held, managed and published by 3rd party
Metadata publishing	Content available via ReCAP website	Content available via 3rd party website
Dataset access	No external access to dataset	Metadata machine readable via OAI-PMH
Full text documents	Held on ReCAP and on third-party sites. Temporary ReCAP URLs	Held on third-party servers/sites Persistent URLs (with optional DOI entries)

6.3.1 Migration pathways

A migration of content from the current system may be characterised as a set of stages outlined in Table 10. This is intentionally a simplification of the stages required, each of which will comprise a number of more detailed tasks.

Table 10 Migration pathways of the RAL transition process

	Stage	Options	Outcome
1	Updating content	Manual and automated data enhancement	Data validated, enhanced where possible and made standards compliant
2	Exposing data	Data export file(s) Creation of a machine-readable interface	Data available to third party systems in a readable digital format
3	Data transfer	Copy via file transfer Data harvesting Data import process	Content transferred to new host.
4	Content publishing	Maintain presence on ReCAP site	Content available to public via 3rd party website(s) or online repository
5	Data publishing	Exposing of RAL via OAI-PMH	Dataset open to all, machine readable.

Stage 1 deals with preparation and is expected to take place on the ReCAP systems, to the extent that the data structures are mappable to an existing minimum standard.

For interim stages 2 and 3, exposing the data and transfer, there are a number of options. The choice of option will depend on availability and complexity of the systems available, and may be driven by the capabilities of the future hosting environment.

The future host may allow for the content to be available directly as a file (e.g. CSV, XML), in native SharePoint format (e.g. via the SharePoint API), or via an alternative protocol.

Given the likely types of organisations and systems that we expect to be candidate future hosts of RAL content, here we consider the preferred alternative protocol to be OAI-PMH, which has gained almost universal uptake in the sector.

Stages 4 and 5 for publishing will be dependent on the ultimate hosting provider, but are included here and required to complete the transition process. In all cases other than Option C2 (below), some work will be needed to maintain the RAL with a user-friendly front-end displaying all the content from one place.

6.3.2 Interim stage options

The possible options for interim stages may be outlined as follows:

A. Enhance the existing systems to provide a documented, standardised **file export**

Third party hosts would map and import RAL content into their own systems.

B. Expose content via a **SharePoint API**

Allows third-parties to develop applications to read the content directly from the RAL SharePoint application.

C. Expose content as OAI-PMH standards-compliant interface

Allows hosts with OAI-PMH harvesters to map and import content using this standard protocol.

To enable this, ReCAP could:

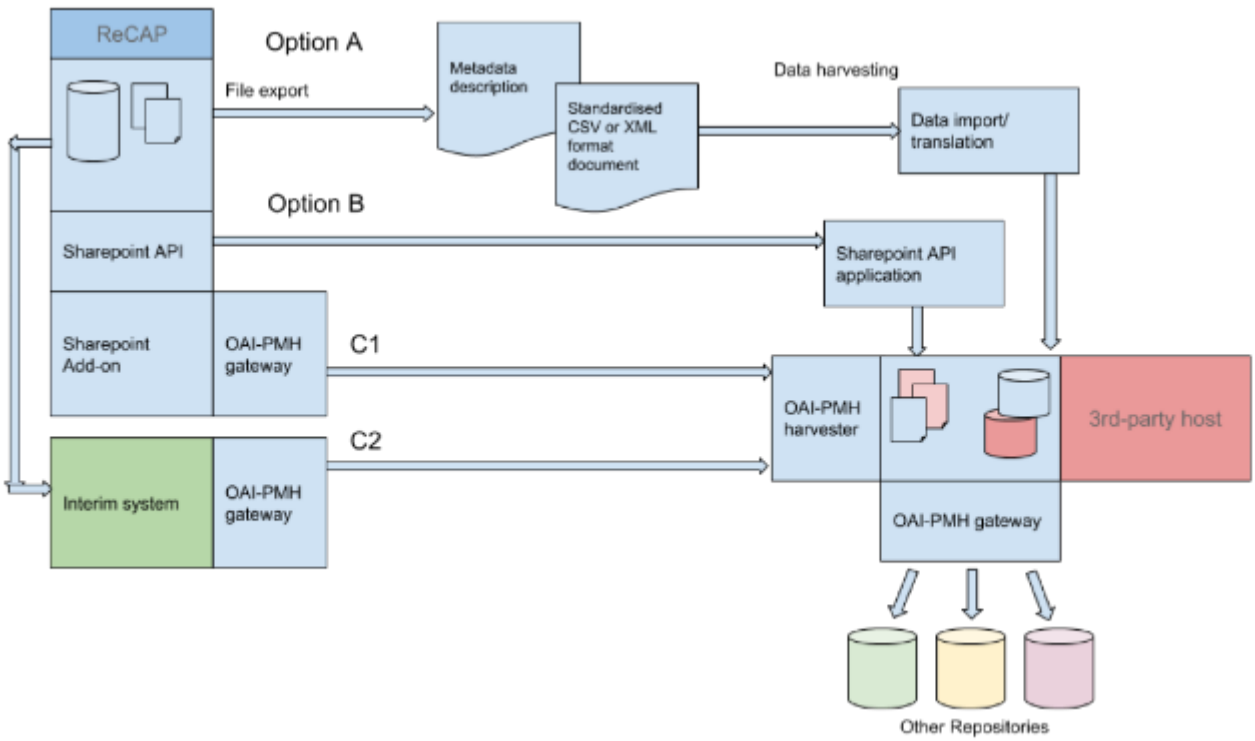
C1. Build an **OAI-PMH wrapper for SharePoint**

Develop or implement a tool which interprets OAI-PMH commands into data queries and return results to external harvesters.

C2. Migrate to an OAI-PMH-enabled **interim system**

Set up an OAI-PMH-enabled application, hosted by ReCAP or other service provider, and migrate content across to this system ready for use by a third-party.

Figure 2 Interim options schematic



6.3.3 Option analysis

Here we give an overview of the advantages (to ReCAP) of each approach, problems which might arise or limit implementation, and possible resource implications.

Option A: File Export

In this option, the content would be exported directly from SharePoint ready for import into a new system. As some re-working of the content will have happened prior to export, following the metadata and content review, the exported files would be well-structured and comprehensive. Although some further transformation, wrapping and labelling of the data could happen post-export, it is expected that the receiving system would do most of the work in import, mapping onto the new internal data structures.

For many systems (such as DSpace), this cost would be mitigated by the fact that the import interfaces are relatively straightforward.

Table 11 Transfer Option A: Benefits, issues, costs

Benefits	Issues	Costs
Relatively simple for ReCAP to implement and therefore will require fewer resources and reduce impact on the existing system.	Relies on 3rd party import tools being able to understand the data structures and be able to map into the new system. Some work is required by the new host. Some more detailed or nuanced attributes of the data may be lost.	For ReCAP, some configuration to ensure all required data is included in the export files and the metadata descriptions are clear. A few days' work for the team is expected. The export could be enhanced with some transformation For the hosting system, a range of implications from basic data mapping through to programming to process and transform content would incur some costs.

Option B: SharePoint API

The Microsoft SharePoint system has a number of options for exposing content via an API. See

https://docs.microsoft.com/en-us/SharePoint/dev/general-development/choose-the-right-api-set-in-SharePoint for more details.

This option would be most interesting in the event that a suitable host could be found which has the capability to use this type of API. Otherwise the costs for the hosting organisation are likely to be prohibitively high.

Table 12 Transfer Option B: Benefits, issues, costs

Benefits	Issues	Costs
The SharePoint API is a component part of the system so all work would take place within the ReCAP SharePoint system, which is already known to the ReCAP team. Content available via a documented interface. In the short-term, publishing could still happen via the ReCAP website.	Need to decide best API approach before we know destination system The number of systems with SharePoint API capabilities will be limited, reducing the choice of potential hosts	For ReCAP, the configuration of the SharePoint API would have a resource implication. For systems which already encompass a SharePoint end-point solution the costs would be relatively low. In the event a new host system would need to develop a new extension to work with SharePoint, a medium to larger cost could be expected.

Option C1: SharePoint OAI-PMH extension

For discussion of the viability of this option, please refer to Annex 10.

Other than these noted examples we have not been able to identify an existing working solution for this option. We have attempted to contact the developer of the first option, but have not had a response at this

stage. Regardless of whether the existing code is available, some development would be needed to tailor the solution for use by ReCAP and for this reason, we would consider this a medium cost option, but with a higher risk.

Table 13 Transfer Option C1: Benefits, issues, costs

Benefits	Issues	Costs
Enables an OAI-PMH endpoint for likely hosts to connect to, which will allow a large part of the data transfer to happen automatically.	This would best support a one-off data transfer, as the interface would not offer incremental updates between systems. From the point of extraction, amendments to the dataset would happen on the new host.	For ReCAP, (re)development of a wrapper for SharePoint to expose data in an OAI-PMH compliant format would incur an initial cost. For a third-party with an OAI-PMH harvester, the cost to import the content would be relatively low

Option C2: Interim OAI-PMH system

In our potential systems document in Annex 11 we discuss the merits and costs of a range of potential interim systems which ReCAP may wish to transition to in the short term.

