



Tracking and Managing Citations: Data Centers and Best Practices

W. Christopher Lenhardt

CIESIN – Columbia University

25 October 2006 – CODATA 2006

Outline

- Summarize the challenges
- Why do data citations matter
- Summarize CIESIN experience
- Related efforts
- Summary of potential best practices
- Additional thoughts



Challenges

- Citing digital data
 - Bits are still ephemeral
 - Standardization still in progress
- Sociology of science
- How to get credit for publishing data
- Rapidly changing technology
- Issue of *How* (theory) versus *Doing* (practice)

Related Issues

- Data quality
- Facilitate usage
- Attribution
- Provenance/authenticity



Address the problem from a different angle

- Potential contribution of data centers
 - Contribute to standards development
 - *Develop and promote best practices*

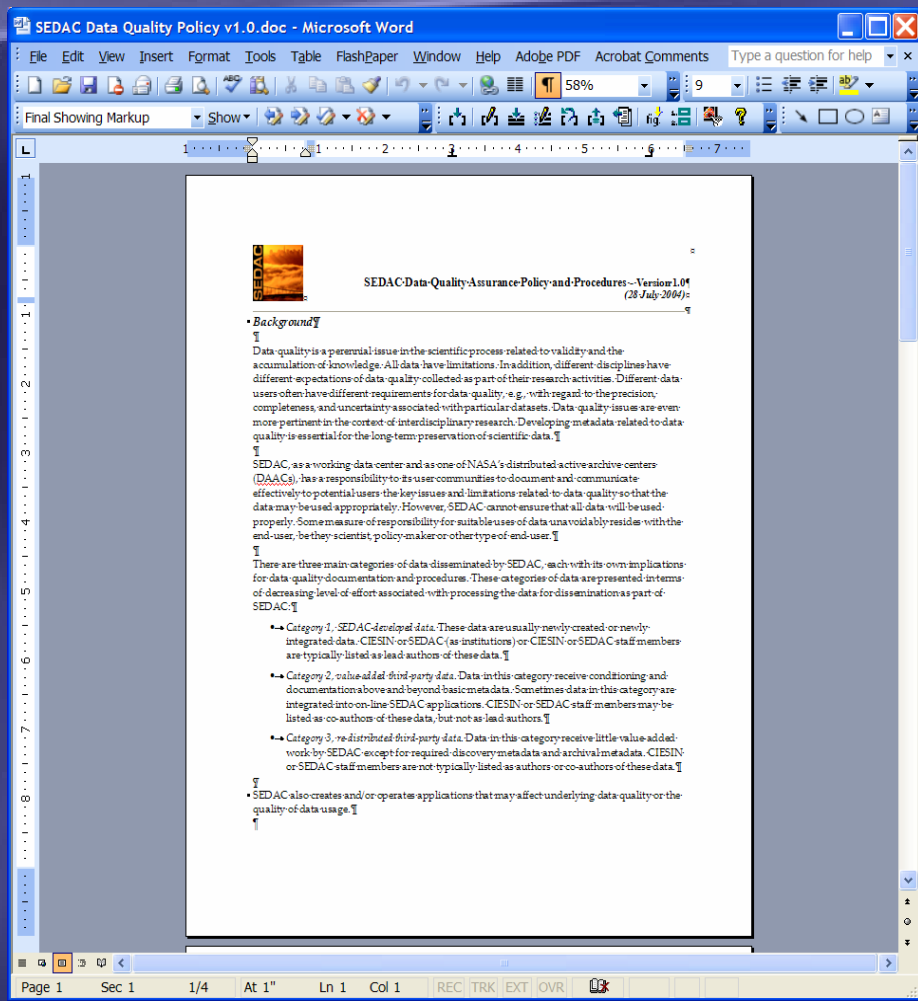
What do you need?

