IASSIST QUARTERLY

The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors

The QUARTERLY is normally published four times per year. Authors are encouraged to submit papers as word processing files (for further information see: http://www.iassistdata.org/iq/instructions-authors) Hard copy submissions may be required in some instances. Manuscripts should be sent to Editor: Karsten Boye Rasmussen.

The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago Press Manual of Style or Kate L. Turabian's Manual for Writers. Graphics: Image files should be sent as separate files in TIF format at 300 DPI resolution and maximum 680 pixels (width) x 800 pixels (height) Announcements of conferences, training sessions, or the like are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event.

Editor

Karsten Boye Rasmussen
Department of Marketing & Management
University of Southern Denmark, SDU
Campusvej 55, DK-5230
Odense M, Denmark
Phone: +45 6550 2115
Email:kbr@sam.sdu.dk

Production

Walter Piovesan
Maps/Data/GI Library
Simon Fraser University
Burnaby, B.C.
Canada V5A 1S6
Phone: (778) 782-5869
Email:walter@sfu.ca


CONTENTS


FEATURES

4 Editor's Notes
Karsten Boye Rasmussen

5 Guest Editor's Notes
Bobray Bordelon

6 The Pedagogical Data Reference Interview
Kristin Parlo

11 Sources for International Trade, Prices, Production, and Consumption
Amy West

16 Data Reference in Depth: Sources of International Labour Data
Walter W. Giesbrecht

22 Financial Crisis Data Resources: A Brief Guide
Mary Tao

25 Data in Development: an Overview of Microdata on Developing Countries
Kristi Thompson

31 The American Community Survey: Benefits and Challenges
Michele Hayslett and Lynda Kellam

40 IASSIST 2011 - Call For Papers

Title: Newsletter - International Association for Social Science Information Service and Technology
ISSN - United States: 0739-1137 © 1997 by IASSIST. All rights reserved.
Welcome to this double issue of the IASSIST Quarterly: volume 33 (number 4, 2009) and volume 34 (number 1, 2010).

The IASSIST Quarterly (IQ) has now published several double issues focusing on a particular theme. Often the issues have arisen from one or more sessions at an IASSIST conference or from workshops and other professional gatherings. Furthermore, the focused issues are often used in workshops and serve as a welcome collection of relevant and related papers for presentation and discussion.

The special issues of the IQ normally have guest editors. On this occasion we are happy to have Bobray Bordelon from Princeton University Library as our guest editor. The special theme is: The subject content and how researchers use the data. Bobray comments that we spend much effort concentrating on various technical aspects of data management and procurement; this special issue however looks into the actual subject matter of the data.

As editor in chief of the IQ I know that encouraging authors to turn their presentations into papers and pulling them together for publication can be a long-winded process with many repetitive steps. So much greater the contentment when the finished product surfaces. Special thanks to Bobray Bordelon and also thanks to Kristin Partlo, Amy West, Walter W. Giesbrecht, Mary Tao, Kristi Thompson, Michele Hayslett and Lynda Kellam for the compilation of this IQ publication.

Articles for the IQ are always very welcome. They can be papers from IASSIST conferences, other conferences, from local presentations or papers especially written for the IQ. If you don't have anything to offer right now, then please prepare yourself for the next IASSIST conference and start planning for participation in a session there. Chairing a conference session with the purpose of aggregating and integrating papers for a special issue IQ is much appreciated as the information reaches many more people than the session participants and will be readily available on the IASSIST website at http://www.iassistdata.org.

Authors are very welcome to contact me via e-mail: kbr@sam.sdu.dk, and should you be interested in compiling special issues for the IQ as guest editor(s) I’d be delighted to hear from you.

Karsten Boye Rasmussen     -    December 2010
In recent years, IASSIST conferences have tended to focus primarily on data documentation, infrastructure, technology, and other technical aspects. All of these topics are crucial to the process. However, the core is the actual data and understanding the subject matter. With the largest number of participants being specialists who help researchers find appropriate data, it was alarming that so little emphasis was placed on a deeper understanding of the subject content and how researchers use the data. A call to arms to focus on the actual data resulted in two sessions collectively titled “Data Reference in Depth” at the 2009 IASSIST Conference, “Social Data and Social Networking: Connecting Social Science Communities Across the Globe”. This special issue contains papers that emerged from these two sessions. The presentations have been expanded to further elaborate on data reference.

In most research settings, data librarians are required to be fluent in a range of sources dependent on the needs of their researchers. The data librarian needs to be prepared to move easily between often diverse subject areas. The data librarian has to transition from answering a question on population to trade to health to terrorism to religion and everything in between typically across all geographic boundaries and languages.

This special issue gives readers a sampling of a data librarian’s “day in the life” by highlighting a few major subject areas and their primary data sources. Readers will see a few of the particular reference challenges by getting an introduction to some key sources and particular challenges that crop up in a sampling of different areas. In addition, the question “how does providing access to data, as a unique format, affect library reference services?” will be answered.

Kristin Partlo examined how the data reference interview is especially crucial in the provision of research assistance, from the viewpoint of working with undergraduates at a small liberal arts college. The lessons learned have a bearing on research institutions of all sizes and provides a model for getting to the heart of what is really needed. Of particular importance is how to balance providing reference with instructing the researcher to learn from the experience. Partlo shares the data reference worksheet she devised and provides clear examples.

The issue then moves on to major resources available in five distinct subject areas. Amy West provides a detailed analysis of the sources and interfaces for international trade, commodities, prices and production. Walter Giesbrecht follows with a guide to labor looking at both national and international sources as well as the problems of harmonization. Mary Tao provides a timely look at data on the credit crisis. She provides starting points for when a researcher asks for data and/or statistical information on mortgages, credit cards, financial institutions, and other financial crisis-related materials. Both fee-based and free resources are included. Kristi Thompson then turns to the developing world as she looks at survey data on international development. She focuses on survey data for developing countries. She looks at some of the different groups that conduct surveys and make the microdata available. She provides helpful comparisons of the major sources, their sample sizes, geographic and temporal coverage, and access issues. The issue concludes with Michele Hayslett and Lynda Kellam focusing on the major changes that will occur with the 2010 United States Census and American Community Survey. A series of examples illustrate the challenges researchers face with comparability.

The authors and I hope that future conferences and issues of this journal will have a renewed emphasis on the actual data. There are many subject areas to explore and sharing makes us all more knowledgeable. Thanks to these authors for answering the call to arms to talk and write about data reference and to Kristi Thompson for serving as a speaker, author, and helping to edit this issue.

Bobray Bordelon, Princeton University Library
The Pedagogical Data Reference Interview

Abstract
This essay reflects on the reference interview on several levels. If we accept that academic reference in general has a pedagogical role, then it is necessary to adjust the standard model of the reference interview to reflect that value. Within the specific context of academic data reference, undergraduates as a group require more instruction during the reference interview because they are less prepared than graduate students and faculty to ask for what they need. A strict service model does not meet their needs. A successful model appropriately balances the tension between instruction and service. This balance will vary from one institution to another based on different user groups and institutional goals, with implications for resource allocation. Data librarians on the one hand and general and subject reference librarians on the other bring distinct sets of knowledge and experience to bear on the challenge of "assessing the user's need," which can be a rich point of collaboration and referral between them.

Keywords: Reference interview, bibliographic instruction, undergraduates

A Data Reference Pedagogy
As a social science data librarian at Carleton College, a small undergraduate campus with a distributed data support model, I consider myself to be situated somewhere between a general reference librarian and a data specialist, with a foot firmly planted in each professional culture. I have found the reference interview to be a rich site for examining and bridging the practices, values and expertise of the two specializations, especially as we have reflected on an appropriate level of research data support for our liberal arts campus. Because my work is at a confluence of many different levels of the organization, namely librarians, staff, students and faculty, I have come to realize that my reflections may be extrapolated to larger campuses where more individuals and organizations are involved in supplying data support across the campus.

In a 2002 article in the journal Portal: Libraries and the Academy, James Elmborg called for a vocabulary and the theoretical underpinnings for discussing reference work as a teaching activity. How does one teach well at the desk? Since what sets academic librarians apart from the rest of the profession is a recognized role in "participating in the teaching roles of their institutions," it is important to conceive of the reference interview not just as providing a service but also as engaging in teaching.

To achieve this end, Elmborg proposes that we draw from the theories of cognitive and social constructivism and the language of writing instructors to reposition our practice so that along with meeting the users' needs it is also our goal to help create self-sufficient learners. He argues, “It is the role of the teacher to identify where the student is in his or her development … and then provide guidance and collaboration in ways the student can internalize” (p. 462). Further, he expands upon the context of the reference interview, framing it not as an isolated interaction at a service counter, but rather as part of the socialization of new researchers into the community of scholarship: …if we accept the central notion that knowledge and meaning get negotiated in social contexts among members of a discourse community, then our responsibility within that community is to participate in discourse, to engage our students with meaningful talk about their research, to help them develop a language of inquiry that will allow them to articulate to themselves how to proceed with present and future research challenges (p. 461).

However, in the field in general, and in the materials from which reference librarians are taught, the reference interview is based on an assumption of identifying a user's existing need and providing them with relevant information – or perhaps a search strategy – suitable to that need. It does not reflect a goal of teaching, only of providing a service that meets an information need.

I find that this dynamic is further complicated when one considers the data reference interview. First, consider the way the reference interview is taught in library schools. The popular reference textbook by Bopp & Smith outlines the following five elements of the reference interview:
• Open the Interview

   Expressing openness and approachability
   Establish that you want to help

• Negotiate the question

   Learn the context of the patron:
   how much detail at what level is needed
   Use of open and closed questions and active listening

• Search for information

• Communicate the information to the user

• Close the interview

   Express willingness to provide further help
   Refer if necessary³

Bopp & Smith's articulation of a service ethic and emphasis on understanding the user's question are central to the way reference librarians conceive of their work. However, this model leaves out such important pedagogical elements as encouraging the student to participate in the process, explaining the judgements and decisions made to determine relevant information, determining the learning stage of the student, attempting to create a dynamic, student-centered conversation (Elmborg p. 460), and fostering exploration and independence as a researcher in the student.

