DataHarvest Workshop Supplementary Documentation for Research Partners

This document provides details of the information provided in the DataHarvest Workshop.

DataHarvest is an investment of the Grains Research and Development Corporation (GRDC). It is led by Curtin University through the Centre for Crop and Disease Management (CCDM).

GRDC code: CUR2401-001BGX

Table of contents

1. GRDC Data Management Policy 2
2. The FAIR Principles 2
3. Your Obligations as Research Partners 3
4. Data Management Plan 3
5. Metadata for GRDC RD&E Datasets 3
6. Metadata Example 4
7. Data Licences and Access Rights 7
8. Data Repositories 7
9. Uploading to a Data Repository (Zenodo) 8
10. The GRDC Data Catalogue 10
11. Registration of Project Metadata to GRDC Data Catalogue 10
12. Accessing data on the GRDC Data Catalogue 11
13. Where to access support – Support Services Centre 12

1. GRDC Data Management Policy

GRDC manages RD&E data in accordance with the FAIR principles. Data should be findable, accessible, interoperable and reusable to maximise its value. For more information about GRDC Data Management plans, visit the GRDC website.

2. The FAIR Principles

FAIR stands for Findable, Accessible, Interoperable, and Reusable. Adopting these principles makes data easier to find and share. When data is easy to access, researchers can quickly build on each other's work, leading to faster breakthroughs and innovative farming practices. Seamless data sharing means better collaboration and more effective use of resources, driving progress in the Australian grains industry.



Findable: Metadata and data can easily be found by humans and computers, using rich descriptive metadata and assigning globally unique and persistent identifiers

Accessible: Once the data has been found, it should be clear how it can be accessed.
This may include authentication or authorisation

Interoperable: Data and metadata are described in a standardised language and format. When data is interoperable, it will integrate well with other data sources

Reusable: Data that can be reused will prevent duplication of research, saving time and resources. Clear data use licences describe how data can be reused.

Resources:

- 'The Fair Principles Explained' Youtube Video
- FAIR Principles ARDC Resource

3. Your Obligations as Research Partners

As a GRDC Research Partner, you will be required to:

- submit at Data Management Plan (DMP)
- upload the data and accompanying metadata produced during the project to a GRDC Approved Repository
- Register quality metadata to the GRDC Data Catalogue

If you are unsure of your obligations, contact your GRDC Investment Manager.

4. Data Management Plan

The Data Management Plan (DMP) clearly outlines how the research partner will manage the RD&E data that is generated during the lifetime of the project. It also states who from the research organisations will be responsible for complying with GRDC's data management requirements.

The GRDC requires all research partners to complete and submit a DMP when they are engaged in GRDC project/investment.

The approved GRDC Data Management Plan template can be accessed and downloaded from the GRDC website. This template includes help notes that will assist in correctly filling out the DMP.

The information that is stored within the DMP will fall under one of three categories, 'Investment Details', 'Project Data Management' and 'Data Storage'. These categories are split by worksheets within the xlsx file. Not all data is worth keeping. Consider the potential re-use purposes, legal compliances and potential value. The data that is to be kept should be decided in consultation with your GRDC Investment Manager.

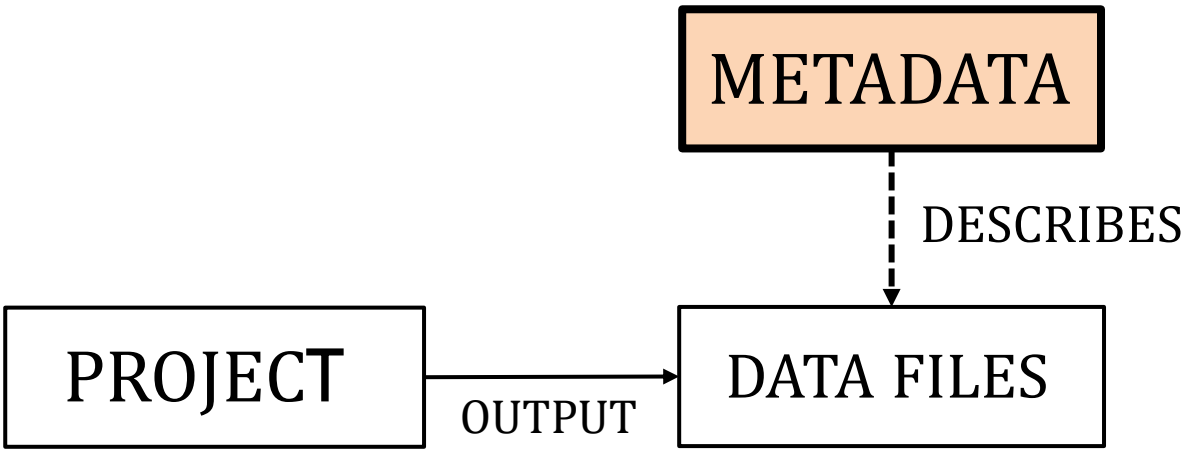
Resources:

- Data Management Plan Template
- GRDC RD&E Data Capture and Storage – Guidelines For Research Partners
- Data Management Plan Background Training Video
- Data Management Plan Tutorial Training Video
- General information about Data Management Plans

5. Metadata for GRDC RD&E Datasets

Metadata is the information that documents and provides context for a dataset. The metadata should describe:

- **What** the data is,
- **Where** it was collected,
- **When** it was collected,
- **Who** was involved
- **How** it was generated.



Effective metadata also describes the contextual knowledge needed to understand and use the data. The more information that is provided, the more findable and useable your data will be. The goal is to ensure that anyone can easily understand the content and purpose of the data through clear, concise, and descriptive metadata.

Resources:

- GRDC RD&E Data Capture And Storage – Guidelines For Research Partners
- General Metadata Information
- AGROVOC Multilingual Thesuarus

6. Metadata Example

An example of clear, good quality metadata is shown below. This is a real example from the GRDC Data Catalogue.

Name	**Optimising Plant Establishment 2020 Hart Wheat Experiment**
Type	Data set
Contributing Organisation	University of Adelaide **Role:** Owner **ROR:** **University of Adelaide ROR**
GRDC Contract Title & Code	Optimising plant establishment, density and spacings to maximise crop yield and profit in the southern and western regions GRDC project number UOA1803-009RTX
DOI	10.25909/24784323.v1
Data Collection dates	**Dataset coverage starts** 31st December 2019 **Dataset coverage ends** 30th December 2020
Spatial Coverage	Other
Dataset Description	Data from the 2020 crop establishment trial at Hart, SA. The trial compared the establishment of wheat at different target seeding rates, under a

	conventional seeder and under a precision seeder. This experiment was run under the UOA1803-009RTX GRDC research project led by the University of Adelaide, and the experiment and data collection were managed locally by the Hart Field-Site Group. **Files:**2020 Hart Wheat.xlsx **Experiment setup:** Experiment managed locally by the Hart Field-Site Group. The trial was organised as a split plot design, with 2 whole plot treatments (conventional seeder vs precision seeder), and 2 subplot treatments for the seeding rate (90 and 180 target plants/m2). The design was arranged spatially in a single column (3 bays long) with 12 rows, 3 replicates, 12 plots in total. **Variables recorded:** stubble (%), stubble (t/ha), establishment count (plants/m2), seedling depth (mm), interplant distance (cm), NDVI, flowering biomass dry matter (t/ha), harvest biomass wet weight (g), harvest biomass dry weight (g), harvest index, yield (t/ha), 1000 grain weight (g), test weight (%), screenings (%), protein (%) **Data collection and assessment:** The paddock was seeded on 14/05/2020, with a pre-seed stubble assessment on 13/05/2020. Plant emergence counts were made on 30/06/2020, 08/07/2020 and 24/07/2020 to assess establishment. Seedling depth in mm was also measured on 24/07/2020 for all plots, with five measurements made per plot. Interplant distances were recorded at each plot on 24/07/2020. NDVI was measured on five separate occasions: 30/06/2020, 24/07/2020, 03/08/2020, 12/08/2020 and 17/08/2020. Biomass cuts were made on 24/09/2020 at flowering and the dry matter weight recorded, and harvest maturity cuts were made on 02/11/2020 to assess weight and harvest index. The crop was harvested on 26/11/2020, with the yield (t/ha), 1000 grain weight and protein content recorded. Additionally, some photos were taken of the trial and stubble during the crop growth, and these are available embedded in the xlsx file.
Contributing People	**Glenn McDonald** **Role: Associated** **ORCID iD: 0000-0002-7120-9042** **Jacky Desbiolles** Role: Associated ORCID iD: 0000-0002-5921-261X **Sarah Noack** Role: Associated **Stefan Schmitt** Role: Associated **Claire Pickles** Role: Associated