- Some policies
- Some procedures/operational practices
- Some content

Potentially Relevant Policies

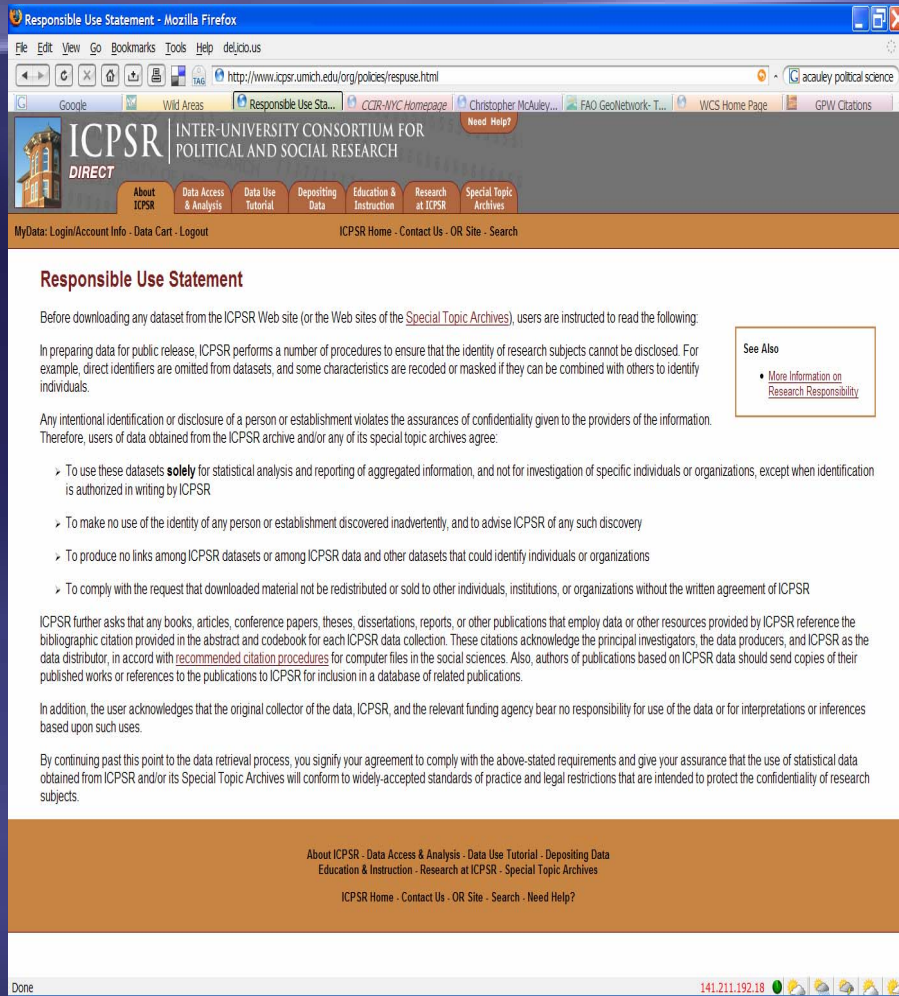
- Data quality policy (and procedure)
- Information quality policy (and procedure)
- Responsible use

Quality Review and Documentation



- What kinds of data and information
- Quality review and documentation
- Making quality information available to end-users