As it involves the building or setting-up of a new system, we consider this to have higher costs, but offer a more appealing transition route for potential hosts. In this case, the host could agree to take on the hosting and maintenance of the new system.

Of the systems identified, those which require a subscription model over an up-front cost are probably more appealing to ReCAP as the overall cost is likely to be lower.

Table 14 Transfer Option C2: Benefits, issues, costs

Benefits	Issues	Costs
Content could continue to be maintained on the interim ReCAP system and harvested by the third-party system as required. ReCAP website would continue to operate the library as long as needed Offering an OAI-PMH interface would appeal to majority of potential hosts. Potentially a complete repository system might be attractive to a future host that required a repository but did not have the resources to initially set it up. The host would take complete control of the developed system	Significant work will be needed to transfer to a system which may only have a short lifespan, given the expected project closure date.	For ReCAP the cost of either commissioning or implementing a new system or transactional costs of purchasing or building and configuring a system would be medium to high (depending on the solution chosen). This would likely need to include ongoing budget until project closure. For a third-party with an OAI-PMH harvester, the cost would be low, to import the content.

6.3.4 Costing summary

Before detailed analysis of potential host systems and their capabilities are known, it is difficult to be accurate about costs. Table 15 gives an indication of relative costs of options and the balance between the work required by ReCAP and potential future hosting partners.

Option A would not require the commissioning of new systems and therefore would have the lowest expected overall cost, leaving more resources for tailoring the data effectively. Although the development of new software is included in some of the options, the overall cost implication is probably best evaluated in terms of paying a third-party to support an interim system.

Table 15 Transfer option cost comparison

	Options			
COSTS	A: File export	B: SharePoint API	C1: SharePoint OAI	C2: Interim system
For ReCAP	LOW	LOW	MEDIUM	MEDIUM - HIGH
For future host	MEDIUM	LOW-MEDIUM[30]	LOW[31]	LOW

Key: LOW < GBP5,000, MEDIUM = GBP 5,000-10,000, HIGH > GBP 10,000.
All costs include an allocation for person time.

The more work which ReCAP undertakes before transition, the lower the costs and barriers for potential hosts, and therefore the greater the chance of attracting a host.

7 Identifying and assessing hosts

7.1 Introduction

This section provides an overview of potential future hosts for the Rural Access Library, an assessment of provision in the sector, and some of the characteristics and potential concerns of different types of hosts.

It is not intended to be an exhaustive list, but to provide a way of scoping the types of likely collaborators and key players in this knowledge space. A broader survey of potential hosts and practitioners was not required at this stage. For further work to be undertaken in this area, we would suggest working with the ReCAP PMU to define a Terms of Reference for a new host. This would take into account possible scenarios, which would include the minimum option of simply statically hosting the repository as an archive, as well as the extended option of actively managing and adding to the content if the future host saw the potential in the collection.

The identification of possible hosts was based on the following activities:

- Internet searches (for Rural Access resources)
- ReCAP network: possible ReCAP steering committee(s) members and contacts
- We Are Potential development sector knowledge
- Document review

A long list of potential hosts was established and can be found in Annex 12.

Three hosts were then contacted to build an understanding of the remit, constraints, barriers and concerns they have over publishing content, especially where it is not their own. These hosts were chosen by WAP and ReCAP as examples of distinct types of potential host.

7.2 Findings

7.2.1 Internet search

Using Google searches with thematic terms, types of provider, and geographic keywords, results were reviewed to identify candidate hosting websites.

These websites may be broadly classified as sector-related, generic (global) archiving services, funder-supported platforms, national and/or regional focussed.

[30] Assuming the host system already has a SharePoint extension
[31] Assuming the host system has OAI-PMH import capabilities

We have also listed some related services which may be of interest as linked systems and users of the library material.

7.2.2 ReCAP network and contacts

At this stage, it was decided to approach one recent ReCAP collaborator and a group call was held with Council for Scientific and Industrial Research (CSIR). It was decided not to contact other partners at this stage.

Issues were raised about the capacity of organisations to be able to maintain and manage required repository systems. Further discussions would need to take place to determine if they were a potential host for the RAL.

7.2.3 Review of existing research

Two specific documents were reviewed:

- Review of *Feasibility Study of Options for Long Term Knowledge Sharing and Management: Final Report* (Paul Starkey and International Forum for Rural Transport and Development (IFRTD), June 2013)[32]
- *Institutional Capacity for Knowledge Management of Transport Research Centres in Africa and Asia, Workshop Report* (Ruud Crul for ReCAP, November 2016)[33]

There is clearly limited capacity in the sector and research institutes and some efforts have been undertaken by funders to address these needs. However, the cyclical nature of funding initiatives means that knowledge services are likely to stop when projects come to an end. It is recognised that the desire to see the RAL sustainability hosted is an attempt to mitigate against this situation recurring.

7.2.4 Suggested criteria for assessment of potential hosts

Suggested draft criteria for assessment of hosts are outlined below in Table 16. These criteria should be reviewed and prioritised to provide a framework for decision-making on the suitability of hosts. The criteria's suitability for the Minimum and Extended hosting options, outlined above, are added below.

Table 16 Suggested Criteria for assessment of potential hosts

Criteria	Minimum option	Extended option	How measured	Notes
Capacity to host	Y		Already hosting other content Technical skills: In-house staff Platform	Hosts would need to evidence that they have the basic capability to deliver what is required. This would include the size of their existing service, number of staff and the functionalities of the technical platform.
Sustainability	Y	Y	Secured funding A policy or commitment to maintain the service	A host should have both the ongoing funding (or at least a commitment to raise funds) as well as an interest in supporting an archive of the library or the ongoing development of the library.
Cost	Y	Y	Total cost for transfer and archiving or Total cost for transfer, hosting and ongoing content management (if	

[32] https://assets.publishing.service.gov.uk/media/57a08a3ae5274a27b20004bf/AFCAP-GEN0-96-Knowledge-Final-Report.pdf

[33]https://assets.publishing.service.gov.uk/media/58d9115940f0b606e3000028/Crul_2016_KMWorkshopReportCaledon_ReCAP_KMN2106A_161214.pdf

			applicable)	
Sector/Knowledge	Y		Thematic fit (e.g. proportion of existing content covering scope).	
Promotion and visibility (including active information dissemination beyond their own reports and outputs) Likelihood of content being used	Y	Y	Popularity (Search engine ranking) Recognition in sector	A repository or service which was well known in the sector would ensure that the content of the RAL would continue to be found and used Active promotion would be necessary if the Extended option was chosen, to disseminate new content added to the RAL
Political considerations and appropriateness	Y		Reputation Support for policies, in particular open access	ReCAP may wish to work with organisations with whom they have successfully partnered with previously, or feel are mandated to play a similar role in the sector. Politically there may be sensitivities working with organisations who have a global remit but who are not seen as having a positive presence in the sector, or those who favour some regions or sectors over others.
Expression of interest/Willingness to host	Y		Mission or aims of the organisation or service	
Geographic location of host organisation / Southern based? (programme aims to build capacity in Africa and Asia).	Y		Location of staff or governing body	It would be preferable if ownership of the dataset was held by organisations which are governed by or provided for those who would most benefit from the contents.
Governance	Y		A consortium model or agreement in place Relevant policies	We might prefer a collaborative approach rather than single institution lead This should include the degree to which the hosts are accountable to the users of the library Policies or commitments should be in place to ensure the library is maintained as open access and is standards compliant.

7.2.5 Preliminary assessment of example potential hosts

The three hosts were contacted to determine:

- Any concerns, constraints or perspectives research publishers face
- Broad scope and remit, including external content policies and existing similar relationships,
- Technological or capacity issues.

The hosts were chosen to loosely represent different types of potential hosts and selected on the basis they were already known to ReCAP or the WAP team and had potential capacity to host the RAL.

- Eldis, as a global development sector service with similar research/practitioner audience, with some sustainability questions
- CSIR, a sector-specific partner of ReCAP, based in global south and hosting research metadata from multiple projects
- Internet archive, as a global service without any sector or geographic connections, but as an example of a generic internet archiving service.

Table 17 Preliminary assessment of example potential hosts

Host	Issues identified
Eldis (IDS, OpenDocs) Eldis provides a user-friendly front-end metadata portal for global development resources beyond those publications produced by IDS. The OpenDocs DSpace-based repository holds independent collections and would be the system likely to host the RAL.	Strategically, there is not a close fit with IDS priorities or partnerships. The technological requirements (to manage the bulk uploads) and editorial capacity (for quality checking and checking of licenses etc.) would not be cost effective for Eldis at this current time. Eldis would be interested in showcasing the content for the RAL should a suitable host be found, by providing editorial and links to the documents. They would not automatically harvest existing metadata.
Internet Archive A global repository specifically designed to archive cultural digital assets.	A scheduled seminar to discuss these issues and potential costs did not happen. At the time of writing, we are waiting for the next opportunity.
CSIR https://researchspace.csir.co.za/dspace/	We have submitted follow up questions to CSIR and at the time of writing are awaiting a response.

