J Comput Aided Mol Des (2014) 28:1043–1048
DOI 10.1007/s10822-014-9768-5

Encouraging data citation and discovery with the Data Citation Index

Megan M. Force · Nigel J. Robinson

Received: 2 April 2014 / Accepted: 23 June 2014 / Published online: 1 July 2014

Abstract An overview of the Data Citation Index is provided. Thomson Reuters developed this resource in response to a stated desire among members of the research community for increased attribution of non-traditional scholarly output. Launched in October of 2012 on the Web of science research platform, its aims include linking published research articles to their underlying data sets and tracking the citation of the data, as well as encouraging bibliographic citation of data. Cross-disciplinary search capabilities in the Index enable new possibilities for data discovery and synthesis. Data repositories are evaluated with respect to various selection criteria, with particular attention to their relevance to scientific and scholarly research. Index content reflects current data deposition practices. As data citation standards and practices continue to move toward widespread formalization and adoption, the initiative seeks to address issues of data citation, reuse, and author credit in a developing climate.

Keywords Data citation · Data Citation Index · Data repositories · Metadata

Introduction

Data preservation and access policies for various research journals, governments, and funding organizations have recently been enacted [1–4], altering the landscape for scientific and scholarly research authors. These policies may include requirements for data deposition, openness, persistence, and resolvable uniqueness in the form of permanent, citable identifiers. While there is evidence to suggest that digital data deposition increases the citation rate of published research [5], many researchers feel that they do not receive proper credit for the non-traditional scientific output they create (Fig. 1). The Data Citation Index aims to solve four key researcher problems:

- Data access and discovery: non-traditional scientific output is growing, yet remains difficult to find through traditional means.
- Data citation: while most researchers agree that non-traditional scholarly output should be cited, a minority of university administrators have policies addressing such citations. Citation to this output is inhibited by the lack of a standard citation style.
- Lack of willingness to deposit and cite data: inconsistent attribution for non-traditional scholarly output inhibits data sharing and citation.
- Lack of recognition and credit: the absence of an adequate peer-review process is a primary barrier to citation as well as to consideration of non-traditional scholarly output in academic rewards such as promotion and tenure.

Increasingly, authors of research deposit their data sets in discipline-specific or multidisciplinary data repositories, including those created by governmental organizations, academic institutions, and private groups. Repository selection on the part of researchers may depend upon a variety of factors, including the resources required for deposition, access (open or restricted), attribution, funding and persistence, and relevance of the created data to the areas of study represented by the data repository.

M. M. Force (✉)
Thomson Reuters, 1500 Spring Garden Street, Philadelphia,
PA 19130, USA
e-mail: megan.force@thomsonreuters.com

N. J. Robinson
Thomson Reuters, Enterprise House, Innovation Way,
Heslington, York YO10 5NQ, UK

Data repository evaluation

The Data Citation Index looks to provide information on data deposited in a source repository from which the data may be obtained. To be indexed in the Data Citation Index, a data repository must be a holder and distributor of data; this excludes some other web resources, such as institutional publications repositories which do not hold data, and metadata catalogues, which reference data stored elsewhere. Data repositories selected for inclusion in the Data Citation Index must also be demonstrably active, whether by continued maintenance and curation of the data sets held, or by addition of new materials, evidenced by data deposition statistics. Other factors of interest include evidence of repository persistence, thoroughness and detail of descriptive information (metadata), and indications of data reuse, such as in the form of established connections between datasets and published literature. Over 1,000 data repositories have been identified; a significant number of these have been rejected as not properly meeting the stated criteria for inclusion [6], while others are unable to deliver metadata or cannot meet requirements at this time. Currently over 150 source repositories are indexed in the resource, where representation of an individual repository varies between a single data record and several hundred thousand records. Data repository coverage in the Index by broad domain and geographical region are shown in Figs. 2 and 3, respectively.