Responsible Use



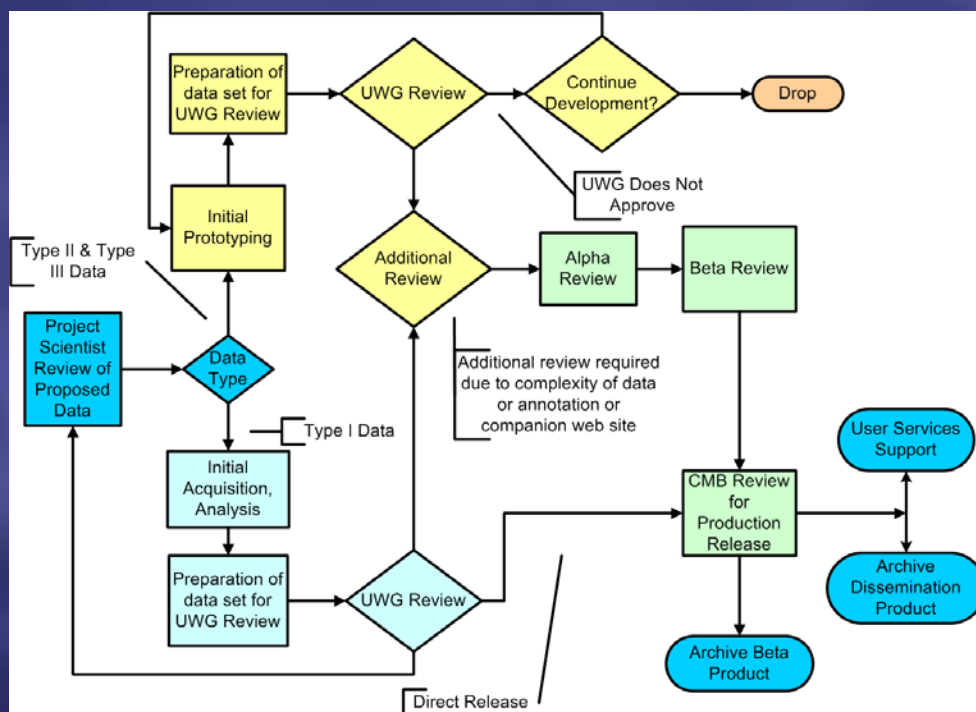
The screenshot shows a Mozilla Firefox browser window displaying the ICPSR Responsible Use Statement page. The browser's address bar shows the URL <http://www.icpsr.umich.edu/org/policies/respuse.html>. The page header features the ICPSR logo and the text "INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH". Below the header is a navigation menu with links for "About ICPSR", "Data Access & Analysis", "Data Use Tutorial", "Depositing Data", "Education & Instruction", "Research at ICPSR", and "Special Topic Archives". The main content area is titled "Responsible Use Statement" and contains several paragraphs of text, including instructions on data use, confidentiality, and citation requirements. A "See Also" box on the right side of the page contains a link to "More Information on Research Responsibility". The footer of the page includes a navigation menu with links for "About ICPSR - Data Access & Analysis - Data Use Tutorial - Depositing Data - Education & Instruction - Research at ICPSR - Special Topic Archives" and "ICPSR Home - Contact Us - OR Site - Search - Need Help?". The browser's status bar at the bottom shows "Done" and the IP address "141.211.192.18".

- Data providers have certain legal and ethical responsibilities related to data stewardship and dissemination
- Opportunity to remind users about issues such as attribution and confidentiality
- Can be a link
- Could pop up prior to a download
- <http://www.icpsr.umich.edu/org/policies/respuse.html>



Operational Practices

- Quality review and documentation
- Recommended citations
- Technical publications about data
- Citation style guides



Provide recommended citations

- Essential reminder/aid to facilitate citation
- Can be non-trivial depending on things like collections versus subsets
- Helpful to users to add a “download to a citation manager link”



A screenshot of a Mozilla Firefox browser window displaying a webpage titled 'Gridded Population of the World - GPW v3'. The browser's address bar shows the URL 'http://sedac.ciesin.columbia.edu/gpw/documentation.jsp'. The page content is under a 'CITATIONS' heading and provides recommended citation text for various GPWv3 data products. The products listed include: 'Gridded Population of the World, version 3 (GPWv3)', 'Population Grids', 'Population Density Grids', 'Land Area Grids', 'Mean Geographic Unit Area Grids', 'Centroids', and 'National Identifier Grid'. Each product has a corresponding paragraph of citation text, all starting with 'Center for International Earth Science Information Network (CIESIN), Columbia University; and Centro Internacional de Agricultura Tropical (CIAT). 2005. Gridded Population of the World Version 3 (GPWv3). Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia University. Available at http://sedac.ciesin.columbia.edu/gpw. (date of download)'. The browser's status bar at the bottom shows 'Done' and the IP address '129.236.39.66'.