8 Next steps: Recommendations for Phase 2

It is a great strength of the RAL collection that it contains good, well-structured metadata that is relatively straightforward to map to existing and well recognised standards. There is generally sound categorisation and high quality content. There is a recognised gap in robust policies around the content and management of the collection, which we recommend should be addressed in Phase 2. In our exploration of cross-repository harvesting options and transfer methods, we have recognised how interlinked all these components are on the capacity and existing set up of any potential future hosts. We therefore recommend concentrating on establishing contact with shortlisted future hosts in Phase 2 to calculate potential costs, and transfer options to enable ReCAP PMU to make an informed decision about where to host the RAL at the end of the programme.

Our recommendations for possible tasks for Phase 2 are listed as follows, broken down by tasks undertaken in Phase 1.

8.1 Metadata standards and adaptations to make the RAL compliant

The next steps would be to improve the metadata structure and content to make it a more attractive offer. This process is relatively straightforward and would include the following.

8.1.1 Meet basic standards

As discussed earlier, the most common and easy to use format is Dublin Core (DC). It is a widely-used and understood format within the scholarly and development communication space. **It is therefore recommended that, as a minimum, RAL metadata is brought in line with Dublin Core Metadata Element Set Version 1.1.**[34]

We recommend that the RAL metadata structure should map to the qualified version - QDC - rather than the more basic DC. QDC offers more flexibility for the third party systems that may harvest metadata from RAL. For example, in simple DC, you could have various occurrences of a date field (such as publication date or accession date) but due to limitations in the data structure and labelling, it is difficult to distinguish these. For the current metadata where we do not have equivalent fields in Dublin Core, fields should be assigned to a local RAL schema.

To meet DC standards for the full text metadata, there is a need for complete description of the file, including a filename, description, the mime type of the file and its size. Filename and description have to be

[34] http://dublincore.org/documents/dces/

manually managed but the other fields can be automatically managed by the implementing platform software system.

We would suggest that the following additional fields, in Table 18, are added and populated.

Table 18 Suggested fields to add and populate to meet basic standards

Field	Description
Identifier	URI/Handle (not the filename, in case the name changes)
Licence	2 fields - name of licence, URL to licence
Provenance	Usually one-off statement added to every record
Access restrictions	If required where research publishers require embargoes or just want the item and/or full text available to a restricted group of users.

8.1.2 Increase metadata scope

With the RAL there could be a need to improve the metadata with fields that are not present in the current Dublin Core Metadata Element Set Version 1.1.

For those fields that are specific to ReCAP and its partners (shown in Annex 4), good practice dictates that these are not added to the standard DC schema but added as a local metadata schema. Most repository software platforms allow for local metadata schema to be created in an easy way so this should be specified now, and implemented later.

We would suggest that the following additional fields, in Table 19, are added and populated.

Table 19 Suggested fields to add and populate to increase metadata scope

Field	Description
Relationships	Determined by project code to link documents which relate to the same project
Language	3 fields – for language of the full text document – for language of the metadata – for unique identifier (or other language versions)

8.1.3 Improve presentation

Although slightly out of the scope of this document, we would also recommend some changes to the RAL web pages (if relevant to the transfer option chosen), to create better presentation, search engine optimisation and findability. We feel it is important that the output of the web page that displays the metadata data is capable of generating the required meta tags that are useful for search engines, and other tools and services such as Zotero[35], Connotea[36] and SIMILE Piggy Bank[37], to correctly pick out item metadata fields. These meta tags are the "Highwire Press tags" which Google Scholar recommends[38], therefore the final system for RAL should be able to do mapping of the metadata from that used in the cataloguing.

The restructuring of the “front end” should also include functionality to display one item per page, in addition to the list pages of filters and searches. This single page would contain all relevant metadata associated with the knowledge item and provide a unique URL for the metadata.

[35] http://www.zotero.org/
[36] http://www.connotea.org/
[37] http://simile.mit.edu/wiki/Piggy_Bank
[38] https://scholar.google.com/intl/en/scholar/inclusion.html

8.1.4 Editorial tidying of content

In terms of content of the metadata, we would recommend the following as a minimum:

- **Removing overlaps and inconsistencies** e.g. rationalising Document Types, adding in missing metadata fields, such as Document Format, de-duplicating Publishers, tidying keywords
- **Augmenting the metadata** e.g. adding the project information, mapping country information to the ISO country standard, and adding language attributes
- **Adding to the content**, by including the ReCAP published Journal Articles.

Table 20 outlines possible tasks for Phase 2 with an indication of cost and benefit.

Table 20 Possible tasks for Phase 2

Content	Issue	Resolution	Benefit	WAP Cost	ReCAP Cost
Document type	Overlap of terms *Conference presentation* *Conference/workshop presentation*	ReCAP PMU to confirm that these do overlap	Fewer choices for end user. Better consistency of labelling	½ day	Minimal
Document Format	5 knowledge items missing Document Format	Add document format to RAL	Completeness of metadata		None
Countries	Full country names	Look up Codes for the Representation of Names of Countries (ISO 3166-1993 (E))	Consistent metadata. More linkable and transferable		1 hour - Add Country code field and apply mapping to data/import new country code data
Language	Not recorded	Add language information Define that we are recording the language of the output as well the metadata	More able to support cross repository harvesting in languages other than English. Improved description of the knowledge item.	½ day	1 hour - Add language fields and apply mapping to data/import new language data
Keywords	A number of inconsistencies which reduce the searchability and effectiveness of this free text field	Tidy up inconsistencies as necessary	Keyword search is more effective. Faceted search would be possible and more accurate Search Engine Optimised	1 day	None - 1 hour Depending on process, this could include importing new keyword data
Publishers	Input errors and	Correct input errors	Better consistency	½	1 hour - could

	inconsistencies Joint publishers recorded as free text Possible duplication and overlap of publishers	and de-duplicate Tidy up inconsistencies and separate publishers, with a semi-colon, to allow for separation at a later stage ReCAP PMU to confirm that these do overlap and rectify as appropriate	of labelling Allows separation of publishers for searching, filtering and faceting.	day	include importing of corrected publisher data
ReCAP Published Journal Articles	Data is not ready for integration with the rest of the RAL dataset	ReCAP PMU to add relevant themes/sub-themes and keywords Other data tidying, including linking to Open Access licence for each item	Includes valuable knowledge outputs in the RAL.	½ day	2 hours - could include importing of new knowledge items
Project information	The project identifier is not recorded in the metadata	Using the project id in the full text file name, a new field will be added and populated with the relevant project identifier	Creates more linking possibilities in the content.	½ day	1 hour - Add Project ID field and import project identifier data
Authors	Currently the authors are added as free text and therefore do not have persistent identifiers	Look up ORCID iD[39] for authors and apply to data	Creates more linking possibilities in the content, as well as persistent identifiers.	3 days	1 hour - Add repeating ORCID field and import author identifier data

Impact on these recommendations for the ReCAP PMU

Early in Phase 2, depending on what tasks are decided upon from the Phase 1 report recommendations, WAP will create a functional specification for the extension of the ReCAP SharePoint platform. The WAP team will work closely with the ReCAP webmaster during this process.

[39] https://orcid.org/

Recommended Actions:

- Add metadata fields to comply with Dublin Core metadata standards; Identifier, Licence, Provenance and Access restrictions and populate
- Add metadata fields to increase interoperability; Project, Language and populate and map country information to the ISO country standard
- Remove overlaps and inconsistencies e.g. rationalising Document Types, adding in missing metadata fields, such as Document Format, de-duplicating Publishers, tidying keywords (NB This would not include any work on searching for and recording persistent identifiers for authors unless required)
- Check list of ReCAP published Journal Articles to ensure these are included in the RAL.

8.2 Identify and define relevant policies

Due to the nature of the way that the RAL has grown, it is apparent that the opportunity to develop clear policies for the RAL collection has not presented itself. The ReCAP PMU have recognised the absence of policies and have therefore included this as an integral part of the *Enhancing the sustainability of the ReCAP Rural Access Library* project.

Table 21 presents a recommended set of policies that should be created by ReCAP. These policies should provide clarity and understanding of the content, its focus, determine how the content is collected, curated and preserved as well as how it may be used, for human users, computer access and other third-party services.

These policies will provide OpenDOAR compliance and help embed best practice in digital library and digital repository provision. Policies prescribed by OpenDOAR are flagged below.

We also feel that these policies will provide the requisite alignment with funder policies on open access as provided through their mandates.

The remaining policies have been excluded as they are considered beyond the need and scope of the project or primarily governing internal use.

We have listed our recommendations for policies to be drafted in Table 21. All the policies listed in this table are **considered essential for the RAL** as they help align the library to good practice.

The development of robust policies will make the RAL more attractive to future hosts, in terms of the quality, provenance and management of the collection.