Data Citation Index editors work to identify data repositories which are useful and relevant not only to the communities of researchers depositing data therein, but to researchers outside of these communities as well. As in a traditional Web of Science tool, queries may be refined by subject category to search for data in a particular area of the Sciences, Social Sciences, or Arts and Humanities. Also, general searches may lead to the discovery of datasets relevant to a particular subject of study, yet housed in a repository which specializes in data from other areas. For example, a topic search for nitrogen-related data in the Data Citation Index returns results from some 62 repository sources, including data related to molecular compound analysis, water and atmospheric concentrations, as well as agricultural statistics on inorganic fertilizer consumption. Research areas currently represented in this topic search can be seen in Fig. 4. In cases where researchers and their immediate colleagues are most familiar with data and repositories particular to their own discipline, this allows for a broadening of scope of a particular research project, as well as collaboration with scholars and laboratories previously unknown to discipline-specific researchers. By providing a search of literature and data together, data are raised to a first class research object alongside the literature.

Index record creation

In order for data records to be created in the Data Citation Index, data repositories provide the metadata describing

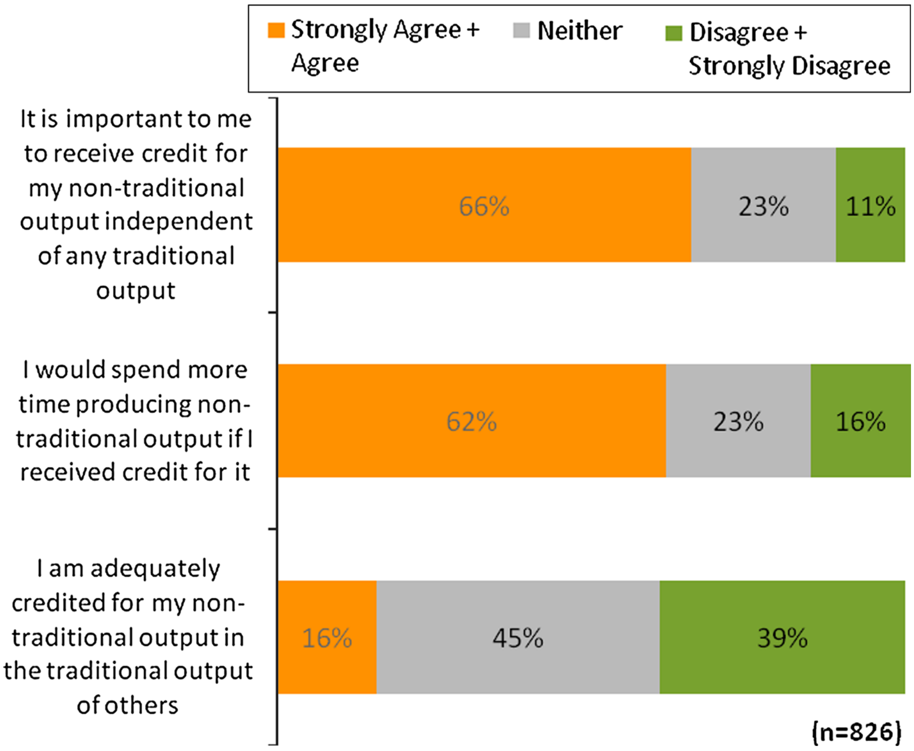


Fig. 1 Results of 2010 Thomson Reuters researcher survey (826 respondents)

