REPOSITORY EVALUATION, SELECTION, AND COVERAGE POLICIES
FOR THE DATA CITATION INDEX℠ WITHIN THOMSON REUTERS WEB OF KNOWLEDGE℠
With the ever increasing amount of digital data being produced and made available, either voluntarily or through policy requirements from grant funding agencies, the need to discover and provide credit for the creation of scholarly research data has never been greater. Starting in 2012, Thomson Reuters will include Data Citation IndexSM within Web of KnowledgeSM, allowing for search and discovery of the scientific research data and links to the published literature with appropriate citation metrics. Thomson Reuters must balance the selection of material for inclusion in Data Citation Index with the ever-increasing abundance of digital web based resources. This essay sets forth the criteria and procedures for inclusion in the new Index.

As always, Thomson Reuters remains responsive to new, innovative developments in data publication which share our mission to bring the existence of the data to the attention of the scholarly community.

**SELECTION**

Research data considered for inclusion include data studies, and data sets deposited in a recognized repository.

**Definitions:**

- **Data repository:** a database or collection comprising data studies, and data sets which stores and provides access to the raw data. Constituent data studies, and sometimes individual data sets, are marked up with metadata providing a context for the available raw data.

- **Data study:** description of studies or experiments held in repositories with the associated data which have been used in the data study. (Includes serial or longitudinal studies over time). Data studies can be a citable object in the literature and may have cited references attached in their metadata, together with information on such aspects as the principal investigators, funding information, subject terms, geographic coverage etc. The level of metadata provided varies between repositories.

- **Data set:** a single or coherent set of data or a data file provided by the repository, as part of a collection, data study or experiment. Data sets may exist in a number of file formats and media types: they may be number based files such as spreadsheets, images, video, audio, databases etc. Data sets can be a citable object in the literature and may include cited references attached in their metadata, but more commonly they inherit the metadata of the overall study in which they are used.

**THE EVALUATION PROCESS**

Repository identification and selection are continuous and ongoing at Thomson Reuters, with repositories added as frequently as weekly. Moreover, existing coverage is constantly under review. Repositories now covered are monitored to ensure that they remain available and are maintaining high standards and a clear relevance to the Data Citation Index product. The repository selection process described here is applied to all resources covered in the Data Citation Index.

Many factors are taken into account when evaluating repositories for coverage, ranging from both qualitative and quantitative. The repository’s basic publishing standards, its editorial content, the international diversity of its authorship, and the citation data associated with it are all considered. No one factor is considered in isolation, but by combining and interrelating the data, the editor is able to determine the repository’s overall strengths and weaknesses.

Thomson Reuters editors who perform the evaluation have educational backgrounds relevant to their areas of responsibility, and understand the data held by the repositories they review.
Primary selection is at the level of the repository where evaluation includes:

- Subject
- Editorial content and repository attributes
- Geographic origin and scope

Once a repository is accepted for inclusion, further evaluation determines the appropriate metadata elements which will be captured to allow discovery and citation.

**BASIC REPOSITORY PUBLISHING STANDARDS**

**Persistence and stability**

Persistence of a repository and the data deposited within it is a basic criterion in the evaluation process. A repository must demonstrate longevity to be considered for initial inclusion in *Data Citation Index*. Thomson Reuters also reviews whether new data is currently deposited; a steady flow of newly deposited data is taken as an indicator that the resource is currently active. Generally, the data should be deposited with the repository, rather than the repository simply holding metadata and a web link to a remote/external source for the data. This ensures robust citation to the data to enable citation metrics and data re-use. A clear definition of the data-publication process with an indication of the data provider/creator’s affiliation should, ideally, be indicated. When a repository is selected for coverage, all deposited data is included in *Data Citation Index*; there is no sub-repository level selection other than to exclude data which is referenced rather than deposited.

**Funding statements**

The *Data Citation Index* aims to promote citation of data and link data to the research literature. To this end, particular consideration is given to repositories which show literature provenance and are accompanied by grant funding information.

**English language metadata**

English is the universal language of science at this time in history. It is for this reason that Thomson Reuters focuses on repositories that publish metadata in English or, at the very least, allow provision of sufficient descriptive (metadata) information in English. Some repositories covered in *Data Citation Index* publish only metadata descriptions in English with the actual data in another language. However, going forward, it is clear that the repositories most important to the international research community will publish data in English. This is especially true in the natural sciences. In addition, all repositories must have metadata and citations in the Roman alphabet.

**Peer review**

While peer review of deposited data is by no means universal, application of the peer-review process is another indication of repository standards and signifies overall quality of the data presented and the completeness of any cited references. It is also recommended that whenever possible, each repository, data study or data set is published with information on the funding source supporting the research presented.

**Age of material**

In addition, Thomson Reuters must form a judgement on the long-term preservation and sustainability of the repository and research data. There are no restrictions on the age of the deposited data. As a multidisciplinary service, the disparate attitudes and requirements of researchers across the various disciplines with regard to “older” data are acknowledged. Timeliness is also no restriction. As grant-funded projects draw to a close, it is accepted that the valued research output presented will not necessarily be updated in future, yet it will continue to be cited and may be reused in current research; there may also be delays in data publication compared to the corresponding research article due to embargos defined by authors and/or funding bodies.

**Links to the research literature**

To promote standards for data citation, and, subsequently, measure the impact of this growing body of scholarship, priority will be given to data repositories that show the provenance relating the data set to the research literature that either produced or re-used the data.

Again, no one factor is considered in isolation, but by combining and interrelating the data, the repository’s overall strengths and weaknesses can be evaluated. The Thomson Reuters staff performing these repository evaluations have advanced-degree-level educational backgrounds relevant to their areas of responsibility.

**EDITORIAL CONTENT**

Thomson Reuters includes research data from three major subject areas: Science & Technology, Social Sciences, and Arts & Humanities. Individual repositories may be multidisciplinary, inter-disciplinary or may have a narrow focus in order to qualify for inclusion. With an enormous amount of data readily available to them, and their daily observation of the international data landscape, Thomson Reuters editors are well positioned to spot emerging topics and active fields.
INTERNATIONAL DIVERSITY

While Thomson Reuters looks for international diversity among the repository’s contributing authors, editors, data producers, and deposited data, with the aim of providing information for an international audience, the importance of local and regional cyberscholarship is also given due consideration. Selection criteria are applied consistently across all repositories, irrespective of geographic coverage (international, national, regional or institutional), or whether the repository is multidisciplinary or has a narrow subject focus.

DATA CITATION & STANDARDS

While the research community has a strong desire to see data citation and attribution, there are no consistent standards and the occurrence of data in cited reference bibliographies of research articles is rare. To this end, Thomson Reuters encourages data citation by providing a standardized citation format for each record. In determining the citation format a number of proposed standards were evaluated. The DataCite citation standard has been adopted by Thomson Reuters due to its general acceptance and its ability to be applied to a wide range of data types and disciplines.

As data citation and the ability to link data repository content to the literature remains of high importance to Thomson Reuters and the research community, repositories are given priority if they provide references which either cite the deposited data, or which are cited by the deposited data record.

RECOMMENDING A REPOSITORY FOR COVERAGE

To recommend a particular data repository for coverage, please send details to tr.datarepository@thomsonreuters.com, and include details such as the URL which provides electronic access.

References:

Available online: http://www.dcc.ac.uk/resources/how-guides/cite-datasets

doi: 10.1087/095315108X254476

DataCite. Why Cite Data? http://www.datacite.org/whycitedata