One paragraph in particular highlights the disconnect between a broader reference model as articulated by Bopp & Smith and the pedagogical model espoused by Elmborg.

It should be obvious that information given to users must be at an appropriate intellectual level and free of jargon from either the field of librarianship or from any other field with which the user is unfamiliar. Users will not always say that the material presented to them is unclear or too difficult for them to master, so librarians must assess the user’s abilities during the search process to avoid giving the user the right information in the wrong package. (Bopp & Smith, p. 58)

This useful advice on meeting the patrons where they are and applying empathetic attention to their context and intellectual level still stops short. The goal here is to meet users where they are, but does not challenge them to become self-sufficient learners and searchers.

Elmborg pushes us to consider instead that the reference librarian should intentionally help students to understand and apply the language of research in their field, rather than protect them from it. I propose that helping students understand disciplinary terminology is especially called for in the case of working with undergraduates and data. It is not the librarian's job to match the information to students' level of understanding. Rather, it is the student's job, with the help of a librarian and their professors, to strive to raise their level of understanding to the terminology and concepts used with data and their documentation.

If we extend this critique to the data reference interview, a similar question arises. In the absence of a standard data reference textbook, I have constructed a composite of the advice I have encountered through fora such as IASSIST presentations and listserv discussions, the IQ, individual data librarians and the ICPSR class, "Providing Social Science Data Services: Strategies for Design and Operation," taught by Jim Jacobs, Chuck Humphrey, and Diane Geraci.

Establish what the patron needs:

• Statistics or data?
• What is the subject or topic?
• What is the unit of analysis?
• Geographic constraints or units?
• Time constraints (a range of years; monthly, quarterly or annually)?
• Do they need cross-sectional or longitudinal data? Time series?
• Opinion or demographic data?
• Financial or administrative data?

These questions that must be answered prior to finding data are highly helpful in establishing goals and a structure for a data reference interview, but similar to the Bopp & Smith model, this advice is focused on service. I still question: what would a data reference model be that blended pedagogy and service, imbuing the data reference interview with the goal to teach?

Undergraduates as Users of Data

One step toward addressing my question is to consider undergraduates as a distinct user group at a particular stage in their learning. We are careful to adapt teaching techniques to suit distinct learning stages in the classroom, so why not the same care at the reference desk?

Frequently, working with undergraduates can feel alien, a bad fit for the data reference interview. My own experience of applying the model by asking the questions referred to above has been to uncover more questions. Students with fluid, emerging research questions can not specify such things as geographic or time constraints. The constraints introduced by the availability of data will have a stronger impact on their research question than vice versa. Students new to quantitative research often do not fully understand the concept of unit of analysis, or the difference between cross-sectional and longitudinal data or, worse, the
difference between data and statistics.

I am sure many of us who have worked with undergraduates have experienced significant breakdowns of communication that catch one off-guard. Reference interactions are rife with problems of semantics and definitions (e.g., what exactly is meant by "raw data"), of process and of expectations (Fig. 1).

It is easy for these experiences to lead to underestimate undergraduates and to let them become caricatures in our minds. We might conclude that undergraduates are too lazy or impatient to pay sufficient attention to detail, that they lack the motivation, that they rarely if ever actually need real data, or that they’re overconfident and always leave their work for the last minute.

It is necessary, though, if we are going to take seriously the task of helping undergraduates learn, to reconceive of some of these patterns and use them as a way to understand where they are developmentally as researchers. These negative qualities are sometimes strategies that students have developed because they have been successful. Working to deadline has provided motivation and adrenaline to boost creativity. Abundant confidence gets them through the constant barrage of new ideas they get on a daily basis in classes. Everyday web searching rewards clicking links quickly instead of reading the page first. New researchers need to be given convincing cues that these strategies are not inherently wrong, but will no longer work in the context of long-term research projects.

Since undergraduates are emerging as researchers they do not have experience to guide them in areas more experienced researchers take for granted. They have not yet mastered the process of literature review and data search strategizing. They are still learning how to form researchable questions. The process of working with data is vague. They lack fluency in the language of quantitative research. In fact, when they come to us, they may have never encountered data "in the wild" before, having always been provided data with their assignments, and having no idea of the amount of decision-making, cleaning and arranging that goes into preparing data once it is found and accessed.

All of these considerations aside, undergraduates often simply are not doing the same thing as advanced researchers. They have different, but equally legitimate, motivations for looking for data. They may be required to find data on a topic not their own. They are likely to be working on short-term projects in which they are investing limited time and attention. They may be looking for data not as evidence per se but rather in order to demonstrate newly-learned statistics skills (i.e., their criteria are structural rather than topical such as a dataset with at least two continuous variables and a categorical variable). Also, format can unevenly influence data selection. Students familiar only with one statistical package may only be willing to look for data files formatted for that package.

Combining Discovery & Instruction
I do not have a formal proposal for building a data reference pedagogy, but I can share some of the strategies used at Carleton College. These strategies are inspired by the idea of a teaching reference interview and follow the general principle of combining discovery with teaching whenever possible.

Recognizing that most novices will not typically have a strategy for searching for data beyond Google, I try to model good search behavior, emphasizing process and continuously narrating my decisions. I use visualizations whenever possible to help speed understanding of complex ideas. I help students take notes by making very concrete suggestions and taking notes with them during the consultation, and suggest approaches for dealing with uncertainty.

On a broader, programmatic level, the Carleton reference librarians have tried to embed data instruction into information literacy instruction whenever possible. The quantitative reasoning initiative on our campus emphasizes teaching QR or numeracy across the

![Figure 2](image)

![Figure 3](image)
curriculum. This cross-curricular emphasis has the added advantage of creating opportunities for all librarians to provide instruction on finding quantitative information. All the librarians integrate quantitative sources into their online research guides so that data are presented as just one type of information among many. Students repeatedly receive the message that employing data as evidence in making arguments is critical to argument for all scholars, not just "quant geeks." For our web based finding aids, we have intentionally prioritized integration into course guides over creation of standalone data-specific materials, which could run the risk of becoming a data silo.

Perhaps most important, the other social science librarian, Danya Leebaw, and I have developed a data reference worksheet (see appendix) that prompts students and the librarian assisting them through a brainstorming process. The worksheet provides a place to jot down the suggested resources into not just a list of places to look, but within a structure that suggests a method. Filling out the form together demonstrates that the librarian doesn't just come up with ideas out of thin air, but rather out of a particular thought process necessitated by the information landscape of data production and publication.

Below are two examples of ways that I regularly prompt students to actually take notes because otherwise they often tend to click and click and go round in circles and get frustrated.

**Data Reference Outside the Data Center**

Although the majority of readers of this article will not share my context of working in a small liberal arts college, I believe there are important reasons for all data specialists to think about how our reference model serves undergraduates. We know that more and more undergraduates are using data. They're being introduced to data production and publication.

Below are two examples of ways that I regularly prompt students to actually take notes because otherwise they often tend to click and click and go round in circles and get frustrated.

**Data Reference Outside the Data Center**

Although the majority of readers of this article will not share my context of working in a small liberal arts college, I believe there are important reasons for all data specialists to think about how our reference model serves undergraduates. We know that more and more undergraduates are using data. They're being introduced to data production and publication.

Below are two examples of ways that I regularly prompt students to actually take notes because otherwise they often tend to click and click and go round in circles and get frustrated.

**Data Reference Outside the Data Center**

Although the majority of readers of this article will not share my context of working in a small liberal arts college, I believe there are important reasons for all data specialists to think about how our reference model serves undergraduates. We know that more and more undergraduates are using data. They're being introduced to data production and publication.

Below are two examples of ways that I regularly prompt students to actually take notes because otherwise they often tend to click and click and go round in circles and get frustrated.

For data specialists who are examining data support across the campus and are thinking about ways to train other librarians, it would help to frame that experience as a collaboration. To truly support undergraduates, it is necessary to combine the expertise of the data specialists with the expertise of subject and general reference librarians, namely their familiarity with the practice of teaching undergraduates and their knowledge of what is being taught in assignments. Data and non-data librarians can bring their complementary areas of expertise to bear on the puzzles of developing a pedagogical data reference model and helping undergraduates find and access data.

In conclusion, I have tried to show that the model of the data reference interview needs to reflect the tension of providing both service and instruction. It is not sufficient for the model to include only the elements of helping a patron to determine their need and then either getting them to the data or pointing them in the right direction. Rather, there is an essential third dimension introduced by the commitment to reference as a site of teaching. Especially in the case of undergraduates, the data reference interview should take a pedagogical approach, helping new researchers develop the capacity to do it on their own, creating self-sufficient learners and nascent social scientists.

**Acknowledgements**

This article would not have been possible without the insightful discussion and generous sharing of ideas by colleagues, Danya Leebaw and Heather Tompkins. I would also like to thank Paula Lackie and Carolyn Sanford who established the groundwork for and continue to provide support for data at Carleton and who provided helpful comments on early drafts of this paper.

**References**


**Notes**

1 Kristin Partlo, Reference and Instruction Librarian for Social Science and Data, Gould Library, Carleton College. One North College Street, Northfield, MN 55057. Email: kpartlo@carleton.edu.

2 Reference librarians have been concerned with defining professionally the practices and goals of reference service since Samuel Green’s article in the first issue of Library Journal in 1876. Specifically, the notion of the reference interview, or the preoccupation with the interaction and communication between the librarian, the patron, and his or her need for information, emerged and crystallized in the 134 years since.

