Getting started with the Data Citation Index

Participating in the Data Citation Index (DCI) is an opt-in service that involves close collaboration between Clarivate Analytics, ARDC and provider institutions to assess records and establish business processes for the production feed. For some providers this may require optimisations to their RIF-CS metadata before a production harvest to DCI can be established. Throughout the process, Clarivate Analytics works closely with the data provider to ensure correct representation of repository information as all material deposited in a given repository is linked to that repository record, thereby raising the visibility of the repository within the Web of Science and positioning data as a first class research object, alongside the scientific research literature to which they relate.

In order for Clarivate Analytics to provide appropriate attribution, certain metadata are needed to create a data citation which can be matched to a data citation in the literature and which provide access to the actual data in the repositories to allow reuse and citation as part of the data lifecycle. The threshold of metadata needed to do this is relatively low:

- Author/creator: the person(s), department(s) or institution(s) who created the data and should receive attribution; data objects without attribution of this type can be assigned as having 'Anonymous' authorship which may result in lack of attribution to the actual authors of the data. Fully parsed and fielded data for each author is preferred rather than including all authors in a single field
- Data object title
- Source: the repository or data provider enabling access to the data themselves. The source listed in Research Data Australia metadata should be the data provider or repository where the data are actually held and which should be included in a data citation, thereby enabling Clarivate Analytics to track the reuse of the data through citation in the scientific research literature; the clickable link provided via a persistent identifier or URL should provide access to the full data deposited, not to a secondary catalogue record or institutional web page with no data access. Where the source of a catalogue record is included as the source of the actual data, these records are not suitable for DCI.
- Source location: URL, or preferably DOI (or some other persistent identifier) which can be tracked and used to link the user to the source of the data in the repository which offers access
- Publication Year: the year the data were published – made available for reuse

Research Data Australia contributors who are able to provide these metadata and fulfill the DCI selection criteria, are eligible for inclusion in DCI and citations to the data objects can be tracked. In return, if selected for inclusion, the data provider will have access to DCI to enable them to review the implementation of their data.

How to get started with the Data Citation Index (DCI)

1. Contact your Outreach Officer or services@ardc.edu.au to express an interest in establishing a DCI harvest.
2. With ARDC, review and discuss record quality and transform as well as the proposed business processes and agree to proceed (See 'Assessing your records for DCI readiness' below)
3. ARDC will provide an initial harvest from the data source to DCI and advise Clarivate Analytics of the nominated contact for the data source.
4. Clarivate Analytics will assess a sample of records in the DCI output against their criteria for inclusion. They will also check quality of content, compliance with the DCI metadata schema and the richness of the record as assessed against the content available in the source repository.
5. Clarivate Analytics staff will liaise directly with the nominated contact for the data source to discuss the metadata assessment and to create a Repository Record for the data source in DCI. This record provides the Repository Name in each DCI record. All collection records for the data source will be linked to this record in DCI. The screenshot below shows an example (see Fig 1).
6. A production harvest from the data source to DCI is established.
7. Clarivate Analytics will provide a DCI admin login for use by the nominated data source contact.
8. Records are re-harvested from Research Data Australia to DCI on a regular basis.

Assessing your records for DCI readiness

An early step in establishing a harvest to DCI is to review the DCI transform of a representative sample of records from your data source. While the focus here is on the transform of records, it is important to also carefully review the accuracy and completeness of content in your records. Incorrect content (for example, misspelling of names) will affect the discoverability and capture of citation metrics for your records. It is also important that the records describe objects that are in scope for the DCI, e.g. they are not secondary records describing data held elsewhere.

To enable you to review your records, ARDC has:

- documented the RIF-CS to DCI transform mapping
- created a simple web service that enables Data Source Administrators (DSAs) to preview records in their data source that have been transformed to the DCI metadata format (XML output) using the mapping. To use the web service, in the Production or Demo environment, you need to have DSA permissions.

Fig 1: DCI Repository record. All records from a data source will be linked to this record in the Data Citation Index