Collect Citation Information



- Gives an indication of usage and quality
- Provides a reminder to users to cite data in their research and publications
- Ideally do this for all your data, but may be valuable for flagship data products
- Potential for automation?
 - Pull and push



Generate or Reference [Peer-reviewed] Publications or Technical Notes About the Data

Wiley InterScience :: Journal :: Article PDF - Mozilla Firefox

File Edit View Go Bookmarks Tools Help de|joo.us

CU Center for International ... Lamont-Doherty Earth ... SEDAC CCR-NYC WDC NBI-NIN Cingular Blackberry MBNA BNY Optimum Online Continental Airlines

INTERNATIONAL JOURNAL OF POPULATION GEOGRAPHY, VOL. 3, 203-225 (1997)

World Population in a Grid of Spherical Quadrilaterals

Waldo Tobler^{1*}, Uwe Deichmann², Jon Gottsegen¹, and Kelly Maloy³

¹ National Center for Geographic Information and Analysis, Geography Department, University of California, Santa Barbara, CA 93106-4060, USA
² Statistical Division of the United Nations, DC2-1764, New York, NY 10017, USA
³ ESRI, Redlands, CA 92373, USA

ABSTRACT

We report on a project that converted subnational population data to a raster of cells on the earth. We note that studies using satellites as collection devices yield results indexed by latitude and longitude. Thus it makes sense to assemble the terrestrial arrangement of people in a compatible manner. This alternative is explored here, using latitude/longitude quadrilaterals as bins for population information. This format also


correlation with global change studies, and more detailed information for some parts of the world. © 1997 John Wiley & Sons, Ltd.

Received 10 August 1996; revised 30 November 1996; accepted 16 December 1996
Int. J. Popul. Geogr. 3, 203-225 (1997)
No. of Figures: 5 No. of Tables: 1 No. of Refs: 58

Keywords: world population; raster; five-minute quadrilaterals

INTRODUCTION

1 of 23



BIOONE Online Journals - The Human Footprint and the Last of the Wild - Mozilla Firefox

File Edit View Go Bookmarks Tools Help de|joo.us

http://www.bioone.org/perferv?request=get-document&doi=10.1641/2F0006-3568%282002%29052%580891%3ATHFATL%5C%20

BIOONE Online Journals - The Hu... Center for International Earth Sci... Socioeconomic Data and Appl... CCR-NYC Homepage Optimum Online MyProfile Sign in Register

The Human Footprint and the Last of the Wild

Article pp. 891-904 | PDF (13.12M)

The human footprint is a global map of human influence on the land surface, which suggests that human beings are stewards of nature, whether we like it or not
ERIC W. SANDERSON, MALANJING JAITEH, MARCA A. LEVY, KENT H. REDFORD, ANTONETTE V. WANNIEBO, GILLIAN WOOLMER

DOI: 10.1641/0006-3568(2002)05200891:THFATL2.0.CO;2

In Genesis, God blesses human beings and bids us to take dominion over the fish in the sea, the birds in the air, and every other living thing. We are entrusted to be fruitful and multiply, to fill the earth, and subdue it (Gen. 1:28). The bad news, and the good news, is that we have almost succeeded.

There is little debate in scientific circles about the importance of human influence on ecosystems. According to scientists' reports, we appropriate over 40% of the net primary productivity (the green material) produced on Earth each year (Vitousek et al. 1986; Rostkizer et al. 2001). We consume 30% of the productivity of the oceanic shelf (Pauly and Christensen 1999), and we use 60% of freshwater run-off (Pielou et al. 1996). The unprecedented escalation in both human population and consumption in the 20th century has resulted in environmental crises never before encountered in the history of humankind and the world (McNeill 2000). E. O. Wilson (2002) claims it would now take four Earths to meet the consumption demands of the current human population, if every human consumed at the level of the average US inhabitant. The influence of human beings on the planet has become so pervasive that it is hard to find adults in any country who have not seen the environment around them reduced in natural values during their lifetimes—woodlots converted to agriculture, agricultural lands converted to suburban development, suburban development converted to urban areas. The cumulative effect of these many local changes is the global phenomenon of human influence on nature: a new geological epoch some call the "anthropocene" (Steffen and Tyson 2003). Human influence is arguably the most important factor affecting life of all kinds in today's world (Lande 1998; Terborgh 1999; Pimm 2001; UNEP 2001).