Table 21 Recommended policies for the RAL

#	Policy	Rationale
1	Collection Policy	It is important for ReCAP to take a strategic approach to its content acquisition, management, preservation and distribution of the content. To aid understanding of what type of content should be collected and disseminated in order to support the organisation's operational needs, meet its objectives and prioritise editorial work. For potential hosts and funders, it provides clarity on the scope and role of the collection in the sector. It offers ReCAP the best way of informing its stakeholders of its commitment to and capability in good knowledge management. Considered essential for any library that collects, curates and distributes any material of a scholarly nature.
2	Metadata Policy	While it is always assumed that metadata is Open Access by default, it is important that a clear statement to emphasize the openness of the metadata is provided.
3	Data Policy	This policy provides clear guidance on how users can use or reuse the full text and/or datasets. This provides legitimacy to the body of evidence presented as well as clearly stating how the research can be used by other researchers, practitioners and stakeholders, which could lead to more uptake and use. [OpenDOAR]

4	Submission Policy	As long as the RAL will continue to source and host content in the platform, this policy is necessary. [OpenDOAR] May not be needed in a future closed (archived) version of the RAL.
5	Content Policy	This policy is essential for providing credibility of the library to third party services that may need to use the content or evaluate the library against others. [OpenDOAR]
6	Preservation Policy	The policy provides confidence that the content will be looked after well with its integrity guaranteed for as long as the library exists. [OpenDOAR]
7	Deposit Licence Agreement	Although the content is mostly produced by ReCAP grantees, there could be situations where the authors might have made agreements with other (journal) publishers or that the content contains sensitive or possible patentable elements. This document, agreed by the RAL and author, demonstrates that the RAL provides mechanisms for embargoes and redactions of the content.
8	Copyright and liability statement	As content is mostly produced by ReCAP grantees this is a useful policy that mitigates certain liabilities. Each ReCAP report has a standard disclaimer, which should be made apparent and public as part of the RAL policies as well as the document itself.
9	RAL decommissioning guidelines	The temporary existence of the RAL under its current hosting (whilst ReCAP is an ongoing project) means that it is probable that the ownership and platform itself will change. It is therefore important that a set of guidelines of how this change is going to be managed without compromising the integrity of the content in it and its discovery, as well as its accessibility is put in place. For any future hosts, it would give an understanding of expectations for long-term sustainability of the library.
10	Open access policy	This would articulate the commitment to building open knowledge for the sector and Open Access principles, to help promote the library to future collaborators. It would explain the relationship with R4D and potentially help secure future funding for the library.
11	System policy	This would contain information about how the repository is managed in terms availability, resilience, backup, disaster recovery and general level of support. [OpenDOAR]

8.2.1 Next steps

Working with ReCAP, the WAP Team will provide templates and guide the drafting of these policies by taking into account the available documentation and obtaining good practice in this area. They should be relatively short documents, written in a clearly accessible language without the use of unnecessary jargon.

We also believe that these policies will form a basis for negotiation with any potential future hosts.

8.2.2 Access to policies

All the policies that will be created should be made available online so that all types of users and ReCAP stakeholders have access to them. Policies, in the absence of OAI-PMH capability, should be accessed online via individual or grouped pages on the RAL website, with a link to the RAL policies placed in a persistent place, such as the footer.

Note that some of the information that OpenDOAR presents regarding a particular service is collected automatically by OpenDOAR using the prescribed OAI-PMH capability of the library. Interoperability of the library via this protocol is discussed in the Cross Repository harvesting and Transfer Option sections.

8.2.3 Recommended Actions:

- WAP to provide draft templates for, guide the writing of and deposit online, the following new policies;
    - Collection Policy
    - Metadata Policy
    - Data Policy
    - Submission Policy
    - Content Policy
    - Preservation Policy

    - Deposit Licence Agreement
    - Copyright and liability statement (A full policy may not be necessary in this case as each Recap report has a standard disclaimer. This statement should be displayed on the website as a catch all).
    - RAL decommissioning guidelines
    - Open access policy
    - System policy
- WAP to assist in the writing and online depositing of the policies
- WAP to assist in the process of validation with OpenDOAR

8.3 Exploring options for cross-repository harvesting and drafting a specification

As stated above, we feel that the approach and specification is entirely dependent on the technical configuration of the host that ReCAP chooses to take on the repository and any chosen method of transfer.

We therefore make no recommendations at the point specifically related to cross-repository harvesting in isolation, although all recommendations in the routemap section take into account the cross-repository harvesting component.

8.4 Routemap for technical transition of management, publishing and sharing of RAL library content

As most modern systems should provide at least a basic file-based import functionality we believe Option A is likely to be the most viable, cost effective and with the lowest risk. The best use of available resources would be to work on the data export, re-formatting any output files generated. This should happen once the host is chosen, and efforts focussed on supporting them in the import process.

At this stage however, ReCAP should not rule out any of the options - the most appropriate solution will depend on the capabilities and preferences of the future hosts and how negotiations with them proceed.

8.4.1 Next steps

Option C1 would potentially offer a more straightforward (and lower cost) route to the RAL being available via an OAI-PMH interface. This has the advantage of offering harvesting for potential hosts immediately, reducing their costs and making hosting more attractive.

In the short-term ReCAP should try the SharePoint Web Part implementation (the 2nd potential solution) to determine the feasibility and likely costs. This would also offer a solution in which the transfer could be designed now and implemented at project closure.

8.4.2 Recommended Actions:

- Engage in small pilot study to test viability of using SharePoint Web Part

8.5 Identifying and assessing hosts

The biggest question to answer in Phase 2 is the choice of future host for the RAL. So much depends on the chosen hosts' technical capabilities, capacity and willingness to take on the dataset. The choice of host will suggest the technical transfer route depending on their existing systems and capabilities. This will also dictate how and when the RAL becomes OAI-PMH compatible, enabling cross-repository harvesting and broader content sharing.

More work should be undertaken in Phase 2, in close consultation with ReCAP, to refine the list of possible hosts and applicable criteria, as well as perform an assessment of each potential host.

A short list of candidate hosts should be drawn up and agreed and individual contact made with the relevant organisations. This would identify potential collaborators and develop detailed costs for each solution.

Recommended Actions:

- WAP to refine potential host long list with further input from ReCAP
- WAP to refine selection criteria for assessment of hosts, based on essential and added value priorities
- WAP to undertake systematic analysis of host long list to draw up short list for ReCAP approval.
- ReCAP to develop a TOR for potential new hosts with support from WAP
- WAP to approach shortlisted candidates to ascertain challenges, preferences and costs and to report on findings to enable ReCAP to make a decision on roadmap and future host.

Annex 1 Dublin Core Metadata

The Dublin Core Initiative

The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description. The name "Dublin" is due to its origin at a 1995 invitational workshop in Dublin, Ohio; "core" because its elements are broad and generic, usable for describing a wide range of resources. The first standard metadata format is defined through the Dublin Core Metadata Element Set Version 1.1[40] and takes the form dc:namespace.

Since 1998, when these fifteen elements entered into a standardisation track, notions of best practice in the Semantic Web have evolved to include the assignment of formal domains and ranges in addition to definitions in natural language and these are now defined as DCTERMS[41] and takes the form dcterms:namespace.

Although going forward and especially if the application will be used in Semantic Web context, use of DCITERMS is encouraged but no advice has been given to stop the use of the DC based definition. The notion of having DC and DCTERMS has caused some confusion as to which of the two implementers should use and so far the advice is as described in Box 1.

Dublin Core metadata

Table A1 DC elements and their qualifiers.

element	qualifier	scope note
contributor		A person, organization, or service responsible for the content of the resource. Catch-all for unspecified contributors.
contributor	advisor	Use primarily for thesis advisor.
contributor	author	Author(s) of the work
contributor	editor	
contributor	illustrator	
contributor	other	
coverage	spatial	Spatial characteristics of content.
coverage	temporal	Temporal characteristics of content.
creator		May be used as an alternative to "contributor.author"
date		Use qualified form if possible.
date	accessioned	This is usually used by the implementing system.
date	available	Date or date range item became available to the public.
date	copyright	Date of copyright.
date	created	Date of creation or manufacture of intellectual content if different from date.issued.
date	issued	Date of publication or distribution.