3 See chapter three of Bopp & Smith, pp. 53-59, which is organized according to these five elements.
Sources for International Trade, Prices, Production, and Consumption

Abstract
Lengthy time-series of international trade, prices, production and consumption of commodities are becoming increasingly easy to find and use via the Internet. Major international organizations see openly available data sources as key elements of their service mission. At the same time, international organizations maintain use policies increasingly out of sync with the features that their technological innovations offer. In contrast, even though the United States government offers a wide range of international data with no use limits, user interfaces are often significantly less well developed. In addition to international and U.S. government sources, one entirely privately produced resource, LexisNexis Statistical Insight, is also included and notable for falling somewhere in between these two poles. This article describes the general data availability and stated use policies for the databases Comtrade, OECD iLibrary, UNCTAD Commodity Price Statistics, Minerals Yearbook, FAOSTAT, PSD Online, and Statistical Insight with respect to data on commodity trade, prices, production, and consumption.

Keywords: Trade; Commodities; Minerals; Agriculture;

Introduction
Thanks to the Internet, librarians and researchers have the best kind of problem these days: sifting through the vast quantities of increasingly freely available statistical time-series for just the statistics needed for their research. While this paper will not attempt to provide an exhaustive catalog of every database, indicator, country or year of coverage in each of the sources discussed, it will provide some broad outlines for each database’s trade, prices, production, and consumption data.

The databases covered are

• United Nations (UN) Comtrade
• Organisation for Economic Cooperation and Development’s OECD iLibrary
• United Nations Conference on Trade and Development (UNCTAD) Commodity Price

Statistics

• United States Geological Survey Minerals Yearbook
• Food and Agriculture Organization (FAO) FAOSTAT
• United States Department of Agriculture (USDA) Production, Supply and Distribution Online (PSD Online)
• LexisNexis Statistical Insight

Sources covered have at least five decades of data and one or more of the following indicators:

• bilateral commodity trade
• international prices
• international production
• international consumption

These and many more links can all be found in the author’s del.icio.us account at http://www.delicious.com/umdatalib.

This paper will focus on the parts of these databases that relate to bilateral commodity trade, prices, production and consumption, but each contains time-series for many additional topics. Where database providers make such features available, the paper will also cover copyright statements, citation tools, alert services, and visualization options.

Terminology note: while data and statistics are different things (specifically, data are what one uses to create statistics), in practice, the two terms are routinely used interchangeably. Most of the databases in this paper use “data” to describe their content and I will follow the same convention.

UN Comtrade
Comtrade contains data from 170 reporting countries and covers 1962 to the present as completely as possible. Comtrade data processing practices include conversion of commodity values from national currency to US dollars using reporter-supplied exchange rates, conversion of reported quantities to metric units, and conversion from the reporter-supplied classification into all versions of the Harmonized System (HS), all versions of the Standard International Trade Classification (SITC), and the Broad Economic Categories (BEC) classification. The United Nations says of its received data, “The data are permanently stored in the UN Comtrade database server.” The Comtrade “About” page (http://unstats.un.org/unsd/tradekb/Knowledgebase/What-is-UN-Comtrade) contains the statement “Browser Warning: UN Comtrade works best with Internet Explorer Web Browser. In other browsers, some pages might not work properly.” Mac users can not use Internet Explorer unless they have a Windows emulator, but they should have no trouble using Comtrade with Firefox. On the other hand, users of Google Chrome and Safari may find some limited functionality.

Anyone may view Comtrade data and anyone may download up to 50,000 records per user per session. If needed, a person or institution may subscribe for additional services including unlimited downloads.

Visualization options are very limited. Comtrade includes some automatically generated pie charts and a map based explorer. However, at the time of writing, the “Explorer with Map” remains broken in all platforms/browsers.

Expansive user access to Comtrade stands in contrast to the placed policies on re-use and dissemination of Comtrade data. There are two slightly different copyright statements plus a separate policy on use and re-dissemination. On the “About” page (http://unstats.un.org/unsd/tradekb/Knowledgebase/What-is-UN-Comtrade) is this statement: “The UNSD holds the copyright to COMTRADE data. All rights reserved. Users cannot disseminate Comtrade raw data without the express permission of the UNSD.” However, Comtrade provides no information on what “COMTRADE raw data” would be on this or any other pages. The footer of every Comtrade page contains this copyright statement: “No part of this material in either its printed or electronic format may be reproduced or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.” Comtrade has a third description of rights and permissions at (http://Comtrade.un.org/db/help/PolicyOnUseAndRedissemination_11Aug2010, pdf) which states that re-dissemination of Comtrade data within an institution does not require permission. So, a student including data or tables in a paper does not need to seek permission.

To recap, the UN says that using data in papers within a single institution or using a few tables or graphs is fine to do without seeking permission. Actually reproducing the data does require permission and depending on the quantity reproduced, a fee as well.

What is dissonant about this use policy is that one of the best things about Comtrade is the ease with which a user may access the data for precisely these kinds of uses. Because librarians ought to direct their users in the proper use of the resources to which they send their users, the author anticipates that in the future she will recommend that, regardless of whether users download data or not, when they cite their sources they should point to the table on the Comtrade website that they generated when downloading data for analysis. In pointing to Comtrade itself, the user is not re-disseminating data and therefore does not need to seek permission.

Because there are so many caveats that come with using data from so many different sources, Comtrade takes the unusual step of inserting a “ReadMe” screen before the display of a table. (http://Comtrade.un.org/db/help/uReadMeFirst.aspx) Much like click-wrap licenses for software purchased online, to get past the screen and to the data, users must check a box indicating that they have read the information provided describing data coverage and limitations before they can see the table that they have just constructed. The ReadMe page also appears when a user saves the URL to a particular table. Thus, if a table is cited in a paper and another user (a reviewer or professor for example) goes to see the table, then that user will also have to indicate that she has read the form.

Finally, for users who would like to develop their own interface to the Comtrade data, the United Nations provides an API. Some aspects of the Comtrade database are freely available via the API, but others are limited to institutions with a license to the database. Details are available at http://Comtrade.un.org/ws/.

OECD iLibrary
The OECD has two versions of its bilateral commodity trade database. SourceOECD is the older version and is in the process of being phased out. Its replacement is the OECD iLibrary, officially launched on 20/9/10.
The OECD is a 33-nation member organization established in 1961. As the OECD's publications serve the needs of member organizations, the OECD iLibrary has a different scope from Comtrade. The reporting countries are the member nations, although for the newest members little data may be found so far. OECD iLibrary also reports data from a few non-member entities including the EU-15 Extra EU\(^1\) and Taiwan (labeled as Chinese Taipei and for which data are hard to find). However, each member reports all of the countries with which it trades, so while the reporters are limited to the OECD membership, the database is still quite large. The overall time period covered is 1960-present. OECD iLibrary classifies its data according to the HS and SITC systems. OECD iLibrary may also include volume and value of trade for price calculation if provided by the reporting countries.

While some OECD iLibrary content is freely available, full access is fee-based. The cost can be quite high. However, the OECD iLibrary also includes all of the OECD's textual materials as well as their statistical databases, so subscribers get a lot of content for the cost.

Several very useful features in the OECD iLibrary make it even more worthwhile for institutions that can afford to subscribe. While providing citations and links to citation management software is pretty standard now in journal databases, similar features are much less common in statistical databases. OECD iLibrary provides citations to whole datasets, links to the major citation management software and will provide citations down to the table level for tables in textual publications. OECD iLibrary uses Digital Object Identifiers (DOI) as part of its classification structure. In doing so, they are adhering to emerging international standards for data citation. OECD iLibrary also provides the markup needed to use browser-based citation tools like Zotero. RSS feeds that alert subscribers to new data as added are another feature of use to researchers studying long-term trends.

The OECD also claims copyright to its materials and has a separate terms and conditions document at http://www.oecd-ilibrary.org/about/terms for subscribers to the OECD iLibrary. However, in each case acceptable uses are clearly defined and for most academic purposes, the OECD maintains a liberal use policy. They summarize acceptable use this way: “OECD encourages the use of its content (textual, statistical, and multimedia). You can copy, download or print content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. When available, you should use the "Cite As" tool within the OECD iLibrary. When the "Cite as" tool is not available, you should cite the Title of the material, © OECD, publication year (if available) and page number or URL (uniform resource locator) as applicable.”\(^5\)

Whereas Comtrade has the greatest volume of trade data, OECD iLibrary provides the most researcher-friendly interface by integrating citation into the database and providing alert services via RSS feeds on a database-by-database or topic-by-topic basis. The OECD iLibrary also has a more clearly worded statement on acceptable use of their content than does Comtrade.

**UNCTAD Commodity Price Statistics**

The UNCTAD Commodity Price Statistics at http://www.unctad.org/templates/Page.asp?intItemID=1889&lang=1 contains commodity prices and price indices for the period 1960-present. The focus here is not on comprehensive lists of commodities, but instead on raw materials, especially mineral commodities. Prices are in US dollars per ton unless indicated otherwise. The Commodity Price Statistics works in multiple browsers and operating systems. The web version of this database is freely available. The UNCTAD Commodity Price Statistics Terms and Conditions document generally "grants permission to Users to visit the Site and to download and copy the information, documents and materials (collectively, "Materials") from the Site for the User’s personal, non-commercial use".\(^6\)

For researchers looking for prices on minerals and raw materials, this is an excellent choice.