Yet despite the broad consensus among biologists about the importance of human influence on nature, this phenomenon and its implications are not fully appreciated by the larger human community, which does not recognize them in its economic systems (Hall et al. 2001) or in most of its political decisions (Soulé and Terborgh 1999; Chapon et al. 2000). In part, this lack of appreciation may be due to scientists' propensity to express themselves in terms like "appropriation of net primary productivity" or "exponential population growth," abstractions that require some training to understand. It may be due to historic assumptions about and habits inherited from times when human beings, as a group, had dramatically less influence on the biosphere. Now the individual decisions of 6 billion people add up to a global phenomenon in a way unique to our time. What we need is a way to understand this influence that is global in extent and yet easy to grasp—what we need is a map.

Until recently, designing such a map was not possible, because detailed data on human activities at the global scale were unavailable. The fortunate confluence of several factors during the 1990s changed this situation. Rapid advances in earth observation, using satellite technology pioneered by NASA and other space agencies, meant that, for the first time, verifiable global maps of land use and land cover were available (Loveland et al. 2000). The thawing of the cold war and calls for efficiency in government meant that other sources of global geographic data, for example, on roads and railways, were released to the public by the US National Imagery and Mapping Agency (Eltis 1997). Improved reporting of population statistics at subnational levels enabled geographers to create global digital maps of human population density (CIESIN et al. 2000). Finally, advances in geographic information systems (GIS) have provided the integration technology necessary to combine these data in an efficient and reproducible manner. Although the datasets now available are imperfect instruments, they are of sufficient detail and completeness that scientists can map the influence of humans on the entire land's surface.

Table of Contents

- Mapping The Human Footprint
- Finding The Last of The...
- Interpreting The Human...
- Implications For Conservation...
- Conclusions
- Acknowledgments
- References Cited
- Tables
- Figures

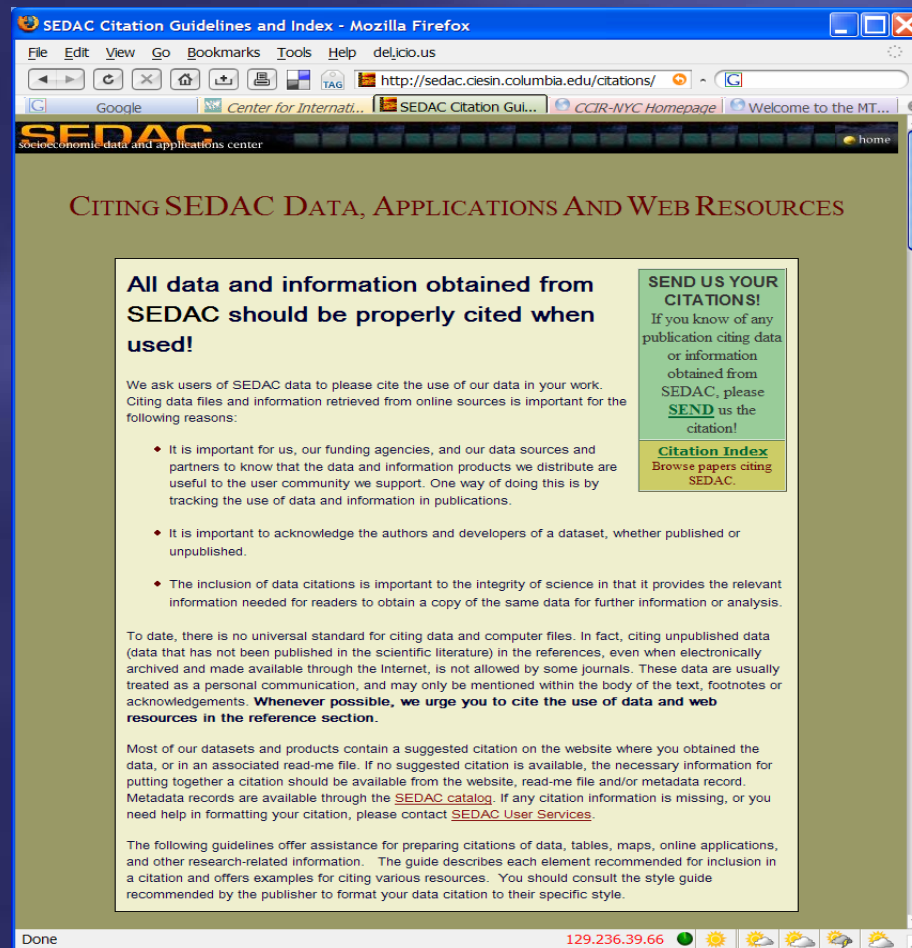
Options:

- Create Reference
- Email This Article
- Copyright Permissions

65.245.134.215

Provide Access to or Develop a Citation 'Style Guide'

- <http://sedac.ciesin.columbia.edu/citations>



The screenshot shows a Mozilla Firefox browser window displaying the SEDAC Citation Guidelines and Index page. The browser's address bar shows the URL <http://sedac.ciesin.columbia.edu/citations/>. The page content includes the SEDAC logo, the title "CITING SEDAC DATA, APPLICATIONS AND WEB RESOURCES", and a main section titled "All data and information obtained from SEDAC should be properly cited when used!". This section explains the importance of citing SEDAC data and provides three reasons: 1) It is important for funding agencies and data sources to know that their data is being used. 2) It is important to acknowledge the authors and developers of a dataset. 3) The inclusion of data citations is important to the integrity of science. A sidebar on the right contains a green box with the text "SEND US YOUR CITATIONS!" and a yellow box with the text "Citation Index Browse papers citing SEDAC.". The page footer includes the CIESIN Columbia University logo and the text "Done" and "129.236.39.66".

SEDAC Citation Guidelines and Index - Mozilla Firefox

File Edit View Go Bookmarks Tools Help deljcio.us

http://sedac.ciesin.columbia.edu/citations/

Google Center for Internati... SEDAC Citation Gul... CCIR-NYC Homepage Welcome to the MT...

SEDAC
Socioeconomic data and applications center

CITING SEDAC DATA, APPLICATIONS AND WEB RESOURCES

All data and information obtained from SEDAC should be properly cited when used!

We ask users of SEDAC data to please cite the use of our data in your work. Citing data files and information retrieved from online sources is important for the following reasons:

- It is important for us, our funding agencies, and our data sources and partners to know that the data and information products we distribute are useful to the user community we support. One way of doing this is by tracking the use of data and information in publications.
- It is important to acknowledge the authors and developers of a dataset, whether published or unpublished.
- The inclusion of data citations is important to the integrity of science in that it provides the relevant information needed for readers to obtain a copy of the same data for further information or analysis.

To date, there is no universal standard for citing data and computer files. In fact, citing unpublished data (data that has not been published in the scientific literature) in the references, even when electronically archived and made available through the internet, is not allowed by some journals. These data are usually treated as a personal communication, and may only be mentioned within the body of the text, footnotes or acknowledgements. **Whenever possible, we urge you to cite the use of data and web resources in the reference section.**

Most of our datasets and products contain a suggested citation on the website where you obtained the data, or in an associated read-me file. If no suggested citation is available, the necessary information for putting together a citation should be available from the website, read-me file and/or metadata record. Metadata records are available through the [SEDAC catalog](#). If any citation information is missing, or you need help in formatting your citation, please contact [SEDAC User Services](#).

The following guidelines offer assistance for preparing citations of data, tables, maps, online applications, and other research-related information. The guide describes each element recommended for inclusion in a citation and offers examples for citing various resources. You should consult the style guide recommended by the publisher to format your data citation to their specific style.