Fig. 2 Data Citation Index coverage by broad discipline

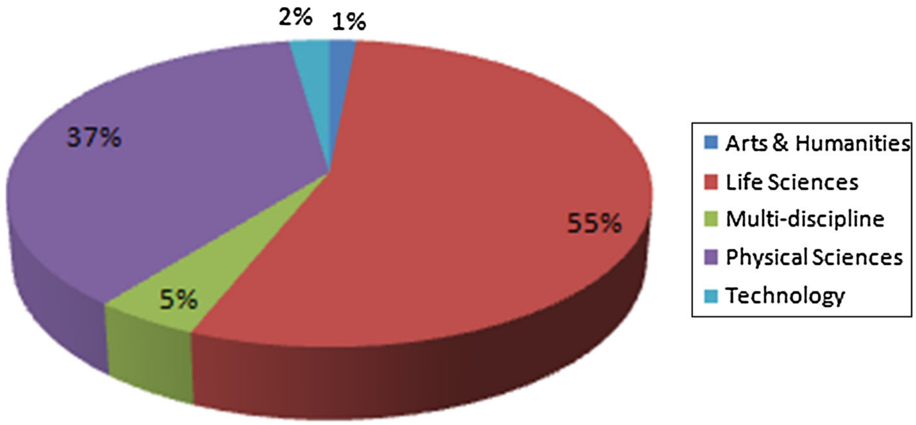
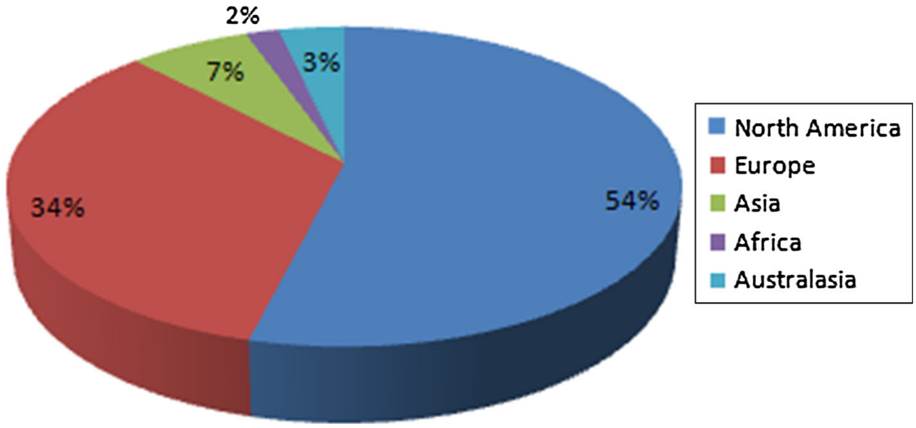