[40] http://purl.org/dc/elements/1.1/
[41] http://purl.org/dc/elements/1.1/

date	submitted	Recommend for theses/dissertations.
description	abstract	Abstract or summary.
description	provenance	The history of custody of the item since its creation, including any changes successive custodians made to it. This is used by the implementing system.
description	sponsorship	Information about sponsoring agencies, individuals, or contractual arrangements for the item.
description	statementof responsibility	To preserve statement of responsibility from MARC records.
description	tableofcontents	A table of contents for a given item.
description	uri	Uniform Resource Identifier pointing to description of this item.
description		Catch-all for any description not defined by qualifiers.
format	extent	Size or duration.
format	medium	Physical medium.
format	mimetype	Registered MIME type identifiers.
format		Catch-all for any format information not defined by qualifiers.
identifier		Catch-all for unambiguous identifiers not defined by qualified form; use identifier.other for a known identifier common to a local collection instead of unqualified form.
identifier	citation	Human-readable, standard bibliographic citation of this item
identifier	govdoc	A government document number
identifier	isbn	International Standard Book Number
identifier	issn	International Standard Serial Number
identifier	sici	Serial Item and Contribution Identifier
identifier	doi	The Digital Object Identifier
identifier	ismn	International Standard Music Number
identifier	other	A known identifier type common to a local collection.
identifier	uri	Uniform Resource Identifier
language		Catch-all for non-ISO forms of the language of the item, accommodating harvested values.
language	iso	Current ISO standard for language of intellectual content, including country codes (e.g. "en_US").
publisher		Entity responsible for publication, distribution, or imprint.
relation		Catch-all for references to other related items.
relation	isformatof	References additional physical form.
relation	ispartof	References physically or logically containing item.
relation	haspart	References physically or logically contained item.
relation	isversionof	References earlier version.

relation	hasversion	References later version.
relation	isbasedon	References source.
relation	isreferencedby	Pointed to by referenced resource.
relation	requires	Referenced resource is required to support function, delivery, or coherence of item.
relation	replaces	References preceding item.
relation	isreplacedby	References succeeding item.
relation	uri	References Uniform Resource Identifier for related item
relation	ispartofseries	Series name and number within that series, if available.
rights		Terms governing use and reproduction.
rights	uri	References terms governing use and reproduction.
source		Do not use; only for harvested metadata.
source	uri	Do not use; only for harvested metadata.
subject	classification	Catch-all for value from local classification system. Global classification systems will receive specific qualifier
subject	ddc	Dewey Decimal Classification Number
subject	lcc	Library of Congress Classification Number
subject	lcsh	Library of Congress Subject Headings
subject	mesh	MEdical Subject Headings
subject	other	Local controlled vocabulary; global vocabularies will receive specific qualifier.
subject		Uncontrolled index term.
title	alternative	Varying (or substitute) form of title proper appearing in item, e.g. abbreviation or translation
title		Title statement/title proper.
type		Nature or genre of content.

This table is an adaptation from
https://wiki.duraspace.org/display/DSDOC6x/Metadata+and+Bitstream+Format+Registries

Annex 2 OpenAire DC metadata example[42]

A complete example record

```
<record>
    <header>
        <identifier>oai:repository.example.com:2003292</identifier>
        <datestamp>2012-11-30T13:40:28Z</datestamp>
        <setSpec>openaire</setSpec>
    </header>
    <metadata>
        <oai_dc:dc
         xmlns:dc="http://purl.org/dc/elements/1.1/"
         xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
                             http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
            <dc:title>Studies of Unicorn Behaviour</dc:title>
            <dc:identifier>http://repository.example.org/2003292</dc:identifier>
            <dc:creator>Jane, Doe</dc:creator>
            <dc:creator>John, Doe</dc:creator>
            <dc:description>
                Lorem ipsum dolor...
            </dc:description>
            <dc:subject>info:eu-repo/classification/ddc/590</dc:subject>
            <dc:subject>Unicorns</dc:subject>
            <dc:relation>info:eu-repo/grantAgreement/EC/FP7/1234556789/EU//UNICORN</dc:relation>
            <dc:relation>info:eu-repo/semantics/altIdentifier/eissn/1234-5678</dc:relation>
            <dc:relation>info:eu-repo/semantics/altIdentifier/pmid/123456789</dc:relation>
            <dc:relation>info:eu-repo/semantics/altIdentifier/doi/10.1000/182</dc:relation>
            <dc:relation>info:eu-repo/semantics/reference/doi/10.1234/789.1</dc:relation>
            <dc:relation>info:eu-repo/semantics/dataset/doi/10.1234/789.1</dc:relation>
            <dc:rights>info:eu-repo/semantics/openAccess</dc:rights>
            <dc:rights>http://creativecommons.org/licenses/by-sa/2.0/uk/</dc:rights>
            <dc:source>Journal Of Unicorn Research</dc:source>
            <dc:publisher>Unicorn Press</dc:publisher>
            <dc:date>2013</dc:date>
            <dc:type>info:eu-repo/semantics/article</dc:type>
        </oai_dc:dc>
    </metadata>
</record>
```

[42] https://guidelines.openaire.eu/en/latest/literature/use_of_oai_dc.html

Annex 3 RAL metadata mapping to Dublin Core (DC) metadata

Table A2 used in this mapping contains a subset of metadata fields from the table in Annex 1 as the metadata fields that are deemed unhelpful or irrelevant in this project have been removed for clarity and simplicity.

Table A2 Metadata mapping

DC element	DC qualifier	RAL Metadata
contributor	author	AUTHOR
contributor	editor	
contributor	other	
coverage	spatial	FOCUS COUNTRIES
date	accessioned[1]	
date	available	
date	copyright	
date	issued	YEAR
description	abstract	
description	provenance[1]	
description	sponsorship	AUTHOR'S INSTITUTION
format	extent	
format	medium	
format	mimetype	
identifier	citation	
identifier	isbn	
identifier	issn	
identifier	sici	
identifier	doi	
identifier	uri	
language	iso	
publisher		PUBLISHER
relation		
relation	isformatof	
relation	ispartof	
relation	haspart	
relation	isversionof	
relation	hasversion	

relation	isbasedon	
relation	isreferencedby	
relation	requires	
relation	replaces	
relation	isreplacedby	
relation	uri	
relation	ispartofseries	
rights		
rights	uri	
subject	classification	
subject	ddc	
subject	lcc	
subject	lcsh	
subject	mesh	
subject	other	
subject		THEME, SUB THEME and KEYWORDS
title	alternative	
title		TITLE and SUBTITLE
type		DOCUMENT TYPE

[1] Used by the implementing system.

Annex 4 RAL metadata not mapped to Dublin Core

The following could not be matched to the Dublin Core metadata standard and they, therefore, could be candidates to add to a local RAL metadata schema.

They will not have DC qualifier if they are added to the local schema.

Table A3 RAL metadata not mapped to Dublin Core

Field	Local Schema	Comment
CITY	Yes But if not added to local schema, see comments	This is the city of publication. It can be mapped to Dublin Core and put in the publisher metadata field together with the Publisher, see https://www.loc.gov/standards/mods/dcsimple-mods.html.
NAME (filename)		This is part of the full text metadata and is usually bundled with the full text metadata fields.
TITLE HYPERLINK		This is not required as the record will have a persistent identifier that can be used on the display interface.
DOCUMENT FORMAT	Yes But if not added to local schema, see comments	The information can be inferred from the file itself by the implementing software system and usually referred to as Mime type. See Annex 5 for examples.
EVENT	Yes But if not use as discussed in the comments	This can be actually coded in the type field as this a Type of output, if not required to be included in the local schema.
(METADATA) AUTHOR	Yes	
CREATED DATE		This is maintained [as dc.date.accessioned] by the system and also in the provenance metadata field. Usually this is maintained by the implementing software system.
MODIFIED DATE		This is maintained by the system and is also part of the provenance metadata field and my also include the user who performed the action, when stored in the provenance metadata field. It is important that this field is always appended to so as to have a complete audit trail of the modification made to the metadata record. Usually this is maintained by the implementing software system.

Annex 5 Mime type examples

Mime type	Short Description	Description	Extensions
application/octet-stream	Unknown	Unknown data format	
text/plain	License	Item-specific license agreed upon to submission	
application/marc	MARC	Machine-Readable Cataloguing records	
application/mathematica	Mathematica	Mathematica Notebook	ma
application/msword	Microsoft Word	Microsoft Word	doc
application/pdf	Adobe PDF	Adobe Portable Document Format	pdf
application/postscript	Postscript	Postscript Files	ai, eps, ps
application/sgml	SGML	SGML application (RFC 1874)	sgm, sgml
application/vnd.ms-excel	Microsoft Excel	Microsoft Excel	xls
application/vnd.ms-powerpoint	Microsoft Powerpoint	Microsoft Powerpoint	ppt
application/vnd.ms-project	Microsoft Project	Microsoft Project	mpd, mpp, mpx
application/vnd.visio	Microsoft Visio	Microsoft Visio	vsd
application/wordperfect5.1	WordPerfect	WordPerfect 5.1 document	wpd
application/x-dvi	TeX dvi	TeX dvi format	dvi
application/x-filemaker	FMP3	Filemaker Pro	fm
application/x-latex	LateX	LaTeX document	latex
application/x-photoshop	Photoshop	Photoshop	pdd, psd
application/x-tex	TeX	Tex/LateX document	tex
audio/basic	audio/basic	Basic Audio	au, snd
audio/x-aiff	AIFF	Audio Interchange File Format	aif, aifc, aiff
audio/x-mpeg	MPEG Audio	MPEG Audio	abs, mpa, mpega
audio/x-pn-realaudio	RealAudio	RealAudio file	ra, ram
audio/x-wav	WAV	Broadcase Wave Format	wav
image/gif	GIF	Graphics Interchange Format	gif
image/jpeg	JPEG	Joint Photographic Experts Group/JPEG File Interchange Format (JFIF)	jpeg, jpg
image/png	image/png	Portable Network Graphics	png
image/tiff	TIFF	Tag Image File Format	tif, tiff
image/x-ms-bmp	BMP	Microsoft Windows bitmap	bmp

image/x-photo-cd	Photo CD	Kodak Photo CD image	pcd
text/css	CSS	Cascading Style Sheets	css
text/html	HTML	Hypertext Markup Language	htm, html
text/plain	Text	Plain Text	asc, txt
text/richtext	RTF	Rich Text Format	rtf
text/xml	XML	Extensible Markup Language	xml
video/mpeg	MPEG	Moving Picture Experts Group	mpe, mpeg, mpg
video/quicktime	Video Quicktime	Video Quicktime	mov, qt