SEND US YOUR CITATIONS!
If you know of any publication citing data or information obtained from SEDAC, please **SEND** us the citation!

Citation Index
Browse papers citing SEDAC.

Done 129.236.39.66





Related Activities

Work at Harvard/MIT

- <http://gking.harvard.edu/files/cite.pdf>

A Proposed Standard for the Scholarly Citation of Quantitative Data¹

Micah Altman² Gary King³

March 2, 2006

¹Our thanks to Caroline Armas, Dale Flecker, Ann Green, Dave Kane, Gerome Miklau, Norman Paskin, Jeri Schneider, Karen Sullivan, Paul Uhler, and Mary Vardigan for helpful comments; and the Library of Congress (PA#NDP03-1), the National Science Foundation (SES-0318275, IIS-98747) and the National Institutes of Aging (P01 AG17625-01) for research support.

²Associate Director, Harvard-MIT Data Center (Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; <http://www.hmdc.harvard.edu/micah.altman/>, micah.altman@harvard.edu, (617) 496-3847).

³David Florence Professor of Government (Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; <http://gking.harvard.edu>, King@Harvard.Edu, (617) 495-2027).



IASSIST

• <http://iassistblog.org/?cat=17>

- Review of styles
- Subgroup working on the issue
- Blog

• <http://www.iassistdata.org/>

IASSIST Communiqué » Data Doc & Citation - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://iassistblog.org/?cat=17

CU CIESIN LDEO SEDAC CCIR-NYC WDC NBII-NIN Cingular Blackberry MBNA BNY Optimum Online

Wizz RSS 2.1.6 Feed Search Help etc. Options etc. Watch List Weather

Google Center for Inte... Socioeconomic... CCIR-NYC Ho... Optimum Online IASSIST Co...

IASSIST COMMUNIQUE

IASSIST
communiqué

A voice for research data and its preservation. Grasp the Issues!

Archive for the 'Data Doc & Citation' Category

Search

You are currently browsing the archives for the Data Doc & Citation category.

Archives

- » October 2006
- » September 2006
- » August 2006
- » July 2006
- » June 2006
- » May 2006
- » February 2006
- » January 2006
- » December 2005
- » November 2005
- » September 2005
- » August 2005
- » June 2005

Categories

- » Access (6)
- » Articles and News (9)
- » Book Review (1)

Cataloging query: how does cataloging data differ from traditional library material?

Monday, June 19th, 2006

In their never ending quest for information, Jen, Tiffani and Paula would like to know about people's experiences cataloging data. Do you use MARC format? Have you tried using DDI? At what level do you tend to catalog collections? How does it depart from cataloging traditional library matter? Enquiring minds want to know.

- [...]

Posted in IASSISTers, Professional Issues, Data Doc & Citation | 2 Comments »

Done 66.98.244.92 Now: Overcast, 41° F Tue: 48° F Wed: 50° F Thu: 52° F Fri: 47° F



Stats Canada

- Gaeton Drolet – Univ of Quebec Laval
- <http://www.statcan.ca/english/freepub/12-591-XIE/12-591-XIE2006001.htm>

How to Cite Statistics Canada Products - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.statcan.ca/english/freepub/12-591-XIE/12-591-XIE2006001.htm

Statistics Canada Statistique Canada

Français	Contact Us	Help	Search	Canada Site
The Daily Census	Canadian Statistics	Community Profiles	Our Products and Services	Home Other Links

How to Cite Statistics Canada Products

This guide has been developed for authors, editors, researchers, academics, students, librarians and data librarians. It describes, in [three steps](#), how to build your reference when citing Statistics Canada products.

Instructions are also available on how to cite [sources other than Statistics Canada](#).

Notice

Authors and researchers must give [full credit](#) for any Statistics Canada data, analysis and other content material used or referred to in their studies, articles, papers and other research works. The rules and examples set out in this citation guide have been developed to help you do this.

To cite this manual:

Statistics Canada. 2006. *How to Cite Statistics Canada Products*. Statistics Canada Catalogue no. 12-591-XWE. Ottawa. Version updated March 31, 2006.
<http://www.statcan.ca/english/freepub/12-591-XIE/12-591-XIE2006001.htm> (accessed date).