	Genevieve Clarke Role: Associated **Rebekah Allen** Role: Associated **Ashley Amourgis** Role: Associated **Jon Midwood** Role: Associated **David Minkey** Role: Associated
Keywords	• plant establishment • precision seeding • germination • wheat
Data licence	CC-BY https://creativecommons.org/licenses/by/4.0/
Access rights	Rights - Open

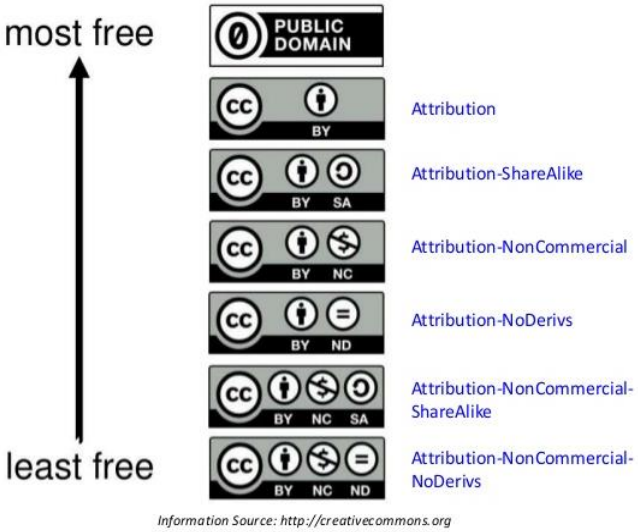
There is a lot of information contained in the metadata, and we have a very good idea about what the dataset will look like, without ever seeing the dataset. Creating metadata records will take time and effort, so plan accordingly.

Below are some other real records available in the GRDC Data Catalogue:

- Optimising Plant Establishment 2018 Skipton Canola Experiment
- 2022 & 2023 Hyper Yielding Crops Paddock Data
- Improved Invertebrate Pest Management in the Northern Grains region – experimental datasets
- Impact of timing and techniques of soil amelioration on crop establishment and profitability in Sandplain soils

7. Data Licences and Access Rights

A data use licence is a legal agreement that stipulates the terms under which data can be accessed, used, and shared. It outlines permissions, restrictions, and obligations for users, including aspects such as whether the data can be modified or redistributed, and how the original creators should be credited.



A common data use licence is the Creative Commons licence. There are six different Creative Commons license types, ranging from least permissive to most permissive.

Access rights detail who may access the dataset, and under what conditions. Access rights can be:

1. *Open* - CC BY
2. *Conditional* – conditions may include embargo period, fee, data available only at a particular physical location, available to Aust citizens only
3. *Restricted* - third party must request access to the RD&E Data from the owners (contains personal data, commercially/ industrially sensitive data etc.)

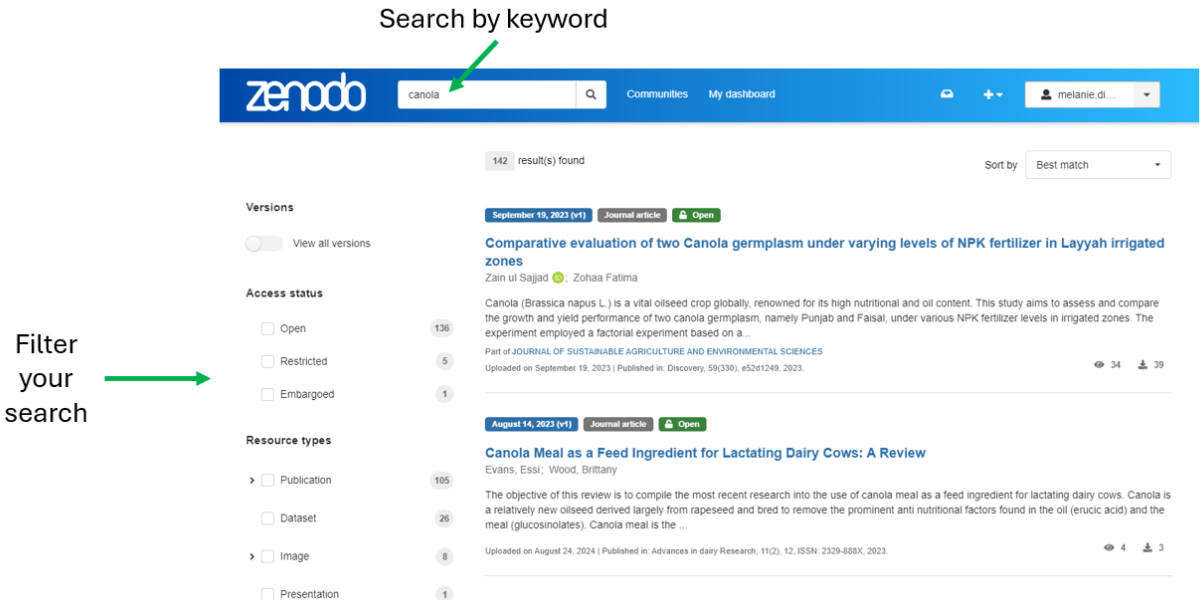
Refer to your GRDC research contract to find out what access rights will apply to your data.