**USGS Minerals Yearbook**

The USGS Minerals Yearbook at http://minerals.usgs.gov/minerals/pubs/myb.html provides price and production data either by country with reports from the mid-90s-present or by mineral with statistics from 1900-present. It is freely available online and, as a United States government publication, is not subject to copyright. The web version of the Minerals Yearbook remains essentially like its print predecessor with a few format modifications. The Minerals Yearbook retains the organization of the print with volume one covering minerals, volume two covering domestic prices and production and volume three covering international prices and production. Most of the content is provided as PDF files or Microsoft Excel files. There is no single database to search across nor are there any built-in citation tools, or APIs. The USGS does provide an RSS feed of new publications on minerals that supports researchers with ongoing needs for data. Individual tables may include recommended citations. While the Minerals Yearbook is not very fancy, it is free, easy to find and use, regularly updated, not subject to usage restrictions of any kind, and backed up with print editions in case the site were to become unavailable.

**FAOSTAT**

The FAO database FAOSTAT contains a wealth of agricultural and resource data. It is the primary source for comprehensive (or as near to comprehensive as can be)
international

- production data (1961-present)
- trade (1961-present) - quantity, unit value, value
- prices (1966-present) - in local currency, standard local currency, US dollars
- food supply (1961-present) - quantity produced, producer price, area harvested, yield per hectare

FAOSTAT units are metric and values are expressed in US dollars unless noted otherwise. Production, price, and supply data can be converted to HS classification from FAO codes, but so far, trade data are only available by FAO code. FAOSTAT is free to use in its entirety as of July 2010, but for researchers in need of very large batch downloads, a free registration is required.

FAOSTAT says of its copyright “The data of the FAOSTAT database shown on this internet site are copyrighted by the Food and Agriculture Organization of the United Nations and are provided for your internal use only. They may not be re-disseminated in any form without written permission of the FAO Statistics Division.” 7 There are no separate Terms of Use. A slightly more positive expression of rights from the FAO is “All rights reserved. FAO encourages the reproduction and dissemination of material published on this Web site. Non-commercial uses will be authorized free of charge, upon request.” 4

FAOSTAT does not have an RSS feed for the whole database or parts thereof. No citation tools are built in either. Indeed, all FAO says about citation is “All references to FAOSTAT data will have to be mentioned with the proper URL and the access date.” 9 Unfortunately, FAOSTAT does not support persistent URLs for individual tables, so presumably the proper URL is the URL for the overall dataset such as TradeSTAT rather than US imports of shelled almonds for 2006 by importing country.

FAOSTAT targets data users with needs for large amounts of data, but it also provides simple graphical representations of trend data for users with casual data needs.

USDA Production, Supply and Distribution Online (PSD Online)
The USDA PSD Online database at http://www.fas.usda.gov/psdonline/psdhome.aspx contains trade, prices, supply, demand, and consumption data for agricultural commodities from 1960-present. The United States is the only reporter, but bilateral data between the US and other countries are available where they exist. Like nearly all US governmental databases, PSD Online is free. There are also no usage restrictions since this database is excluded from copyright like the Minerals Yearbook by virtue of being a US government work. It appears to work in all browsers and platforms. In Firefox, there are some display quirks, but they do not appear to affect the database’s function or its return of data. Users may create free accounts in the database and thereby save their queries for re-use. Users may also download datasets for each commodity for analysis in their preferred software. PSD Online also offers a data availability lookup at http://www.fas.usda.gov/psdonline/psdAvailability.aspx that, if the user finds it first, can save them the effort of constructing a query only to get no results.

LexisNexis Statistical Insight
LexisNexis Statistical Insight is very different from the other resources mentioned here. It is probably closest in many ways to a traditional journal database. It does contain some tables in Excel or *.gif, but its value lies in abstracts of print publications from an exceptionally wide range of sources. Even when the data are not directly available in Statistical Insight, it still helps users to figure out what kinds of data might exist and the most likely sources for that data. LexisNexis claims that Statistical Insight includes data from 1801-present. Statistical Insight includes all kinds of statistical publications including those with prices, production, and consumption. At the University of Minnesota, Statistical Insight helps to fill a gap in the Libraries collection of market research. Market research is very, very expensive and the University of Minnesota is unfortunately not able to purchase market research reports directly. However, Statistical Insight can identify sources of data from trade associations that can cover similar territory to market research. Thus, it helps us serve researchers that otherwise would have to look elsewhere for such data.

Statistical Insight is less expensive than market research reports, but it is still very expensive, especially if an institution licenses access to additional modules that incorporate more data into the database. No part of it is freely available. While there are limits on claims to copyright of factual material in US law, license agreements can overcome those limits. LexisNexis licenses are typically fairly restrictive. On the other hand, to the extent that Statistical Insight serves to identify data sources rather than as a source of data itself, rights issues seem a little less important in this case.

Conclusion
Not only are more data sources appearing online, but they are doing so with the explicit goal of making more data available to the greatest number of users. Data sources that may have launched with short time series find that users always want more and the data source producers strive to meet those desires as openly as possible. Many sources provide researcher-friendly features such as update alerts via RSS feed and built-in citation tools. At the same time, data source rights statements do seem to be out of sync with the spirit of expanded open access. In an environment in which data source producers enable features like mass
downloads, the expectation that permission must be acquired on a case-by-case basis in order to use those downloads seems odd. However, these are minor quibbles. The increased quantity of data, increased ease of access and improved interfaces far outweigh any drawbacks identified in this paper.

Notes:
1 Amy West, Data Services Librarian. 10 Wilson Library, University of Minnesota, 309 - 19th Avenue South, Minneapolis, MN 55455-0414. (612) 625-6368. westx045@umn.edu.

2 Morteo, R., 2010. Question about the copyright statement for UN Comtrade.


Data Reference in Depth: Sources of International Labour Data

Abstract
Most countries provide access to national labour data on the Internet, but finding it can be a frustrating exercise, especially if the country in question does not provide the information in a language understood by the researcher. Foreign labour data are also often sought when attempting to make international comparisons of particular labour statistics. This article reviews Internet-accessible multi-country compilations of labour data that provide access in multiple languages, with attention given to those that permit international comparisons to be made.

Keywords: labour data, labor data, comparative data, international data

While data and statistics on labour and employment for wide variety of countries are now readily available on the internet, finding statistics for a particular country and/or on a particular parameter for multiple countries can be a challenge. Even more frustrating can be trying to find statistics that are comparable across countries. In this article, I will describe a number of sites that provide easy access to both national labour statistics and international and comparative labour statistics.

International Labour Organization (ILO)
<http://www.ilo.org>

The ILO should be the first stop for labour data beyond one’s own borders. It is the hub for labour statistics within the UN system. The ILO is "... devoted to advancing opportunities for women and men to obtain decent and productive work in conditions of freedom, equity, security and human dignity. Its main aims are to promote rights at work, encourage decent employment opportunities, enhance social protection and strengthen dialogue in handling work-related issues."12

The ILO provides, in English, French and Spanish, official core labour statistics and estimates for over 200 countries since 1969. Besides the data, it also provides considerable metadata, definitions and methodological descriptions of main national statistical sources. The data provided use mostly national definitions, except in limited cases; the difficulties this poses will be discussed at the end of this paper.

On the ILO home page, a link to ‘Statistics and databases’ is given on the right side. This page shows a list of databases, both of statistics and literature, with brief descriptions. The ones that are primarily data-related are listed under the ‘Statistics’ subheading:

1. LABORSTA - database of labour statistics
2. Key Indicators of the Labour Market (KILM)
3. Statistical Information and Monitoring Programme on Child Labour (IPEC-SIMPOC)
4. Labour Force Surveys

Plus, separate from these four sites, but devoted primarily to data that are more current, is

5. ILO Global Job Crisis Observatory

LABORSTA
<http://laborsta.ilo.org/>

LABORSTA covers official core labour statistics and estimates for over 200 countries since 1969. Also provides methodological descriptions of main national statistical sources. [3] 3 At the time of writing, there were 224 entries for countries, of which some are subnational areas included for historical reasons (e.g., the four entries for Germany: Germany as it is now, the former FRG and GDR, and the 5 new Länder plus what was East Berlin), or overseas territories (e.g., St. Pierre & Miquelon, an overseas territory of France). The first page offers the user a choice of statistics for 11 general topics, in most cases available in both annual and monthly frequencies:

- Total and Economically Active Population
- Employment
- Unemployment
Following the link to 'employment', for example, leads to six links to data sources, some of which go to other databases. The hover text for each link lets the user know how far back the data go, at least as an upper limit. There is also a ‘by topic’ link on the left-hand menu bar that leads to a page that lists all the available topics. The user can also see what is available by country, or by publication (for those familiar with the range of topics covered by ILO Yearbook of Labour Statistics, ILO Bulletin of Labour Statistics, and ILO October Inquiry). Ultimately, the data can either be viewed online, or downloaded to Excel. The online version has links to relevant metadata embedded in it; the Excel spreadsheet has limited metadata. The user needs to look at the ‘view data’ page to have a clear idea of what the data represent; if nothing else, the spreadsheet should contain links to the necessary metadata.

On the LABORSTA home page (and on the left menu on all subsequent pages) is a set of metadata links:

- **Definitions** - of terminology used (e.g., economically active population, employment, etc.) In each case, an essay is provided with the definitions used, potential sources of data on the given concept, plus alerts as to potential problems with interpretation;

- **Classifications** - detailed information on the various international standard classification schemes used:
  - International Classification by Status in Employment
  - International Standard Classification of Education
  - International Standard Classification of Occupations
  - International Standard Industrial Classification of all Economic Activities
  - System of National Accounts 1993

- **Sources and methods** - information on the scope of the statistics, their definitions and the methods used by the national statistical services in establishing the data published. This is the full text of the ten-volume series Sources and Methods: Labour Statistics. Perusal of these metadata links is critical to a full understanding of the data and their limitations for making comparisons.