Table is an adaption from https://wiki.duraspace.org/display/DSDOC6x/Metadata+and+Bitstream+Format+Registries

Annex 6 Content Review

Document Types

Document Type	No of occurrences
Activities report/general overview	149
Conference presentation	11
Conference/workshop presentation	270
Evaluation report	31
Policy Brief	24
Progress/Milestone report	262
Research paper	63
Standard, guideline or manual	75
Training resource	177
Workshop report	81

Document format

Document Format	No of occurrences
PDF	1125
Word	13
Unknown	5

Spatial coverage

Regional coverage

Region	No of occurrences
Africa	204
Asia	60
Global	107

Country coverage

Country	No of occurrences
Afghanistan*	3
Bangladesh*	21

Botswana	3
Cambodia	76
Central African Republic	1
China	1
Congo, Dem. Rep. of the*	10
Ethiopia*	101
Ghana*	58
Kenya*	43
Lao People's Democ. Rep.	71
Liberia*	10
Madagascar	3
Malawi*	70
Mozambique*	65
Myanmar (ex-Burma)*	7
Nepal*	21
Nigeria	8
None	8
Pakistan*	1
Sierra Leone*	13
South Africa	5
South Sudan*	4
Tanzania*	150
Uganda*	15
Vietnam	99
Zambia*	27
Zimbabwe	3

* denotes current ReCAP focus countries as per January 2018

Thematic coverage

Top level theme

Top Level Theme	No of occurrences
Rural roads and infrastructure research	488
Transport knowledge management, education and	327

dissemination	
Transport research uptake and policy	87
Transport services research	241

Sub theme

Sub theme	No of occurrences
Agriculture and access	56
Asset Management and Road Condition	332
Capacity building	270
Children, older persons and marginalised groups	21
Climate Resilience and Environment	128
Construction and Upgrading	137
Design	229
Disability, access and universal design	9
Economics, Value for Money, CBA	71
Education and access	60
Footpaths, trails and trail bridges	2
Gender and mobility issues	55
Health services and access	52
Information and communications technology (ICT) and mobile phones	24
Integration of transport (including waterways)	2
Intermediate means of transport (IMTs)	27
Knowledge management practices	55
Maintenance and rehabilitation	199
Measuring access and isolation and policy issues	41
Monitoring and evaluation	80
Motorcycle taxis and three-wheelers	38
Needs assessment	29
None	48
Planning, including Integrated Rural Accessibility Planning	66
Research methodology	81
Road materials and aggregates	122
Road safety and security	77

Rural transport advocacy	80
Rural transport services (rural taxis, buses and minibuses)	83
Seals and surfaces for low volume roads	243
Structures	57
Trial/demonstration sites	103

Keyword Analysis

Keywords that have occurred in 50 or greater knowledge items

Keyword	No of occurrences
RURAL	332
ROAD	194
TRANSPORT	184
MAINTENANCE	121
DESIGN	102
RESEARCH	100
VIETNAM	98
MANAGEMENT	96
LVR	87
TRAINING	79
SURFACING	77
LAO PDR	69
ROADS	63
PAVING	62
CAMBODIA	60
SERVICES	58
ACCESS	55
LOW VOLUME ROADS	52

Publisher

Publisher	No of occurrences
AFCAP	1

Amend	1
ASANRA	1
CAPSA	2
Cardno	4
Cardno Emergin Markets UK	1
Cardno Emerging Markets UK	1
Cardno IT Transport	1
CNCTP	1
Crown Agenta	7
Crown Agents	628
DART	1
DFID	2
Eco-Logica	1
Elsevier	1
ERA, Ethiopia	10
Ethiopian Roads Authority	7
GFDRR	1
Government of Ghana	1
Government of the Republic of the Union of Myanmar	1
gTKP	8
Helvetas	2
ICE	1
ICTA 2015	1
ICTA 2015Pa	1
ILO Cambodia	2
Indian Roads Congress	2
InfraAfrica	1
Intech Associates	1
IRF	2
iTRARR	21
LGED	1
MHSW, Tanzania	1
Ministry of Federal Affairs and Local Development	1
Ministry of Public Works & Transport, Lao PDR	1

Ministry of Rural Development, Cambodia	2
Ministry of Transport, Infrastructure, Housing & Urban Development, Kenya	1
Ministry of Works, Transport & Communication	4
MRB South Sudan	1
MRD Cambodia & PIARC	27
MTPW Malawi	1
MTPW, Malawi	1
PIARC	2
PIARC - World Road Association	1
PIARC & ILO	1
RDA Zambia	1
ReCAP for DFID	313
RFB, Tanzania	1
Riders for Health	1
Road Fund Mozambique	1
Roughton	1
Roughton International	1
SAICE	1
SSATP	1
Steering Committee	1
Sustainable Mobility for All	1
T2 Conference	2
T2 Conference 2017	18
Taylor & Francis Group	1
Transport Publishing House, Hanoi	1
Transport Publishing House, Hanoi	1
TRB	6
TRL	1
TRL and ILO	1
TRL Ltd	11
TRL Ltd & ILO	3
TRL Ltd & ILO Cambodia	4
UNCDR	1

UNESCAP	4
UNOPS	1
WHO	2
World Bank Group	1
World Conference on Transport Research	1
World Transport Policy & Practice Journal	1

Publication Year

Publication Year	No of occurrences
2000	6
2001	1
2002	8
2003	4
2004	33
2005	47
2006	63
2007	15
2008	50
2009	81
2010	85
2011	37
2012	43
2013	60
2014	218
2015	54
2016	127
2017	175
2018	36

Language

Language	ISO_639-1 code[43]	No of occurrences
	CHECK[1]	1
English	EN	1055
Spanish	ES	9
French	FR	22
Khmer (Cambodian)	KM	11
Lao	LO	17
Portuguese	PT	10
Vietnamese	VI	15

[1] There is one phantom record, that contains some metadata, but the full text is currently inaccessible.

Bilingual knowledge items

Language	ISO_639-1 code	No of occurrences
English and French	EN;FR	1
English and Khmer	EN;KM	1
English and Lao	EN;LO	1

[43] International Organisation for Standardization code: https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

Annex 7 Comparative Taxonomies

Document Type

R4D Doc types	Default DSpace doc types
Book	Animation
Book Chapter	Article
Briefing	Book
Case Study	Book Chapter
Conference Paper	Dataset
Country Report	Learning Object
Dataset	Image
Discussion Paper	Image, 3-D
Evaluation Report	Map
Journal Article	Musical Score
Journal Issue	Plan or blueprint
Lessons Learned	Preprint
Literature Review	Presentation
Manual	Recording, acoustical
Protocol	Recording, musical
Research Paper	Recording, oral
Systematic Review	Software
Technical Report	Technical Report
Thematic Summary	Video
Tool kit	Working Paper
Training Materials	
Working Paper	

Theme

SSATP themes

Taken from https://www.ssatp.org

Top Level theme	Sub theme

Integration and Connectivity	
	Regional Coordination
	Corridors Management
	Practical Solutions
	Corridor Performance Monitoring
	Policies and Strategies
Transport Management	
	Road Management and Financing
	Railways
	Urban Mobility and Accessibility
	Rural Transport and Mobility
Cross-Cutting Issues	
	Road Safety
	Governance and Integrity
	Climate Change
	Gender and Inclusion
	HIV & AIDS
	Learning
	Tools
	Toolkits & Methodologies

Annex 8 RAL entry on OpenDOAR[44]

Repository Information

Repository Name	Rural Access Library [English]
Repository Type	Disciplinary
Description	This site provides access to the reports and research outputs of the program. ReCAP looks into rural road infrastructure and transport services in Africa and Asia. The interface is available in English
Repository URL	http://www.research4cap.org/SitePages/Rural%20access%20library.aspx
Software Name	Other (HTML)
Languages:	English
Content Types	Journal Articles Conference and Workshop Papers Books, Chapters and Sections
Subjects	Arts and Humanities General > Geography and Regional Studies Social Sciences General > Management and Planning

Organisation

Organisation Name	Research for Community Access Partnership (ReCAP) [English]
Organisation URL	http://www.research4cap.org/
Country	United Kingdom

Metadata Policy

None

Data Policy

None

Content Policy

None

[44] http://v2.sherpa.ac.uk/id/repository/3693

Submission Policy

None

Preservation Policy

None

System Policy

None

Annex 9 OAI-PMH, OAI-ORE and metadata formats

OAI-PMH definitions

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [45]is a simple, common set of rules for computers to exchange content. It is primarily used to share information about the contents of Open Access repositories, the metadata. This metadata describes the full text documents and other articles held in the repository.