Fig. 3 Data Citation Index coverage by geographical region

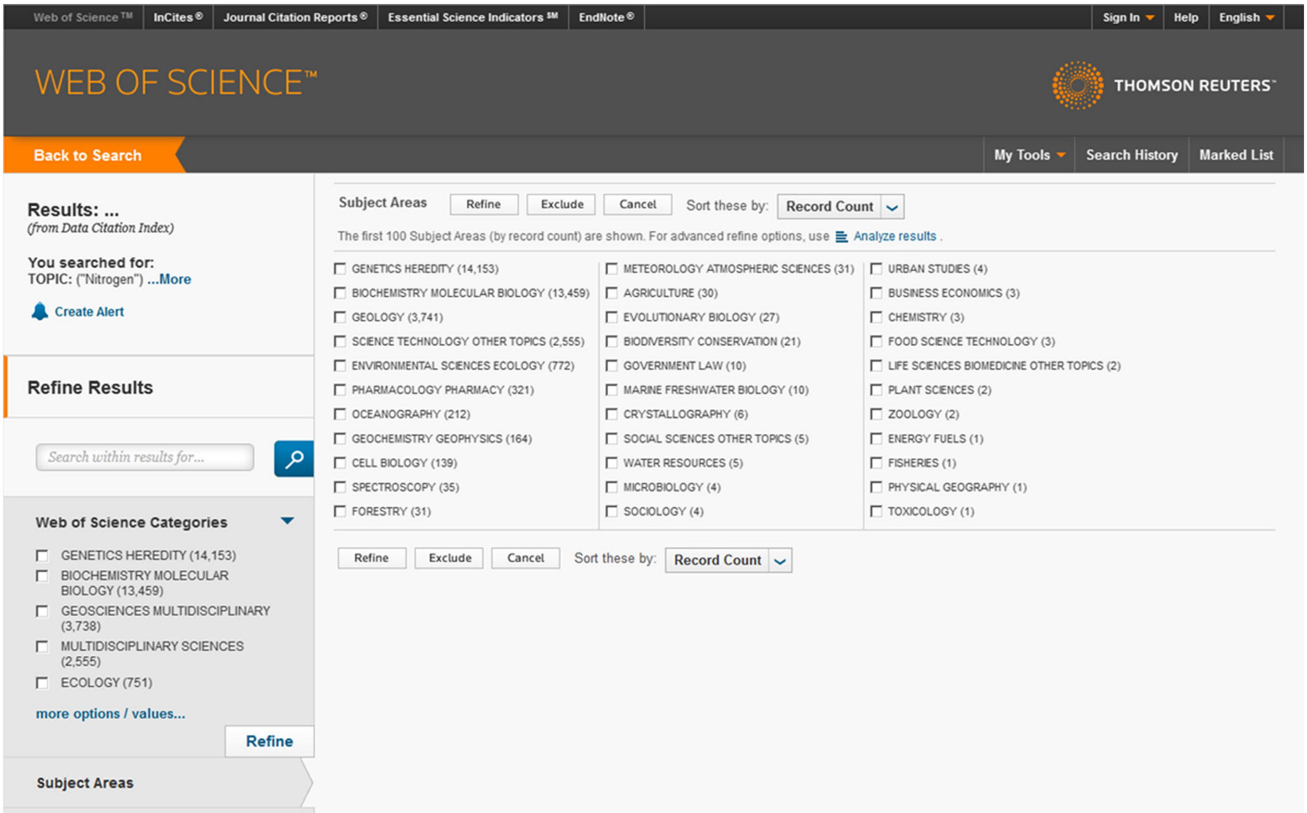


their data to Thomson Reuters; the data themselves remain at the repository, while their retrieval and reuse remain subject to the policies and restrictions of the repository and owners of the data. The provided feeds must contain certain required metadata elements, sufficient for creating citations to datasets. Additional, optional indexing terms enrich the Data Citation Index records, to enhance search and discovery. The metadata from each source repository is analyzed to establish the extent of these required and optional elements. Metadata may be provided in a variety of formats; incoming metadata may need to be normalized in order to have uniform structure for import into the editorial system, as well as to allow consistent search indexes to be built.

Where metadata are not provided by the repository, some manual indexing and text mining is applied to gain additional terms for discovery and description. In the case of some elements, Thomson Reuters taxonomies/thesauri are employed to provide additional terminology to allow broadening of searches and placement of terms in context in order to assist with disambiguation (e.g. subject categories, organism classifications, geographical terms). Data Citation Index records are created at three levels depending on the repository:

- Repository/source: comprises data studies and/or data sets. Stores and provides access to the raw data.
- Data study: descriptions of studies or experiments with associated data which have been used in the data study. Includes serial or longitudinal studies over time.
- 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.

Records created at the level below Repository/Source are dependent upon repository content. In some cases, selected data repositories are best represented at the source level alone, due to the structure and organization of the data, the level of metadata available, and the retrieval mechanism for data download/display, such as through a structured, non-static database query. As a result, content in the Index reflects variation in metadata practice by discipline or sub-discipline. Data Citation Index editors engage in detailed discussions with repositories with respect to modeling the data they hold, in order to establish the appropriate level of coverage/granularity for data citation purposes, as well as to ensure the data is accurately represented in the resource. As of the writing of this article, nearly 4 million records for data repositories, data studies, and data sets are available in the Index to be searched.



Results: ...
(from Data Citation Index)

You searched for:
TOPIC: ("Nitrogen") ...More

Create Alert

Refine Results

Search within results for...

Web of Science Categories

□ GENETICS HEREDITY (14,153)
□ BIOCHEMISTRY MOLECULAR BIOLOGY (13,459)
□ GEOSCIENCES MULTIDISCIPLINARY (3,738)
□ MULTIDISCIPLINARY SCIENCES (2,555)
□ ECOLOGY (751)

more options / values...

Refine

Subject Areas

Subject Areas Refine Exclude Cancel Sort these by: Record Count

The first 100 Subject Areas (by record count) are shown. For advanced refine options, use Analyze results .