The LABORSTA homepage also contains links to short term indicators of the labour market; these are drawn from official national statistical sources, based on national definitions, and are not seasonally adjusted. They contain monthly data for the past twelve months, quarterly data for the past five quarters, and annual data for the past three years. The data are available by topic or by country; in addition, complete country profiles and selected series by country are available in PDF or XLS formats.

**Key Indicators of the Labour Market (KILM)**

Available in print, online and as a standalone software package, the KILM has been published every two years since 1999. Currently in its sixth edition, the KILM

- is a comprehensive database of country-level data on 20 key indicators of the labour market from 1980 to the latest available year;

- is a source of the latest ILO world and regional estimates of employment and unemployment indicators. A training tool on development and use of labour market indicators;

- highlights of current labour market trends;

- provides analyses of key issues in the labour market.

The online version, known as KILMnet (still in beta) offers 32 tables in six groupings:

- Participation in the world of work (3 tables)
- Employment indicators (11 tables)
- Unemployment indicators (6 tables)
- Educational attainment (2 tables)
- Wages and labour costs (7 tables)
- Performance and poverty indicators (3 tables)

The interface is similar in some ways to that of the World
Development Indicators, except that all the data parameter selections can be made on the same screen.

The differences between the KILM and LABORSTA are addressed in the document "Guide to understanding the KILM". Basically, LABORSTA (and its print equivalent, the Yearbook of Labour Statistics) are the best source of nationally-reported labour statistics, whereas KILM supplements these data from other sources when those other sources are considered more accurate or more complete, and offer better international comparability. The KILM is not restricted to using national data as reported; it makes efforts to use indicator series that are more comparable across time and geographies. The KILM offers three comparable or harmonized series: labour force participation rates, employment-to-population ratios, and the inactivity rate. Other series have been made as comparable as possible; anomalies in definition and methodologies are clearly indicated in the table notes.

In other words, when national data are desired, the user should start with LABORSTA; when comparisons between countries are to be made, KILM should be the starting point.

Statistical Information and Monitoring Programme on Child Labour (IPEC-SIMPOC)

This is the statistical arm of the International Programme for the Elimination of Child Labour (IPEC). It claims to offer:

- Specific questionnaires for child labour surveys
- Manuals and training kits on how to carry out child labour data collection in households, schools and at the workplace
- Guidance on how to properly process and analyse the collected information
- Micro datasets and survey reports from around the world
- Research on critical statistical issues
- Regular trend reports

Microdata sets for 30 countries (as of 2010.09.26) are available under the left menu option "Surveys"; the data are often available for multiple years, and are provided in ASCII, SPSS or Stata, along with metadata.

Labour Force Surveys

This page lists links to the websites for the labour force surveys (defined as "a standard household-based survey of work-related statistics") of all the countries in the ILO. It does not provide access to the microdata, and is inconsistent as to what it does link. Using examples from countries the author of this paper is currently familiar with,

- in the case of Canada, it links to the press release associated with the most current release of the Labour Force Survey; it does not mention the Survey of Labour and Income Dynamics;
- for the United States, it links to the "Current Labor Statistics" page from the Monthly Labor Review; it provides access to labour data from both the Current Population Survey and Current Employment Statistics;
- for Australia, it simply provides a link to the "Statistics by Topic: Labour" page on the Australian Bureau of Statistics website, with no direct access to the Household, Income and Labour Dynamics in Australia (HILDA) survey.

In each case, however, the ILO does provide access to a page describing the methodology of the main labour force survey of each country. Also, the user gets a link to the country’s relevant website - -this in itself can be valuable, as it enables one to follow up with the country itself on any questions, as well as potentially providing access to reports on data unavailable from the ILO.

Global Statistics on the Labour Market
website, this site offers "the latest national data for indicators which have been selected for their ability to reflect recent and short term changes." Data are available by topic or country, and are not seasonally adjusted or otherwise altered by the ILO. Updates of indicators and associated publications are usually monthly, but can be more or less frequent.

OTHER DATA SITES
The ILO is not the only site from which one can get compiled labour data from a variety of countries. The following section of this paper describes some others.

International Labor Comparisons (US)
<http://www.bls.gov/data/#international>
This site provides labour data for selected countries using a variety of indicators, all of which are adjusted to U.S. concepts unless otherwise noted. Data series are presented as: a group of most requested series; either a one-screen or multi-screen data search process; a series of static tables (in HTML, PDF or XLS formats). Some series, such as the supplementary tables comparing manufacturing productivity and unit labour cost trends, go back to 1950; most, however, start sometime in the 1990s.

OECD Labour Statistics
<http://www.oecd.org/>
This section of the article is divided into two parts: for OECD iLibrary (formerly SourceOECD) subscribers and those who do not subscribe.

Non-subscribers to OECD iLibrary
Finding labour and employment statistics on the OECD site can be a little confusing, as there are three different pages that provide related information.

(1) From the above address, you can reach the OECD Statistics Portal
<http://www.oecd.org/statsportal/0,3352,2825_293564_1_1_1_1_1,00.html>
This has a section on Labour:
<http://www.oecd.org/topicstatsportal/0,3398,2825_495670_1_1_1_1_1,00.html>
The Labour portion of their portal has two sections: "Labour statistics" (13 indicators) and "Unemployment Statistics" (6 indicators), along with links to various reports, definitions and other metadata. The links for each indicator can take one to a publicly accessible (i.e., to non-subscribers) portion of OECD.Stat, or one of the OECD's many other labour- and employment-related databases. Much of the data are inaccessible to non-subscribers.

(2) The OECD also offers access to their content by general subject area; via this route, there is a section on "Employment"
<http://www.oecd.org/topic/0,3373,2649_37457_1_1_1_1_37457,00.html>
Clicking on "Statistics" in the menu bar on this page leads one to a long list of discrete statistical publications on employment, many of which, but not all, are also under the "Unemployment statistics" heading on the OECD Statistics Portal page on Labour.

(3) The OECD also has a Directorate for Employment, Labour and Social Affairs
<http://www.oecd.org/departement/0,3355,2649_33729_1_1_1_1_37457,00.html>
The link to "Statistics" within this section leads one to a page that is actually more helpful, in many ways, that the other two: a set of links to relevant pages of statistics are given at the top of the page. The relevant one is to Employment <http://www.oecd.org/els/employment/data>

- The Employment database offers current statistics for international comparisons and trends over time.
- Key Employment Statistics has summary tables for OECD countries with indicators on labour market outcomes and policies and how they compare with the OECD average.

In both cases, links to sources and relevant metadata are readily accessible.

While the written reports may be unique to the OECD, the data are derived from national labour force surveys. Given this, you may find it easier to use LABORSTA.

Subscribers to OECD iLibrary
<http://www.oecd-ilibrary.org/>
For subscribers, the quest is much simpler. From the OECD iLibrary home page, follow the link to "Statistics" at the top of the page, and then find "OECD Employment and Labour Market Statistics" in the list of available databases. The abstract states that this database includes a range of annual labour market statistics and indicators from 1960 broken down by sex and age as well as information about part-time and short-time workers, job tenure, hours worked, unemployment duration, trade union, employment protection legislation, minimum wages, labour market programmes for OECD countries and non-member economies.

The data are more accessible, and many preconfigured tables are available. Related OECD publications are clearly
indicated. A nice feature of the new iLibrary is a link, for each available dataset, to a page that provides a citation and links to download the citation information to a variety of citation managers.

OTHER RELATED SITES
A source that can be generally useful in finding or interpreting labour data is How to find labour statistics 11. The focus is naturally on the ILO’s resources, but it does offer a selection of resources available to the public from other international organizations and partner institutions. Resources, including a variety of metadata, are listed by topic as well as either globally or regionally.

PROBLEMS OF COMPARING DATA FROM DIFFERENT COUNTRIES
As alluded to earlier, problems can occur when attempting to compare data across geographies. In the ILO definitions of terminology, one frequently encounters phrases such as “National definitions of [desired criterion] may differ from the recommended international standard definition” or “National practices vary between countries” or “The comparability of the data is hampered by the differences between countries and even within a country”. A concrete example: in Canada, a part-time worker is one who usually works less than 30 hours per week; in Australia and the U.S, the criterion is less than 35 hours or more per week, and in the EU, it is whatever the individual being asked considers part-time work. In addition, in terms of classifying one’s status as employed vs. employer, most countries classify managers and directors of incorporated enterprises as employees, while in some others they are classified as employers. One wonders if comparability is possible at all.

Efforts are being made to create labour data that can be compared across national boundaries. As previously mentioned, the KILM database incorporates efforts to create harmonized variables that are comparable between countries. Other efforts include the Cross-National Equivalent File (CNEF project based at Cornell University12, which includes “equivalently defined variables” for data from the U.K., Australia, Korea, U.S., Switzerland, Canada and Germany. The Comparative Perspectives Database, an offshoot of the Gender & Work Database 13, is a project that is attempting to create a variety of cross-tabulations on the subject of precarious employment using data from the U.S., Canada, Australia and the countries of the European Union. These efforts, as well as others, will make the task of comparability much simpler, albeit not as comprehensively as one might like.

Making variables have equivalent definitions is not necessarily an ideal solution either. The legal and regulatory frameworks surrounding labour and employment are dependent on national definitions; creating equivalent definitions removes the data from its national context, thereby, in some ways, making the data less relevant to the people and environment it represents. It is an issue one has to keep in mind, depending on the reason for comparing the data.

The important thing to remember, when attempting to compare statistics on similar parameters from different countries, is to examine the metadata, especially the definitions, closely to make sure apples are being compared to apples, not apples to pineapples, as it were. Just because something is called the same thing does not mean it is the same thing. Ultimately, it is the responsibility of the user of the data/statistics to ensure that their use is appropriate, but the data professional can, at least, alert her/him to the issues.