Home | Search | Contact Us | Français

Date modified: 2006-06-27

Important Notices

142.206.64.31 Now: Overcast, 41° F Tue: 48° F Wed: 50° F Thu: 52° F Fri: 47° F

Summary of Potential Best Practices

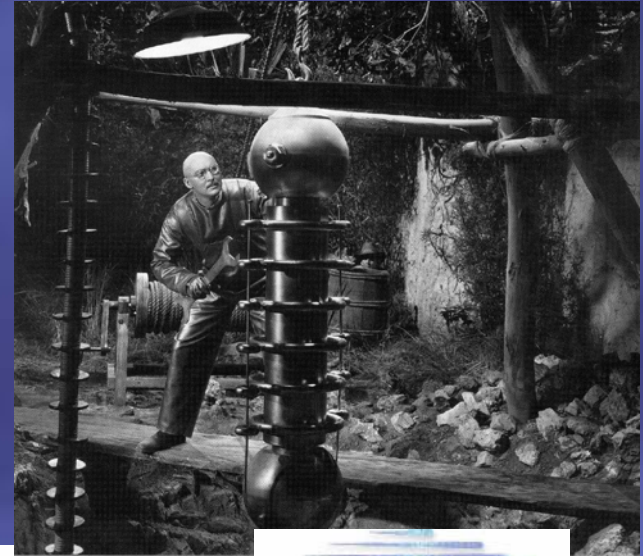
- Provide a recommended citation
- Provide access to guides on citation
- Encourage responsible use
- Publish about data in peer reviewed literature
- Collect citations to the data from other researchers and users

Additional Challenges

- Downloads of whole data sets versus subsets of data
- Composite data sets
 - Collections
 - Aggregations
- Resources may be limited; Can you do this for all of your holdings?
- It may not make sense to develop your own style guide, may be more efficacious to utilize a pre-existing guide
- Location and naming – for citations to be useful, the location must be stable
 - URNs/DOIs etc.

To Address the Larger Challenge Need to Involve

- Funders
- Publishers
- Professional associations
- Creators of data
- Other data centers



Founded 1903



Should we treat data more like a traditional publication?

- Research data is messy
- Persistence: Are data sets analogous to books?
- Do we need unique identifiers and/or catalog numbers for data sets?
 - ISBN v. catalog number

The diagram illustrates the structure of a library call number and an ISBN. The call number is **JK 216 T 7 1945 v.1**. Labels point to various parts:

- JK 216**: Specific Subject (United States) and Classification Number.
- T 7**: Aspect of the Subject (Constitutional History, 1821-1865) and Book number (usually derived from author's name).
- (.)**: Decimal Point indicating following information is filed as a decimal number.
- 1945**: Year of publication of the edition.
- v.1**: Volume number (in multivolume sets).

 Below the call number is the ISBN **90-5584-067-X** and its corresponding barcode. The barcode is labeled with the number **9 789055 840670** and the number **90000** in a box.

Photo # NH 96566-KN First Computer "Bug", 1945

92

9/9

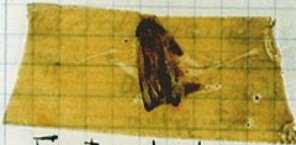
0800 Antam started
 1000 " stopped - antam ✓

13'00 (032) MP-MC ~~1.582140000~~ 1.2700 9.037 847 025
 (033) PRO 2 2.130476415 ~~2.130476415~~ 9.037 846 995 connect
 connect 2.130476415 4.615925059(-2)

Relays 6-2 in 033 failed special speed test
 in relay " 11.00 test.

(Relays changed)

1100 Started Cosine Tape (Sine check)
 1525 Started Multi-Adder Test.

1545  Relay #70 Panel F
 (moth) in relay.

1700/150 Antam started.
 1700 closed down.

Relay 3145
 Relay 3370



Thanks...

謝謝您

