**UK Research Data Discovery Service (RDDS) - Technical Use Case**

**Research outputs discovery using the Organisation Profile Document (OPD)**

****

**What is an OPD?**

An [OPD](http://opd.data.ac.uk/) is a machine readable rdf file embedded in an institution’s website containing the organisation's full name, homepage, logo, dataset location and contact information for open access datasets. The OPD is an essential component of the autodiscovery process in the National Equipment Portal, equipment.data, a service endorsed as strategically significant by HEFCE.

**What is needed to enable OAI-PMH end point to be discovered on an OPD?**

Research output datasets use <http://purl.org/openorg/theme/ResearchOutputs> to describe the availability of a research outputs dataset and the most common format <http://purl.org/dc/terms/conformsTo> is a OAI-PMH endpoint with the uri <http://www.openarchives.org/OAI/openarchivesprotocol.html>. Below is an example of the script required to publish the availability of a “Research outputs” end point on an OPD.

In addition the Research Outputs entry on the OPD lists the setSpec(s) of appropriate set(s) available via the end point. E.g. a mixed repository might list the setSpec for Datasets.

<http://eprints.soton.ac.uk/cgi/oai2>

 dcterms:conformsTo <http://www.openarchives.org/OAI/openarchivesprotocol.html> ;

 dcterms:license <http://www.nationalarchives.gov.uk/doc/open-government-licence/> ;

 dcterms:subject <http://purl.org/openorg/theme/ResearchOutputs> ;

 oo:contact <mailto:eprints@soton.ac.uk> ;

 oo:corrections <mailto:eprints@soton.ac.uk> ;

 oo:organization <http://id.southampton.ac.uk/> .

Using the [OPD checker service](http://opd.data.ac.uk/checker) to display the human readable version the above script will provide the information below.

****

The autodiscovery programming developed and used by the equipment.data service harvests the entire contents of the OPD therefore offering the opportunity to filter by dataset. This will enable a full list of published endpoints to be filtered and compiled on a separate webpage/sub-domain. The research outputs end points are currently available at <http://opd.data.ac.uk/data/researchoutputs>.

Using this route will provide two possible delivery options:

**Option 1.**

A third party service uses the list of OAI-PMH end point locations to interrogate all openly published outputs to provide a dataset which can be filtered as required e.g. to a standardised set (via the setSpec provided) or using the research outputs metadata profile (Ball A. et al) (annex a) and displayed by the selected web service provider.

It is likely subject to the data requirements of the RDDS web service provider this option may be achieved by either of two dataset query options:

1a. <http://opd.data.ac.uk/data/researchoutputs> will provide filtered OPD data consisting of organisation ID and end point for all institutions publishing OPDs (including setSpec(s) where appropriate). This will be available for interrogation by downloading the aggregated datasets from the site.

1b. The full OPD dataset discovered by <http://opd.data.ac.uk> is interrogated by the web service provider allowing the capture of items such as institutional logo, official name, organisation ID and the OIA-PMH end point (including setSpec(s) were appropriate). The web service provider can achieve this through a choice of two methods:

1. Downloading full OPD dataset spreadsheet in .csv or JSON and running a query of the data in a separate location
2. Running SPARQL queries against the complete RDF graph available at opd.data.ac.uk.

**Option 2.**

This option is essentially option 1 with the addition of end point interrogation and metadata aggregation. This will include a establishing a new sub domain, <http://researchoutputs.data.ac.uk>, on the website data.ac.uk which includes a data store of the discovered research outputs. Alternatively an option such as an ePrints community hosting service could be set up to store all discovered outputs.

The new subdomain or service could offer an either/or option of raw data discovered by the discovery service following interrogation of end points or a dataset filtered using the research outputs metadata profile/schema.

As is some cases, repositories may be used to store many collections, only part of which could be datasets, so the data profile should support the selection of a setSpec defined sub set of the outputs.

 An api (with preferred data format – suggest OAI-PMH) would then be established which could be queried and displayed by the UK RDDS web service.

 **Options for metadata provision**

<http://opd.data.ac.uk/data/researchoutputs> (OAI-PMH list) - queried by 3rd party system

<http://researchoutputs.data.ac.uk> - Data harvested using discovery service - Validated against a research outputs metadata profile.

The autodiscovery programme uses the metadata profile to filter the open access records content.

To appropriately achieve all metadata discovery needs a set(s) with “standardised” metadata schema(s) e.g. Dublin Core (OAI\_DC), MODS or DataCite.

**1a**

**1b**

**OPTION 1**

**opd.data.ac.uk/data/research outputs**

List of published OAI-PMH endpoints

**opd.dat.ac.uk**

Stored information discovered from institutional OPDs

Institutional OPDs

SPARQL query of rdf

**OPTION 2**

Institutional OAI-PMH endpoints interrogated by autodiscovery programme

Institutional repository end-points

New sub-domain established as repository for research outputs metadata

**Researchoutputs.data.ac.uk** (repository of discovered metadata)

Agreed research outputs metadata profile as validator

UK RDDS Web service interrogates API (OAIM-PMH end point)

Option: Discovery program filters data using metadata profile

UK RDDS Web Service front end

Process for metadata provision:

1. “Research outputs” end point data is filtered from current discovered institutional OPDs and listed on <http://opd.data.ac.uk/data/researchoutputs>
2. data.ac.uk’s autodiscovery programme is used to interrogate end points either:
3. aggregating all open access research outputs
4. Program interrogates end points and filters data validating against metadata profile, aggregating specified metadata only.
5. Discovered and aggregated metadata is either hosted on the sub-domain <http://researchoutputs.data.ac.uk> or other open access web domain.
6. An OAI-PMH end point will provide access allowing services to interrogate or download the datasets.

Alternative approach (SPARQL Query)

1. RDDS Web service provider runs a SPARQL query on the RDF graph available from <http://opd.data.ac.uk> extracting required fields including list of OAI-PMH end points (and defined setSpec(s) if appropriate)
2. The end points are interrogated and validated (against an agreed metadata profile) with the results linked to the institutional repository.

**Benefits of solution**

* Uses existing institutional and national infrastructure
* Interrogation of end points and data provision is automated
* Mechanisms for data integrity already developed i.e. System generates a warning email to institutions if ingest fails i.e. data location moves without OPD being updated.
* Lightweight only requiring publishing format complying with OAI-PMH.
* Other data services can access the metadata e.g. Intelligent Brokerage Tool or Gateway to Research.

**Considerations**

1. **All publishing institutions require an OPD**
2. **Institutions have OAI-PMH end point and defined setSpec(s) if needed**
3. **Is there a need to establish a standard OAI set name and defined specification? Who would lead this development?**
4. **Metadata profile (core fields) is adopted by all data collections – Is current consensus sufficient? Using OAI – DC is the widely used current standard**
5. **Is there a need for a CASRAI Working Group if harmonising range of profiles?**

**Author: Adrian J Cox, University of Southampton**

**Date: January 16**