REFERENCES
1 Author contact information: Walter W. Giesbrecht, 203E Scott Library, York University, Toronto, ON, Canada M3J 1P3 (email: walterg@yorku.ca)


Financial Crisis Data Resources: A Brief Guide

The U.S. housing boom was brought to a halt by the subprime mortgage crisis in 2007. As the housing market hit a bad patch, it affected other areas as well. The situation had evolved into the credit crisis by 2008. Banks’ balance sheets took a few hits, leading to a liquidity crisis.

The United States was not the only country hit hard by events; Greece and Iceland, among others, saw their share of pain and suffering. Ireland is the latest country in the news, with the government stepping in to save two major Irish banks.

Here’s an example of how closely intertwined the fortunes of global companies are. The troubles of an American company, Lehman Brothers, meant that it was unable to repay the money due to its creditors, including Depfa Bank, an Irish firm. As a result, Depfa was unable to pay its own creditors, eventually causing major problems for its parent company, Hypo Real Estate Group, based in Germany.

A lot has been written about the financial crisis and a lot more are yet to come. Listed here are just a few resources to use as starting points.

**Banking Statistics**

**CRSP-FRB Link**

- allows researchers to compare companies over time, taking into account mergers and acquisitions. The dataset matches regulatory entity codes to CRSP permcos for publicly traded banks and bank holding companies from January 1990 to December 2007.

http://www.newyorkfed.org/research/banking_research/datasets.html

**Assets and Liabilities of Commercial Banks in the United States - H.8**

- this weekly release provides an estimated aggregate balance sheet for all commercial banks in the United States including U.S. branches and agencies of foreign banks.

http://www.federalreserve.gov/releases/h8/about.htm

Reports of Condition and Income (Call Reports)

- this database allows one to obtain financial and structural information for most FDIC-insured institutions and compare banks with its peer group. The data goes as far back as March 31, 2001 and can be downloaded as (PDF), Semicolon Delimited Format (SDF), or eXtensible Business Reporting Language (XBRL) format. Bulk data is also available for all reporters in the form of Tab Delimited format or (XBRL) format.

https://cdr.ffiec.gov/public/

**QUARTERLY SUMMARY OF BANKING STATISTICS**

- quarterly synopsis of balance sheet and income statement developments for all U.S. commercial banks. The banks are broken into two categories: those held by the 8 largest domestic bank holding companies (BHCs) - JPMorgan Chase, Bank of America, Citigroup, Wells Fargo, PNC Financial Services Group, US Bancorp, Bank of New York Mellon, and Suntrust - and all other commercial banks.

http://www.newyorkfed.org/research/banking_research/quarterly_summary.html

Quarterly Summary of Banking Statistics: 3rd Quarter 2009
Fee-based resources include SNL Financial and Bloomberg (WDCI - writedowns and credit losses versus capital raised )

**Housing Statistics**

*Federal Housing Finance Agency (FHFA)*
- housing market indicators


*U.S. Bureau of the Census*
- datasets for decennial census housing files 1940-2000, housing vacancies and homeownerships, housing starts and building permits.

http://www.census.gov/cgi-bin/briefroom/BriefRm

*U.S. Department of Housing and Urban Development*
- access to the original datasets including the American Housing Survey, as well as microdata from research initiatives on topics such as housing discrimination, the HUD-insured multifamily housing stock, and the public housing population.

http://www.huduser.org/portal/datasets/pdrdatas.html

**ICPSR**
- offers American Housing Survey datasets, Census of Population and Housing as well as the underlying data cited in research publications.

http://www.icpsr.umich.edu/icpsrweb/ICPSR/

**Mortgage Statistics**

*Mortgage Market Statistical Annual*  
- print and CD (XLS format). Available data include monthly average mortgage rates calculated for different mortgage terms, monthly new home sales inventories, house prices, mortgage markets, total securities issuance volume, and more.

*U.S. Credit Conditions: Mortgages*  
- contains maps of delinquent mortgages around the country by state and county and uses publicly available data on mortgage delinquencies and foreclosures.

http://data.newyorkfed.org/creditconditions/

Other useful sources include fee-based sources such as First American CoreLogic LoanPerformance, ABSNet, Equifax, Haver Analytics, and Bloomberg.

**Mortgage-backed Securities Statistics (MBS)**

*Fannie Mae*
- monthly reporting data on Fannie Mae’s MBS


Other financial sources include fee-based resources such as First American CoreLogic LoanPerformance, ABSNet, Equifax, Haver Analytics, and Bloomberg.

**Freddie Mac**
- a variety of data relating to mortgage securities


**Inside Mortgage Finance**
- publisher of Mortgage Market Statistical Annual and various newsletters containing the latest data on MBS.

**Credit Cards Statistics**

Major providers of credit card statistics tend to be fee-based sources such as Equifax,

Bloomberg, and SNL Financial.

**SNL Financial**
- information on the major credit card issuers (American Express, Bank of America, JPMorgan Chase, Citigroup, Discover)
  - Credit Card Delinquency Data - 30+ day delinquencies; chart; spreadsheet of underlying data

**Federal Reserve Liquidity Facilities**

The following programs were authorized by the Board of Governors of the Federal Reserve System under Section 13(3) of the Federal Reserve Act to provide credit and liquidity during a time of financial stress.

- [Commercial Paper Funding Facility (CPFF)](http://www.newyorkfed.org/markets/epff.html)
- [Primary Dealer Credit Facility (PDCF)](http://www.newyorkfed.org/markets/pdcf.html)
- [Term Asset-Backed Securities Loan Facility (TALF)](http://www.newyorkfed.org/markets/talf.html)
- [Term Securities Lending Facility (TSLF)](http://www.newyorkfed.org/markets/tslf.html)
- [Forms of Fed Lending chart](http://www.newyorkfed.org/markets/Forms_of_Fed_Lending.pdf)

**Other Financial Data**

- Credit Ratings (of companies and countries)
  - Moody's, Standard & Poor's, Fitch
Moody's Corporate Default and Recovery Rates, 1920-2009

- www.moodys.com/corporate_default_and_recovery_rates_02_10
- Free registration for some reports; a subscription is needed to acquire most datasets

Bloomberg, SNL Financial $ - RATT - historical ratings trends 2000-2010

- Credit Risk – data on the financial stability of nations
  - Moody's, Standard & Poor's, Fitch, Bloomberg $
  - Financial Soundness Indicators (FSIs)-IMF source that measures the capital adequacy of deposit takers. Meaning does a certain country’s financial institutions have enough capital at hand to withstand shocks to their balance sheets?

Further Reading


Notes
1 Mary Tao, Federal Reserve Bank of New York, Contact: mary.tao@ny.frb.org. The views expressed are those of the author and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.
Data in Development: an Overview of Microdata on Developing Countries

Abstract
Finding quality microdata on developing countries can seem problematic as their national infrastructures may not support large-scale surveys. In fact a variety of organizations are collecting and distributing data, though the types of data and reasons for collection often differ from those in the most developed countries. Much of the data is collected by groups involved in, or interested in researching, the field of international development. This paper provides an introduction to the different groups involved with collecting data on developing countries and to the data they collect, including population health and welfare surveys, program assessments, finance, and opinion data.

Keywords: Developing Countries; Development; International; Foreign Aid; Survey data

As all data librarians know, a good datum can be hard to find. When the data being sought deal with countries generally referred to as developing, less developed, or low and middle income, finding good data can seem like even more of a challenge. The less developed countries form an extremely heterogeneous group, and from a data perspective, they are united primarily by what they lack. In developed countries, the librarian can usually expect national governments and related institutions such as central banks to collect demographic and business data, and national archives or major research institutions to archive surveys. In less developed countries, national governments and institutions may not collect much detailed data beyond the census and other material strictly required for internal use, or may not have developed the infrastructure to process and distribute the data they do have for public consumption. And yet, "there has been a spectacular increase in the availability and quality of data from developing countries in recent years." (Bureau for Research in Economic Analysis of Development). Data is being collected and distributed, but in many cases the groups collecting the data, and the purpose behind the collection, are specific to the field of international development. This paper looks at data on developing countries and the field of international development which concerns itself with them. For the sake of keeping this overview to a manageable length and also of highlighting the sources will be of use to the most people, the discussion will be limited to microdata available from sources that do not require a paid subscription.

Development is a process, not a permanent state. Developing countries are not united only by their lack of the economic advantages that distinguished the most developed countries. They are united by the fact that they are subject to development. "Developing countries is an international practice. The essence of this practice is the mobilization and allocation of resources, and the design of institutions, to transform national economies and societies, in an orderly way, from a state and status of being less developed to one of being more developed." (Gore) This practice of developing countries has developed itself into a highly diverse international field, with its own sets of standards and practices and key players large and small. People and institutions active in the practice of development collect data, both to assist with their own operations and for fundraising purposes, to "demonstrate that they can perform effectively and are accountable for their actions." (Degomme and Guha-Sapir)

An initial challenge is simply to define the countries under consideration. "There is no established convention for the designation of 'developed' and 'developing' countries or areas in the United Nations system." However, "in common practice, Japan in Asia, Canada and the United States in northern America, Australia and New Zealand in Oceania, and Europe are considered 'developed' regions or areas,"(United Nations Statistics Division) leaving most of the planet still under development. A more fine-grained categorization is the United Nations’ Human Development Index ranking, which divides member states into Very High, High, Medium and Low human development. The World Bank uses Gross National Income to divide countries into High Income, Upper Middle, Lower Middle and Low Income. The UN measure takes into account a broader range of factors, including life expectancy and education as well as income. For purposes of this paper, and where relevant, I will use the groupings from the UN 2009 Human Development Report to distinguish level of development.