□ GENETICS HEREDITY (14,153)
□ BIOCHEMISTRY MOLECULAR BIOLOGY (13,459)
□ GEOLOGY (3,741)
□ SCIENCE TECHNOLOGY OTHER TOPICS (2,555)
□ ENVIRONMENTAL SCIENCES ECOLOGY (772)
□ PHARMACOLOGY PHARMACY (321)
□ OCEANOGRAPHY (212)
□ GEOCHEMISTRY GEOPHYSICS (164)
□ CELL BIOLOGY (139)
□ SPECTROSCOPY (35)
□ FORESTRY (31)
□ METEOROLOGY ATMOSPHERIC SCIENCES (31)
□ AGRICULTURE (30)
□ EVOLUTIONARY BIOLOGY (27)
□ BIODIVERSITY CONSERVATION (21)
□ GOVERNMENT LAW (10)
□ MARINE FRESHWATER BIOLOGY (10)
□ CRYSTALLOGRAPHY (6)
□ SOCIAL SCIENCES OTHER TOPICS (5)
□ WATER RESOURCES (5)
□ MICROBIOLOGY (4)
□ SOCIOLOGY (4)
□ URBAN STUDIES (4)
□ BUSINESS ECONOMICS (3)
□ CHEMISTRY (3)
□ FOOD SCIENCE TECHNOLOGY (3)
□ LIFE SCIENCES BIOMEDICINE OTHER TOPICS (2)
□ PLANT SCIENCES (2)
□ ZOOLOGY (2)
□ ENERGY FUELS (1)
□ FISHERIES (1)
□ PHYSICAL GEOGRAPHY (1)
□ TOXICOLOGY (1)

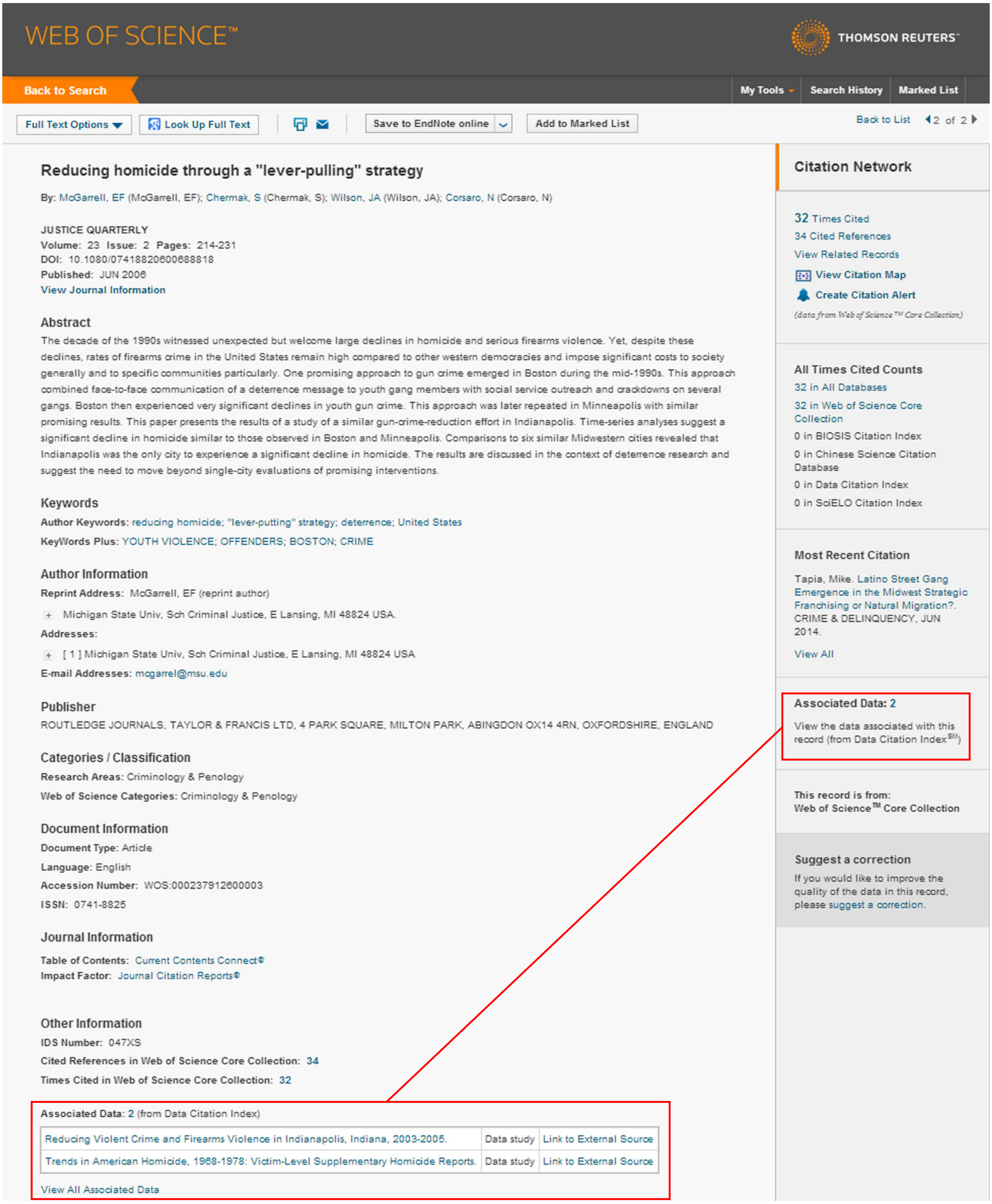
Refine Exclude Cancel Sort these by: Record Count