There is a relatively low technical barrier to implementing and using OAI-PMH which has contributed to its popularity and most systems which are designed to hold and share data now implement OAI-PMH.

In technical terms, OAI-PMH is an HTTP-based protocol that defines methods and structures for sharing, publishing and archiving metadata from repositories over the Internet which supports and enhances repository interoperability. Data Providers are repositories that expose structured metadata via OAI-PMH and Service Providers then make OAI-PMH requests to harvest that metadata[46].

It is important to note that OAI-PMH does not describe the exact format in which the metadata is held, although there are many common metadata formats in use with OAI-PMH. These are listed below.

Accessing data via OAI-PMH

A data provider will publish an 'OAI Base URL' usually in the form http[s]://data-provider/oai/request? and such a URL is used in machine-to-machine communications between data consumers and data harvesters. When a harvesting request is made using the OAI Base URL plus an appropriate Verb, the data provider returns metadata as an XML (eXtensible Mark-up Language) formatted response.

OAI-PMH comprises of a set of six verbs or services that are invoked within HTTP, and appended to OAI base URLs in order to access different repository contents.

These are:

1. **Identify:** fetches descriptive information about the data-provider itself
2. **ListMetadataFormats**: returns a list of available metadata formats supported by a data provider
3. **ListIdentifiers:** lists structure and record identifiers
4. **ListSets:** retrieves the set structure of the repository
5. **ListRecords:** gets a list of complete metadata of the content held in the repository or part of the repository
6. **GetRecord:** retrieves individual metadata of a record held in the repository

An OAI data provider can prevent any performance impact caused by harvesting by forcing a harvester to receive data in time-separated chunks. If the data provider receives a request for a lot of data, it can send part of the data with a resumption token. The harvester can then return later with the resumption token and continue.

[47]Data providers are encouraged to use OAI 2.0, a Java implementation of an OAI-PMH data provider interface developed by Lyncode that uses XOAI, an OAI-PMH Java Library. This implementation therefore allows for projects like OpenAIRE[48], and Driver[49] that have specific metadata requirements (to the published content through the OAI-PMH interface). The OAI-PMH protocol, on the other hand, does not.

[45] https://www.openarchives.org/pmh/
[46] https://wiki.duraspace.org/display/DSDOC6x/OAI+2.0+Server
[47] https://wiki.duraspace.org/display/DSDOC6x/OAI+2.0+Server
[48] https://www.openaire.eu/
[49] http://www.driver-support.eu/

Beyond metadata

For harvesting of metadata, OAI-PMH can be used. However, to share and access the full text documents and assets, the data provider must also implement Open Archives Initiative Object Reuse and Exchange (OAI-ORE)[50] in addition to OAI-PMH.

This extended implementation therefore allows migration of all content from one repository to another.

More and more services (for example the CORE[51]) are opting to also be able to harvest both metadata and full-text documents. Even the services that use ResourceSync[52] also utilises the OAI-ORE implementation of the data provider.

Metadata formats

OAI-PMH implementations expose metadata in a number of formats, many based on library formats. The most basic metadata format is the simple Dublin Core (oai_dc) format, as well as the extended version Qualified Dublin Core (qdc).

The table below lists some of the metadata formats, for example a DSpace based OAI-PMH implementation provides.

ketd_dc	Namespace: http://naca.central.cranfield.ac.uk/ethos-oai/2.0/ Schema: http://naca.central.cranfield.ac.uk/ethos-oai/2.0/uketd_dc.xsd
qdc	Namespace: http://purl.org/dc/terms/ Schema: http://dublincore.org/schemas/xmls/qdc/2006/01/06/dcterms.xsd
didl	Namespace: urn:mpeg:mpeg21:2002:02-DIDL-NS Schema: http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-21_schema_files/did/didl.xsd
mods	Namespace: http://www.loc.gov/mods/v3 Schema: http://www.loc.gov/standards/mods/v3/mods-3-1.xsd
ore	Namespace: http://www.w3.org/2005/Atom Schema: http://tweety.lanl.gov/public/schemas/2008-06/atom-tron.sch
mets	Namespace: http://www.loc.gov/METS/ Schema: http://www.loc.gov/standards/mets/mets.xsd
oai_dc	Namespace: http://www.openarchives.org/OAI/2.0/oai_dc/ Schema: http://www.openarchives.org/OAI/2.0/oai_dc.xsd
rdf	Namespace: http://www.openarchives.org/OAI/2.0/rdf/ Schem: http://www.openarchives.org/OAI/2.0/rdf.xsd
marc	Namespace: http://www.loc.gov/MARC21/slim Schema: http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd

[50] https://www.openarchives.org/ore/
[51] https://core.ac.uk/
[52] http://www.openarchives.org/rs

xoai	Namespace: http://www.lyncode.com/xoai Schema: http://www.lyncode.com/schemas/xoai.xsd
dim	Namespace: http://www.dspace.org/xmlns/dspace/dim Schema: http://www.dspace.org/schema/dim.xsd
etdms	Namespace: http://www.ndltd.org/standards/metadata/etdms/1.0/ Schema: http://www.ndltd.org/standards/metadata/etdms/1.0/etdms.xsd

Annex 10 OAI-PMH Connector on SharePoint

These options may provide a mechanism for querying SharePoint content using the OAI-PMH protocol, and take two different approaches.

1. SharePoint 2010 to Primo data connector

Built as a thesis submission at the National University of Ireland, this implementation has a main use case to be able to pull simple format metadata of records in a SharePoint server that are then passed through to a Primo discovery service platform[53]. The implementation example is on SharePoint 2010 and connects to a Primo data connector. The aim of this implementation was to create a bridge between the two formats that the systems use which are ATOM XML on the SharePoint side and OAI-PMH client on Primo's side. Please note that primo is fully OAI-PMH compliant. The solution thus provided suitably named ATOM2OAI-PMH was built using PHP, XML, XSLT, CURL, and SharePoint REST API using oData.

While the implementation satisfied the use case it was created for, it has to be noted that it is not a fully-fledged implementation of an OAI-PMH server. ATOM2OAI-PMH only implements two OAI-PMH verbs namely **Identify** and **ListRecords** and exposes the records as simple Dublin Core (oai_dc). The implementation also lacks a Resumption Token capability. Such an implementation would be adequate as a short-term or data-exposing solution for the RAL.

Downloaded from: https://developers.exlibrisgroup.com/blog/SharePoint-2010-to-Primo

The gateway was developed by Cillian Joy[54] at the National University of Ireland, Galway[55].

In theory it could be extended to deliver a fully OAI-PMH compliant server implementation:

It has however to noted that OpenDOAR and a majority of known and reputable repository aggregators, compliance validators and repository ranking services require a fully functioning OAI-PMH server.

2. SharePoint Web Part implementation

Online searching has also uncovered a SharePoint webpart which theoretically provides functionality to search for and export content via the OAI-PMH protocol.

Downloaded from https://archive.codeplex.com/?p=OpenArchivesWP

[53] http://primodb.org
[54] https://www.linkedin.com/in/cillianjoy/
[55] https://library.nuigalway.ie/digitalscholarship/about/staff/

Annex 11 Potential Interim systems

Summary

This review aims to give a summary of the pros and cons of systems which may be suitable for hosting RAL content in the interim.

We have included viable systems which:

- Are OAI-PMH compliant, as a minimum requirement;
- Have functionality to import content from CSV or SharePoint;
- Could be hosted by RAL or as online software as a service (SaaS);
- Have a good adoption rate in the sector (as a proxy for sufficient functionality and support networks)
- Have a relatively low-cost

This review is undertaken in the context of the system likely to be an interim solution as the ultimate solution will be based on the preferred hosts existing set up. ReCAP is not looking to develop its own longer-term repository.

System types

1. Repository and digital library types - specialised content management system that are used as repository and digital library management system
2. Journal publishing - systems which support the workflow associated with submission, collation and review of submitted academic articles. These are too specific in focus and not a good fit for the RAL
3. Research information systems - systems primarily focussed on managing content within a research institution,
4. Aggregation and discovery services - systems or portals which harvest and categorise metadata, usually providing a search engine
5. Web CMS - systems which are designed to generate rich user-facing content over data management functionality, but which can be extended to incorporate the management of metadata and documents. SharePoint is an example of this type.
6. Enterprise content management (ECM) and document/record management systems (DAMS) - aimed at the management of documents, audio, visual and other media, usually for large organisations.