The different groups active in collecting data on developing
countries include national governments, intergovernmental organizations (IGO's), non-governmental organizations (NGO's) and other – this last being a catchall category containing such groups as academics and for-profit private sector firms. The bulk of the data considered in this paper comes from intergovernmental and nongovernmental organizations. In the interest of brevity, data from individual governments is considered here only insofar as it appears in other compilations. Intergovernmental organizations include such familiar large and well-funded organizations as the World Bank, the United Nations and the World Health Organization. These are valuable sources both because they conduct surveys and collect data, and because they serve as compilers and standardizers of country-level macroeconomic data. Intergovernmental organizations tend to conduct large-scale, nationally representative surveys that are standardized across a number of countries. Non-governmental organizations are more heterogeneous, and include larger and better funded organizations such as Demographic and Health Surveys and the International Food Policy Research Group along with a myriad of smaller and less known organizations. The data they collect is similarly heterogeneous, but non-governmental organizations are more likely to provide subnational surveys targeted towards a particular context or issue. “NGO surveys aim at assessing a local situation for needs and programming... (while) UN surveys tend to be large scale snap shots of a situation that serves as a point of reference” (Degomme and Guha-Sapir) Surveys conducted by academics, or collaborating groups of academics, range from the large-scale, standardized, cross-national World Values Survey to the small, very specific local assessments available from MIT’s Poverty Action Lab.

Survey Catalogs
There are two databases that compile surveys on developing countries, the International Household Survey Network catalog, and the Bureau for Research and Economic Analysis of Development (BREAD) / McArthur survey database. The International Household Survey Network catalog, which is maintained by the World Bank Data Group, contains 4147 surveys at the time of writing. Surveys include major population welfare surveys, censuses, firm-level economic surveys, and others, and the catalog can be browsed by country or survey series. The other database, which is available from the BREAD website, is smaller, claiming to hold only about 500 surveys, and lists as its focus surveys on poverty and health. It can be searched by survey location and module. Both of these databases provide information and links to access survey microdata where available.

A third database, the Complex Emergency Database (CE-DAT), has information about specialized surveys carried out by non-governmental organizations in smaller populations such as refugee camps, often under crisis conditions. At the time of writing it contained 2713 surveys. While microdata is not available from the database, the citations will assist in finding the appropriate contact person to inquire about availability, and the database includes summary statistics for the indicators and has an interface for mapping and charting them.

Major Population Health and Welfare Surveys
There are three major, long-running international data collection programs that focus on surveying developing countries: the Demographic and Health Surveys (DHS), the World Bank's Living Standards Measurement Study (LMS) surveys, and UNICEF’s Multiple Indicator Cluster Surveys (MICS). The World Health Organization’s World Health Surveys cover both developing and developed countries and are also worth a look. Each has advantages and drawbacks in terms of geographic coverage and topical focus. I have limited this comparison to these four series because each provides relatively uniform data collection across a number of countries, enabling cross-national comparisons. I have excluded survey programs such as the Food Policy Research Institute household surveys, which are not uniform enough to consider as a group, and the World Fertility Surveys, which were last conducted in the 1980’s and are now quite dated. The data from all these surveys is largely available to researchers as microdata, though individual surveys may be unavailable and registration or application may be required.

The Demographic and Health Surveys (DHS) program is the largest of the major population welfare survey series. The DHS program is primarily funded by the U.S. Agency for International Development, and as such has a broadly defined focus on aid, development and policy improvement. They have conducted over 240 surveys in about 90 High, Medium and Low human development index countries in Africa, Asia, Latin America, the Middle East and Europe. The earliest surveys date to the mid-1980’s, and multiple waves have been done in many countries. The primary surveys focus on fertility, family planning, maternal and child health, gender issues, and nutrition, and the surveys also cover household and respondent characteristics including education and school attendance, employment, and income and family wealth, making them useful for a range of analyses. DHS also conducts special modules on key topics such as AIDS and Malaria, and scaled-down surveys called the Key Indicators Surveys that are used to assess smaller sub-national populations that may be targeted by special initiatives. The focus is on women and children and men are often excluded.

The Living Standards Measurements Study surveys are conducted by the World Bank, and perhaps not surprisingly, it has excellent coverage of consumption and income. Around 90 surveys have been completed in over 40 countries between 1985 and the present. Most of the surveys cover countries with High or Medium Development, with only one (Malawi) classified as Low,
and another, Iraq, that is currently not classified. While the focus is on economic measures ranging from income and employment to debt and purchasing behavior, there are a number of health, demographic and social measures. Some of the surveys are available directly from the World Bank web site, others are distributed by the government of the country in which they were conducted.

The World Health Surveys were done in 2002, and follow-up studies such as the WHO Study on Global Ageing and Adult Health are being conducted. Microdata has been released for all the countries in the original surveys and is available upon signing a data use agreement. There are detailed questions on health, mortality, health insurance and use of health services along with some basic demographic and wealth variables. Employment is included, but not income. A different questionnaire is used for countries in the highest income group.

The Multiple Indicator Cluster Surveys, done by UNICEF, are currently on their fourth wave, and have been done every five years since 1995. The MICS are particularly notable for their coverage of the poorest countries; in the 2005 wave, 11 of the 55 countries they surveyed were in the low development group. They also surveyed some sub-national populations. However, the surveys are limited in their usefulness for general analysis due to the selection of variables available. The MICS were developed specifically to track progress on the World Fit for Children Plan of Action and the Millenium Development Goals, and cover indicators relating to child health and welfare, women’s reproductive health, and some basic household variables, but do not include the usual economic and demographic variables such as income and employment.

### Comparing the Major Population Health and Welfare Survey Series

<table>
<thead>
<tr>
<th>Survey</th>
<th>Time Period</th>
<th>Focus</th>
<th>Demographics</th>
<th>Labour and Income variables</th>
<th>Countries Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>DHS</td>
<td>1985-present</td>
<td>Health, particularly reproductive</td>
<td>Good</td>
<td>Some</td>
<td>90+ low, medium and high development</td>
</tr>
<tr>
<td>LSMS</td>
<td>1985-present</td>
<td>Consumption and income</td>
<td>Good</td>
<td>Good</td>
<td>40+ mostly high and medium development</td>
</tr>
<tr>
<td>WHS</td>
<td>2002</td>
<td>Health and health systems</td>
<td>Good</td>
<td>Some, no income</td>
<td>70 ranging from low to very high development</td>
</tr>
<tr>
<td>MIC</td>
<td>Every 5 years from 1995</td>
<td>Child health and welfare, reproductive health</td>
<td>Limited, mostly household head</td>
<td>Limited</td>
<td>65+ low and medium development</td>
</tr>
</tbody>
</table>

### Data for Assessing, Modeling and Targeting Development

While much of the data collected on developing countries is used to provide evidence to assess developmental progress, data focusing on the evaluation of specific approaches is rarer. The Abdul Latif Jameel Poverty Action Lab was formed at MIT as a network of professors around the world who use Randomized Evaluations to answer questions on poverty alleviation. “What makes J-PAL's work innovative is that such randomized studies haven't typically been used in evaluating poverty-alleviation programs, or even in the wider field of economics.” (Standish) They offer training on randomized evaluation and maintain a database that currently holds information about 234 randomized assessments, with microdata available for 13 of them, on topics ranging from textbook provision to microfinance.

The International Food Policy Research Institute (IFPRI) is another source for random evaluations and has released data for studies such as the Comparing Food versus Cash for Education Program, along with their more standard household welfare datasets.

The International Food Policy Research Institute also collects data to construct Social Accounting Matrices, which are currently available for about 30 countries. Social Accounting Matrices are used as the basis for various economic models that can be used to estimate the effects that different policy or aid approaches will have before
Actually carrying them out.

**Finance**

The World Bank claims to provide “the world's most comprehensive company-level data in emerging markets and developing economies” (World Bank Group) and I would not attempt to dispute that claim. They have conducted surveys in over 125 countries, including survey projects such as the World Business Environment Surveys, as well as more specialised projects such as the recent three-round Financial Crisis Surveys and the Management, Organisation and Innovation survey work. Firm-level microdata is available to researchers at no cost from these surveys, in contrast to the International Monetary Fund, which also conducts surveys but only releases macrodata, and some of that only to paying customers. The World Bank’s Enterprise Survey Portal also allows users to construct charts online.

A couple of regional sources for financial microdata are the Economic Research Forum, focusing on the Middle East and North Africa, and Oxford University’s Centre for the Study of African Economies, which focuses on Sub-Saharan Africa. The Economic Research Forum has released data on micro and small enterprises in Egypt, Lebanon, Morocco and Turkey, as well as an Egypt labour market panel study. The methodology documents for the micro and small enterprises data provides interesting insights into some of the difficulties in collecting and analysing data on developing economies; for example, in Turkey the researchers were unable to weight the rural sample because no more authoritative data on rural enterprises existed. The Centre for the Study of African Economies has comparative cross-national firm-level datasets primarily of manufacturing firms, as well as a couple of household panel surveys, and some more specialized datasets from working papers.

**Opinions, Attitudes, and Values**

The data collections discussed so far have been factual, consisting of theoretically objective measures of demographic and economic variables that can be used to implement and assess development programs. Opinion and values surveys may not appear to have the same practical, development-oriented application that most of the data discussed so far have. Development is done, in the end, for people, to improve their lives as they actually experience them, not merely to increase GNI or some other measure. Opinion data can “delve deeper than, for example, official poverty data, by reporting on people’s experiences in obtaining basic human needs and their own perceptions of whether or not they feel poor.” (Corporacion Latinobarometro) In addition, development happens in a political as well as an economic context. It is ideally done in cooperation with and with the support of governments and the people in the area being developed.