Fig. 4 Research areas represented in the Data Citation Index through a search for topic 'Nitrogen'

Data citation, linking, and future developments

Formal bibliographic data citation has yet to be widely adopted across various academic and research disciplines [7]. As there are no well-defined data citation standards commonly in use, the Data Citation Index encourages data citation by providing a recommended citation for each data object, following the DataCite recommendations [8]. Other scenarios unique to non-traditional scholarly output also must be modelled and affect the process of creating and updating records in the resource. Data, for instance, may be updated or added to after their initial publication. This is unlike traditional research literature where new or corrected information is presented in the form of a separate article or erratum. While some data repositories preserve and number all versions of the data to allow for individual citation, others may preserve only the most recent version, or may accumulate various versions in a single record. Accuracy in data citation, whether formal or informal, requires sufficient information for data retrieval and reuse, as well as an understanding of the level of information necessary for the proper interpretation of the data presented or discussed.

These issues are central to a primary goal of the Data Citation Index: connecting data sets to the scientific literature they inform. In addition to bibliographic data records being created from repository metadata, the initiative has gathered approximately 2 million literature citations from repository depositions and curations. These citations are employed to provide links between records for published articles in the Web of Science and their associated data source records in the Data Citation Index (Fig. 5). In its next phase, the initiative looks to expand these literature citations by gathering data citations from published articles. Currently, most data citations are informal mentions in the full text of the article. While a number of standards and recommendations have been proposed by bodies such as the Research Data Alliance and the Force 11 group [9], the adoption of formal data citation is not yet commonplace. The goal of the Data Citation Index is to harvest both formal and informal data citations from the literature, thence to provide recommended formal citations for these data objects, in addition to links to the full text of the citing publications and access to the data themselves. This will provide the basis for new data citation metrics and analytics.



Reducing homicide through a "lever-pulling" strategy

By: McGarrell, EF (McGarrell, EF); Chermak, S (Chermak, S); Wilson, JA (Wilson, JA); Corsaro, N (Corsaro, N)

JUSTICE QUARTERLY
Volume: 23 Issue: 2 Pages: 214-231
DOI: 10.1080/07418820600688818
Published: JUN 2006
View Journal Information

Abstract

The decade of the 1990s witnessed unexpected but welcome large declines in homicide and serious firearms violence. Yet, despite these declines, rates of firearms crime in the United States remain high compared to other western democracies and impose significant costs to society generally and to specific communities particularly. One promising approach to gun crime emerged in Boston during the mid-1990s. This approach combined face-to-face communication of a deterrence message to youth gang members with social service outreach and crackdowns on several gangs. Boston then experienced very significant declines in youth gun crime. This approach was later repeated in Minneapolis with similar promising results. This paper presents the results of a study of a similar gun-crime-reduction effort in Indianapolis. Time-series analyses suggest a significant decline in homicide similar to those observed in Boston and Minneapolis. Comparisons to six similar Midwestern cities revealed that Indianapolis was the only city to experience a significant decline in homicide. The results are discussed in the context of deterrence research and suggest the need to move beyond single-city evaluations of promising interventions.