Resources:

- GRDC RD&E Data Capture And Storage – Guidelines For Research Partners
- General information about Creative Commons Licences

8. Data Repositories

A data repository is a centralised location designed for storing information. It functions by **collecting, storing**, and **managing** data, ensuring its preservation and facilitating sharing. The data within a repository is standardised and organised, making it easily searchable and accessible.



Using a data repository is crucial for ensuring the long-term preservation and accessibility of your data. Repositories provide a secure and organised space for storing, managing, and sharing data, making it easily searchable and usable. They support collaboration by enabling data sharing with colleagues and the wider community, increasing the impact of your work.

OneDrive is not a data repository. OneDrive is primarily a cloud storage service and not a traditional data repository. While it allows for storing, sharing, and syncing files, it lacks key features typical of a data repository, such as the ability to create detailed metadata, assign persistent identifiers like DOIs for citation, and ensure long-term data preservation.

Resources:

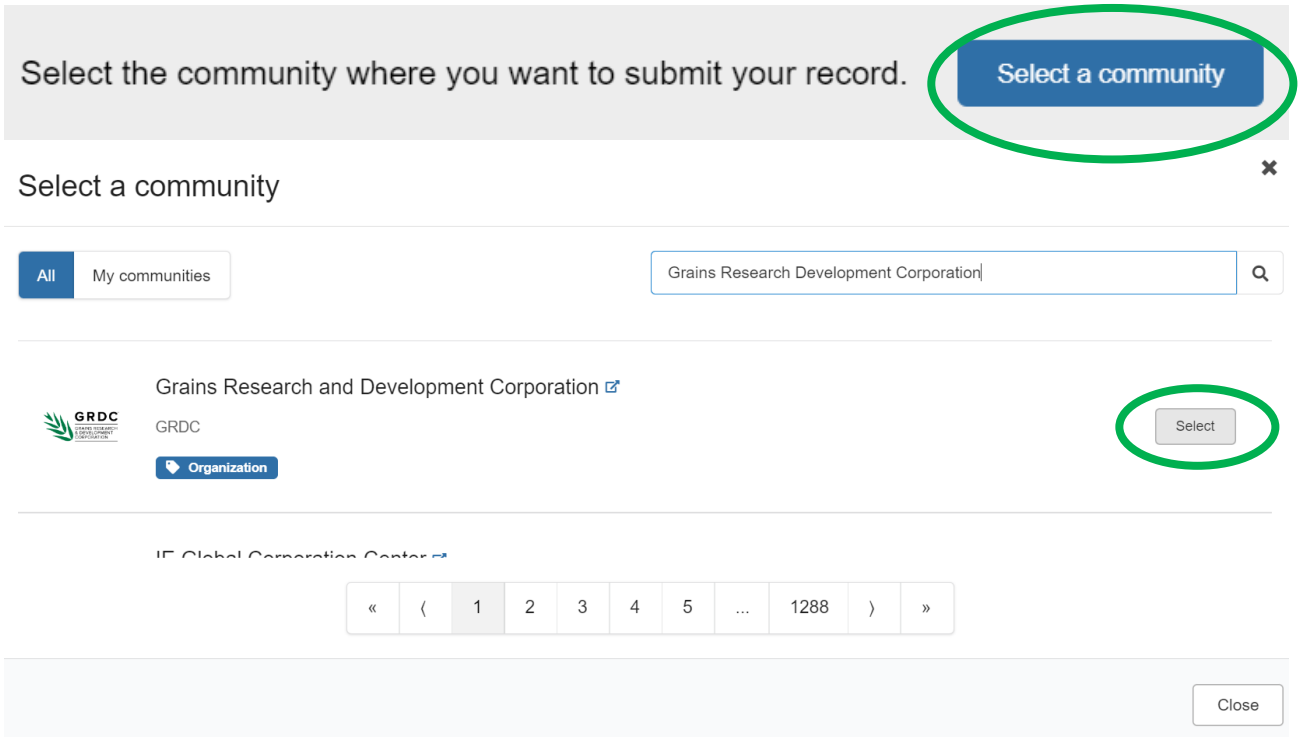
- GRDC RD&E Data Capture And Storage – Guidelines For Research Partners