Candidate platforms

System (Type)	Description	Pros	Cons
DSpace[56]	A turnkey open-source digital asset management system from Duraspace[57] and is considered by far the most popular and tested repository solution available.	It provides most of the attributes required for a digital repository in terms of metadata; interoperability; Embargoes, Versioning and Preservation; and many more. Supports AIP imports/exports	While it is a versatile platform the total cost of ownership can be quite high without good enough skill sets.

[56] https://duraspace.org/dspace/
[57] https://duraspace.org/

		Can be locally hosted or can be used as SaaS.	
Drupal[58] (Web CMS/framework)	Drupal is an open-source web-based content management system which provides a framework for software development beyond front-end website publishing implementations.	Flexible and extensible with good support for interoperability Offers a mature Biblio module with integrations for OAI-PMH	Complex to set up and support Latest version of Drupal only has limited support
WordPress[59] (Web CMS)	WordPress is a complex and advanced open-source web content management system used to power a number of information systems and websites. However, its primary function is a web publishing tool, rather than a repository system.	Easy to set up, and maintain. Many developers familiar with the system and therefore support is readily available and cheap.	There are a few plugins, which are used to extend the core WordPress functionality, that offer a repository-like system e.g. Tainacan[60] and Document Management System[61]. These systems however are not used widely or mature enough to recommend.
EPrints[62]	EPrints is a turnkey free and open-source software package originally developed by researchers at the University of Southampton School of Electronics and Computer Science in 2000 (making it the oldest of the platforms in this report). It was designed specifically for archiving research papers, theses and teaching materials, though it can accept any content.	It provides most of the attributes required for a digital repository in terms of metadata; interoperability; Embargoes, Versioning and Preservation; and many more. Capable of using a controlled vocabulary and authority lists Can be locally hosted or can be used as SaaS.	It is UK centric; only has support for simple DC Does not support AIP imports/exports
Fedora[63]	Fedora is a digital asset management (DAMS) architecture upon which institutional repositories, digital archives, and digital library systems might be built.	Flexible, modular, with native linked data support and all the other attributes listed for DSpace and EPrints.	It has very high total cost of ownership as it is only a framework that has to be built on. There are however turnkey implementations that leverage it, such as Hydra (now called Samvera[64]) and Islandora[65].

[58] https://www.drupal.org/
[59] https://wordpress.org/
[60] https://wordpress.org/plugins/tainacan/
[61] https://wordpress.org/plugins/dms/
[62] http://www.eprints.org/uk/
[63] https://duraspace.org/fedora/
[64] ttp://samvera.org/

Alfresco[66]	Alfresco is an open source document and knowledge management system with a commercial organisation behind it, providing add-on services. Can be used with Web CMS front-end to provide complete services	Has an OAI-PMH add-on module	Relatively expensive to host and implement Not widely used in the academic sector
CKAN[67]	CKAN (Comprehensive Knowledge Archive Network) has been primarily developed a turn-key solution for multi-provider open data repositories. Used for many national (government) data initiatives	A modern interface with excellent tools for presenting datasets in multiple formats	Limited support for document metadata, may be too specialist for ReCAP needs.
Invenio[68]	An open-source framework for large-scale digital repositories developed and managed by CERN - the European Organization for Nuclear Research.	It provides most of the attributes required for a digital repository in terms of metadata; interoperability; Embargoes, Versioning and Preservation; and many more. Can be locally hosted	EU centric and hasn't got a large installation base. A lack of a large user and developer community may add to the total cost of ownership.
Zenodo[69]	Provides a free repository space, primarily for EU research and datasets. Aimed at researchers who do not have an existing institutional or thematic repository they can deposit their publications and data in. Zenodo is based on Invenio. Run by CERN - European Organization for Nuclear Research	It provides most of the attributes required for a digital repository in terms of metadata; interoperability; Embargoes, Versioning and Preservation; and many more. Includes DOI and is available via OpenAIRE It is used as PaaS offering and is free to use. Users can create their own collections on it.	May only provide services to EU based research projects and programmes

[65] https://islandora.ca/
[66] https://www.alfresco.com
[67] https://github.com/ckan/ckan
[68] https://invenio-software.org/
[69] http://zenodo.org

Annex 12 Long list of potential hosts

Types of repositories

- Research - institutional (university) or departmental
- Research - Multi-institution repository
- Research - Cross-Institutional
- e-Journal/Publication
- e-Theses
- Database/A&I Index
- Research Data/Open and Linked Data
- Learning and Teaching objects
- Demonstration
- Web Observatory

Provider/host types

- Academic
- Private sector
- Funder
- State (government)
- Network/partnership

Sectoral archives/portals

Rural transport, development sector or social services related

Name and URL	Host	Scope	Notes
Sub-Saharan Africa Transport Policy Program (SSATP) https://www.ssatp.org	World Bank funded	180+ publications hosted at https://www.ssatp.org/en/publications	*Not a viable option in my view* Caroline Visser (email 16/7/18)
PIARC knowledgebase https://www.piarc.org/en/knowledge-base	World Road Association (PIARC)	800 publications in virtual library	Recommended by Caroline Visser (email 16/7/18)
Eldis http://www.eldis.org https://opendocs.ids.ac.uk/opendocs	Institute of Development Studies (IDS) http://www.ids.ac.uk	Eldis (metadata index of 40K+ social science and development research outputs) and Open Docs open access repository, which hosts content from IDS and other research institutes	Good reputation in the development sector. Primary focus is on research papers so may not be a close enough fit.
The International Forum for Rural Transport and Development (IFRTD) http://www.ifrtd.org	Global network of individuals and organisations working together towards improved access, mobility and economic opportunity for	Strong correspondence with thematic and geographic focus. Signposts to external resources and doesn't have its own online library.	Could be considered as a potential vehicle for governance, asking one of the members to become the

	poor communities in developing		host.
global Transport Knowledge Partnership (gTKP) https://www.gtkp.com	Funded by World Bank Group, run by IRF Geneva (see below)	Website was down or no longer online at the time of review, although Internet Archive searches reveal a close match in scope..	Hosts contacted to determine the sustainability of the platform
International Road Federation https://www.irf.global/			Global website says: 2016: Ended unification efforts with IRF Geneva and established IRF Global
IRF Geneva https://www.irfnet.ch/			
Practical Action https://practicalaction.org	A long standing UK-based but now global organisation with a mission to tackle poverty.	Provides a technical information service, including transport coverage as part of Practical Answers, good geographic match. They have also been working on building knowledge repositories and have good capacity in this area.	Strong candidate for further discussion

Generic archiving services

Service	Description	Approximate costs
Preservica https://preservica.com/digital-archive-software/products-pricing	Private sector option Ingest pack for SharePoint OAI-PMH and CMIS content query API	*£9.5k p/a*
Internet Archive http://archive.org https://archive-it.org	Runs Wayback Machine which archives general Internet content. However, it can be used to harvest metadata and provides an API to the content. Describes itself as a web based application, enabling institutions to create collections of archived web content. An annual Archive-It subscription includes hosting, access, and storage. Tends to be for "cultural" content.	Do not publish prices online.
Figshare https://figshare.com/		
Zenodo	Provides a free repository, primarily for EU research and datasets. Aimed at researchers	Free[70]

[70] This needs to be verified in discussion

http://zenodo.org	who do not have an existing institutional or thematic repository they can deposit their publications and data in. Includes DOI. Also available via OpenAIRE Run by CERN - European Organization for Nuclear Research.	

Funder repository and archiving services

Repository / Service	Description	Notes
World Bank Open Knowledge Repository https://openknowledge.worldbank.org		Just for World Bank publications: https://openknowledge.worldbank.org/pages/about-en
Wellcome Trust https://wellcomeopenresearch.org		Only for Wellcome funded research.
DfID R4D https://www.gov.uk/dfid-research-outputs	It is part of the ReCAP contract with DFID that output from the programme are deposited in R4D so good thematic match. Full text items may be deposited in or signposted from the R4D site.	No longer seems to be available as a dataset[71] and there is a very limited front-end for searching.

National and/or government, partnership

Organisation	Description	Notes
Southern African Transport Conference (SATC) www.satc.org.za	South African Universities Uses University of Pretoria (dspace) repository	
New Partnership for Africa's Development (NEPAD) www.nepad.org/	The technical body of the African Union Knowledge base of approx 500 documents, as well as its own publications.	
Ethiopia		

[71] https://ckan.integration.publishing.service.gov.uk/dataset/research-for-development-gateway

Ethiopian Roads Authority www.era.gov.et/ Addis Ababa University - Institutional Repository http://etd.aau.edu.et		
Kenya Kenya Rural Roads Authority www.kerra.go.ke		Website was not working at the time of auditing
Tanzania Tanzania Transportation Technology Transfer (TanT2) Centre http://www.tant2centre.or.tz/	2200 records Clearing-house for transportation information	

Related services

Services which compliment research

Service	Notes
Research Gate https://researchgate.net	
ResearchFish https://www.researchfish.net/thehub	Collection system for (mainly UK) research impact, uploaded by researchers.