Keywords

Author Keywords: reducing homicide; "lever-putting" strategy; deterrence; United States
KeyWords Plus: YOUTH VIOLENCE; OFFENDERS; BOSTON; CRIME

Author Information

Reprint Address: McGarrell, EF (reprint author)
+ Michigan State Univ, Sch Criminal Justice, E Lansing, MI 48824 USA.
Addresses:
+ [1] Michigan State Univ, Sch Criminal Justice, E Lansing, MI 48824 USA
E-mail Addresses: mcgarrel@msu.edu

Publisher

ROUTLEDGE JOURNALS, TAYLOR & FRANCIS LTD, 4 PARK SQUARE, MILTON PARK, ABINGDON OX14 4RN, OXFORDSHIRE, ENGLAND

Categories / Classification

Research Areas: Criminology & Penology
Web of Science Categories: Criminology & Penology

Document Information

Document Type: Article
Language: English
Accession Number: WOS:000237912600003
ISSN: 0741-8825

Journal Information

Table of Contents: Current Contents Connect®
Impact Factor: Journal Citation Reports®

Other Information

IDS Number: 047XS
Cited References in Web of Science Core Collection: 34
Times Cited in Web of Science Core Collection: 32

Associated Data: 2 (from Data Citation Index)

Reducing Violent Crime and Firearms Violence in Indianapolis, Indiana, 2003-2005.	Data study	Link to External Source
Trends in American Homicide, 1968-1978: Victim-Level Supplementary Homicide Reports.	Data study	Link to External Source

View All Associated Data

Citation Network

32 Times Cited
34 Cited References
View Related Records
View Citation Map
Create Citation Alert
(data from Web of Science™ Core Collection)

All Times Cited Counts

32 in All Databases
32 in Web of Science Core Collection
0 in BIOSIS Citation Index
0 in Chinese Science Citation Database
0 in Data Citation Index
0 in SciELO Citation Index

Most Recent Citation

Tapia, Mike. Latino Street Gang Emergence in the Midwest Strategic Franchising or Natural Migration?. CRIME & DELINQUENCY, JUN 2014.

View All

Associated Data: 2

View the data associated with this record (from Data Citation Index℠)

This record is from:
Web of Science™ Core Collection

Suggest a correction

If you would like to improve the quality of the data in this record, please suggest a correction.

Fig. 5 Web of science article record with links to underlying data

Conclusion

Thomson Reuters has launched a new initiative to solve key issues within the scientific and scholarly research community. The Data Citation Index continues to build content and develop infrastructure in the interest of improving attribution for non-traditional research output and enabling data discoverability and access. By encouraging data citation and facilitating connections between datasets and published literature, the resource elevates datasets to the status of citable and standardized research objects.

References

1. Canada’s Action Plan in Open Government. http://data.gc.ca/eng/canadas-action-plan-open-government. Accessed 16 March 2014
2. The Open Government Partnership: Second Open Government National Action Plan for the United States of America. http://www.whitehouse.gov/sites/default/files/docs/us_national_action_plan_6p.pdf. Accessed 16 March 2014
3. NSF Data Management Plan Requirements. https://www.nsf.gov/eng/general/dmp.jsp. Accessed 16 March 2014
4. Bloom T, Ganley E, Winker M (2014) Data access for the open access literature: PLoS’S data policy. PLoS Biol 12(2):e1001797. doi:10.1371/journal.pbio.1001797
5. Piwowar HA, Day RS, Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. PLoS One 2(3):e308. doi:10.1371/journal.pone.0000308
6. Repository Evaluation, Selection, and Coverage Policies for the Data Citation Index within Thomson Reuters Web of Science. http://wokinfo.com/products_tools/multidisciplinary/dci/selection_essay/. Accessed 21 May 2014
7. Downs RR, Chen RS (2010) Self-assessment of a long-term archive for interdisciplinary scientific data as a trustworthy digital repository. J Digit Inf 11 (1)
8. DataCite http://www.datacite.org/whycitedata. Accessed 16 March 2014
9. Joint Declaration of Data Citation Principles http://www.force11.org/datacitation. Accessed 16 March 2014