9. Uploading to a Data Repository (Zenodo)

Zenodo is a free (and GRDC Approved) repository that would suit the needs of a research organisation that does not their own, preferred repository.

A clear, step by step guideline has been created to assist research partners for the upload of data and metadata creation in Zenodo, and is available on the GRDC website. Following this guide will ensure that a high-quality metadata record is successfully harvested into the GRDC Data Catalogue.

When uploading your data to Zenodo, you **must** select **"GRDC Community".** If you forget to select this community, your data will not be automatically harvested onto the GRDC Data Catalogue and therefore, you will not fulfill your obligations as a research partner.



After linking your new Zenodo upload to the GRDC community, ensuring that it can be automatically harvested onto the GRDC Data Catalogue, simply fill out the required metadata information. Resources are available to provide assistance in uploading project data to Zenodo.

Resources available:

- GRDC RD&E Data Capture And Storage – Guidelines For Research Partners
- GRDC Guideline for Research Partners for data upload and metadata creation in Zenodo
- Zenodo Upload Tutorial Videos

10. The GRDC Data Catalogue

The GRDC Data Catalogue is a public platform that is used to consolidate, reference and ensure discoverability of Australian Grains Industry datasets. The GRDC Data Catalogue was launched in April 2024, and more metadata will be added to the catalogue over time.

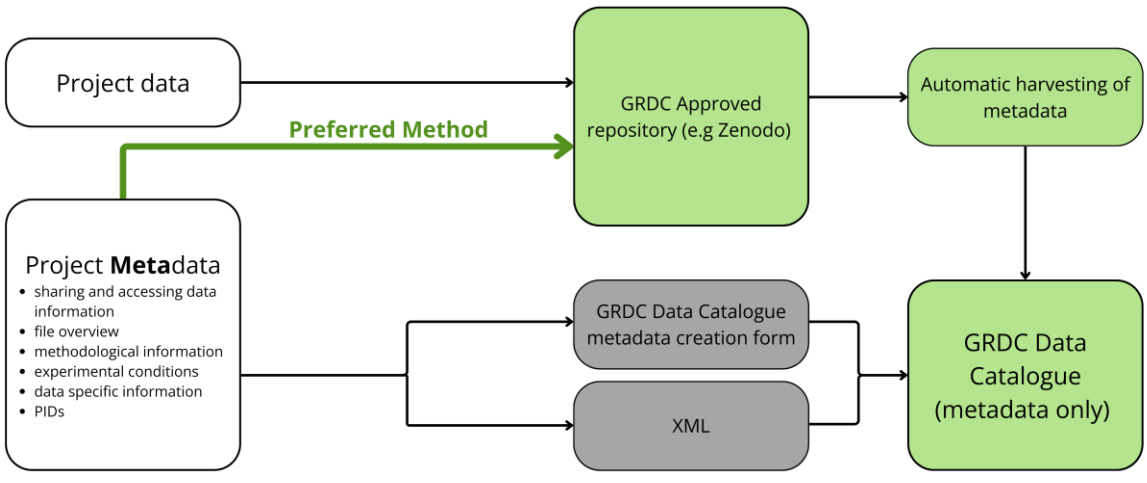
The GRDC Data Catalogue will only ever store project metadata. This metadata will describe the data in detail, including information like (but not limited to) keywords, authors, institutions, regions and collection period. Project data will be stored on an approved GRDC data repository.

The GRDC Data Catalogue has a search function that appears on the GRDC Data Catalogue home page. There are also advanced search capacities, allowing users to filter by item type, status, registration authority, update date and creation date.

If you would like to request data from the GRDC Data Catalogue, you must submit the 'Data Request Form', which is available in the toolbar on the GRDC Data Catalogue website.

11. Registration of Project Metadata to GRDC Data Catalogue

There are several ways to register Project Metadata describing RD&E Datasets with the GRDC Data Catalogue. The easiest, and preferred way to register Project Metadata with the GRDC Data Catalogue is through automated harvesting from an Approved Repository.



If you have uploaded the project data and metadata correctly to an approved repository there should be no issues in registering the project metadata to the GRDC Data Catalogue. You will not receive a notification regarding the registration of project metadata to the GRDC Data Catalogue through automated harvesting.

Resources:

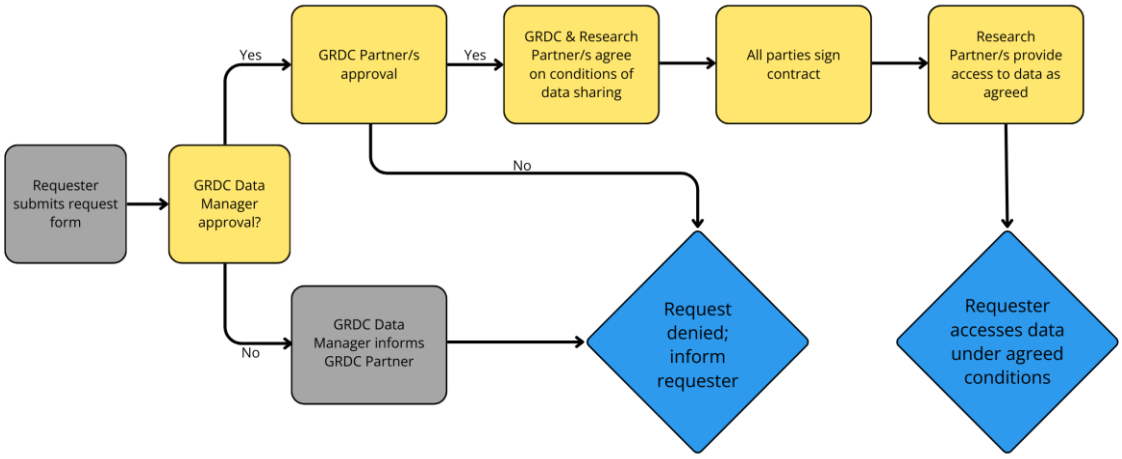
- GRDC RD&E Data Capture And Storage – Guidelines For Research Partners

12. Accessing data on the GRDC Data Catalogue

Research partners will be notified of all dataset access requests on the GRDC Data Catalogue. Decisions on all co-owned RD&E data are made by the co-owners. Upon receiving a request for data reuse, GRDC will assess whether to share the data and inform the co-owner, who will also decide.

If your data has been made available as Open access, your own organisational or institutional repository staff may be able to identify who has accessed your data. Open access means RD&E Data are publicly available and accessible by a third party without permission from the data owners.

If your data has been made available as Restricted access, no one can access your data without the approval of both the GRDC and all data owners (the research organisation) – in that approval process, you will be given the details of the requestor. The approval process for restricted access data is shown below.



If you would like to access data on the GRDC Data Catalogue, you must complete the GRDC RD&E Data Request Form. The GRDC Data Manager will get back to you about your intended use and seek permission/access from the relevant Research Partner.

Resources:

- GRDC RD&E Data Capture And Storage – Guidelines For Research Partners
- GRDC RD&E Data Request Form.
- Guideline for processing RD&E Data Access requests for re-use by Research Partners

13. Where to access support – Support Services Centre

Support is always available for any questions you might have.
Contact email: dataharvest@grdc.com.au
Contact phone: 0419 739 948

The GRDC website also contains information about the GRDC Data Management policy, approved repositories and metadata collection.

DataHarvest is an investment of the Grains Research and Development Corporation (GRDC). It is led by Curtin University through the Centre for Crop and Disease Management (CCDM).

GRDC code: CUR2401-001BGX