Two of the largest and best-known cross-national opinion surveys are the World Values Surveys and the Global Barometers. The World Values Surveys are an international collaboration among academics who are attempting to survey the “basic values and beliefs of the publics of more than 80 societies.” (World Values Survey) Five waves have been completed, conducted between 1981 and 2008, with a sixth currently in progress. “Since each national group funded its own survey, its first wave was largely limited to relatively developed societies.” (World Values Survey) By the second wave, the researchers behind the project had decided that “it was important to include societies across the entire range of development, from low income societies to rich societies,” (World Values Survey) and additional researchers and sources of funding were found.

The Global Barometers are done more modularly, with separate Afrobarometer, Arab Barometer, Asian Barometer, East Asian Barometer, and Latinobarometro surveys, and a Eurasia Barometer under development. (Another one, America’s Barometer, covers some of the small Latin American nations and is available for subscription through the Latin American Publib Opinion Project (LAPOP).) While the Global Barometers were inspired by Eurobarometer, there is no formal association. The Barometers are designed to be a comparative survey of attitudes and values toward politics, power, reform, democracy and citizens’ political actions. They are repeated at approximately three year intervals, and include a core module of questions asked across regions, plus additional region-specific questions. “Whereas the WVS addresses deep-seated, semi-permanent cultural values, the GB is concerned with tracking emerging political and economic attitudes, which are often subject to rapid change.” (Corporacion Latinobarometro)

The Pew Global Attitudes Project is conducted by the Pew Research Centre, a U.S. based non-governmental organization. It is a series of opinion surveys, each covering anywhere between five and 50 countries, that have been conducted between 2001 and the present. Key areas of interest include “attitudes toward the U.S. and American foreign policy, globalization, terrorism, and democracy.” (Pew Research Center) While there is a decided focus on American foreign policy (for example, the data formed a basis for the book *America Against the World: How We Are Different and Why We Are Disliked*), the data also includes questions of broader interest such as global reactions to issues in the news.

The multinational surveys mentioned above are generally too large and slow to implement to track responses to local issues under rapidly changing circumstances. Local polls, whether done by local media organizations, academics, or others are more likely to track these highly situational opinions, but the decentralised nature of local polling as well as language barriers make them particularly difficult to
access. Worldpublicopinion.org, which is managed by the Program on International Policy Attitudes at the University of Maryland, compiles and analyzes some of these opinion polls, and they also conduct their own local polls. Their own studies are performed by a network of research centres in 25 countries, and many of their datasets are available to download.

**Conclusion**

This overview has only touched on some of the sources of data available on the developing world. The choice to focus on microdata means that macroeconomic and many finance sources have been neglected; limiting myself to freely available sources means that resources in archives such as ICPSR or subscription services such as Polling the Nations were left out; most smaller projects covering only one or a few countries are included only insofar as they appear in one of the survey databanks. Still, it should be clear by now that a great variety of data on developing countries and development is available, whether it is collected by practitioners of development trying to improve their outcomes, researchers studying developing countries and various types of comparative social science, or other interested parties from both within and outside the developing world. While much of this data is gathered to serve relatively narrow purposes, collectively it can provide the researcher with a window, however imperfect, into the minds and lives of a large and often unheard portion of the world's population.

**Appendix: list of data sources in the order they were mentioned.**

**Survey catalogs**

BREAD/Mcarthur/CCPR Survey Database: http://ipl.econ.duke.edu:8080/survey/


**Major Population Health and Welfare Surveys**

Demographic and Health Surveys: http://www.measuredhs.com/


Multiple Indicator Cluster Surveys: http://www.childinfo.org/mics.html

**Data for Assessing, Modeling and Targeting Development**


International Food Policy Research Institute (IFPRI) Surveys: http://www.ifpri.org/datasets

**Finance**

World Bank Enterprise Surveys: http://www.enterprisesurveys.org/


Oxford University’s Centre for the Study of African Economies: http://www.csae.ox.ac.uk/

**Opinions, Attitudes, and Values**

World Values Surveys: http://www.worldvaluessurvey.org/

Global Barometer: http://www.globalbarometer.net/

Worldpublicopinion.org: http://www.worldpublicopinion.org/


**References**


Available online: http://www.globalenvision.org/2009/12/03/rigor-science-now-economics-too


Endnotes

1 Contact: Kristi Thompson, Data Librarian, Leddy Library, University of Windsor, Windsor, Ontario, N9B 3P4, Phone: 1 519 253 3000 x3858 Email: kathomps@uwindsor.ca

This paper is based on the presentation Data in Development given by the author, IASSIST 2010.


4 See http://www.globalbarometer.net/background.htm

5 See http://pewglobal.org/americaagainsttheworld/
The American Community Survey: Benefits and Challenges

Abstract
In the United States' decennial census, all persons living in the US are asked to fill out a short form asking basic questions such as age, race, and number of people living in a housing unit. In addition to the short form, starting in 1960 a sample of housing units were asked to fill out a long form with both the basic demographic questions plus questions about socioeconomic topics, such as education, income, housing characteristics and more. Although this sample survey is not constitutionally mandated, it serves an essential function for policy-makers and planners. In 2010 the United States will conduct its constitutionally mandated census of the population, but a major change will occur. The long form will no longer be distributed and in its place will be the American Community Survey (ACS). This article discusses the development of the survey and its benefits and challenges. The ACS will provide researchers and policymakers more timely information of the characteristics of areas.

Nevertheless, there are still some questions and concerns about how to use the data and challenges for the implementation of the survey.

Keywords: Population, Demographics, Census, Socioeconomics

Every ten years the United States is required by its Constitution to conduct a census of the population. Article 1, Section 2 of the Constitution of the United States maintains that:

Representatives and direct taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers...The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.²

In the United States' decennial census, all persons living in the US are asked to fill out a short form asking basic questions such as age, race, and number of people living in a housing unit. These demographic data are used for the apportionment of Congressional seats. In addition to the short form, starting in 1960 a sample of housing units were asked to fill out a long form with both the basic demographic questions plus questions about socioeconomic topics, such as education, income, housing characteristics and more. Although this sample survey is not constitutionally mandated, it serves an essential function for policy-makers and planners. In 2010 the United States will conduct its constitutionally mandated census of the population, but a major change will occur. The long form will no longer be distributed and in its place will be the American Community Survey (ACS).

This article will discuss the development of the survey and its benefits and challenges. The ACS will provide researchers and policymakers more timely information of the characteristics of areas. Nevertheless, there are still some questions and concerns about how to use the data and challenges for the implementation of the survey.

Development and Design of the American Community Survey
Efforts to create the American Community Survey began in 1996 when the survey was launched at four test sites. With the 2000 Census, a test form of the ACS was conducted as the Census 2000 Supplement Survey (C2SS) and was launched in 1,200 counties. The purpose was to test “the feasibility of collecting ACS statistics in a decennial census year.” (Herman, 2008) Full nationwide implementation of the ACS began in 2005 except for group quarters data which began in 2006.

The ACS is a selfenumeration survey with questionnaires sent by mail to chosen survey households. Enumerators conduct follow up telephone calls and visits to addresses that have not mailed in their questionnaires. Approximately 250,000 addresses receive a questionnaire each month totaling about 3 million households each year, resulting in a sample size of approximately one in eight households. The costs of conducting a monthly survey prevent an increase in the sample size to match the Census long form sample size. Because of the smaller sample size and because the sample is accumulated progressively over time, the release of the data is tiered based on the size of geographic areas (Mather, Rivers, Jacobsen, 2005). Hence estimates for geographies
with larger populations (more than 65,000 people) can be calculated on the sample accumulated within just one year, but estimates for geographies with smaller populations must wait to be calculated until three years or five years of data have been collected.

ACS data prior to 2005 are available for geographies with 250,000 people or more and are considered test data. In 2006, the Census Bureau published ACS data collected in 2005 for geographies with at least 65,000 people and data for these large geographies with over 65,000 people will be available on an annual basis. For a geographic area with between 20,000 and 65,000 people, three year estimates first became available in December 2008 using data collected from 2005 to 2007. In 2009, the three year estimates for 2007 through 2009 were released. For a geographic area with fewer than 20,000 people, a five year estimate will be required. The first five year estimates for the period 2005-2009 will begin to be released in late 2010.

**Figure 1: ACS Release Dates**

As with the long-form sample in the decennial census, the ACS is sample survey data and will have margins of error and confidence intervals. The Census Bureau maintains that the estimates are within the range of a 90% confidence interval. For example, in the 2005-2007 three-year estimate the population of Greensboro, NC is 237,423 with the margin of error of +/-2,958. This statement tells us that the Census Bureau is 90% certain that the population of Greensboro is between 234,465 and 240,381.

Another defining characteristic of the ACS is the collection of data over a period of time. This is in direct contrast to the data collection for the decennial census. Whereas the decennial census has a reference point of April 1 for determining residency, the ACS’s reference period for residency varies depending on the month in which the specific household receives the questionnaire. This has numerous effects on understanding data related to specific reference periods especially employment, income, and school enrollment. Although the ACS replaces the long form in the conduct of the Census, these differences related to reference periods affect the comparability of ACS data to decennial long-form data.

**Benefits**

The Census Bureau developed the ACS in response to users’ demands for more timely data. Although the Supreme Court determined that only 100% data can be used for apportionment of Congressional seats, planners and policy makers needed more frequent data releases to make better decisions and determine whether programs were successful and working as intended. Thus, the immediate benefit of ACS data is that it is collected every year and released the following year. Businesses, government agencies at all levels and the public will no longer have to wait ten years to find out how the country and local communities have grown and whether planning and public policy is meeting people's needs. No longer will they have to wait two to three years after the decennial census is taken for data on income, education and housing characteristics to be released. Moreover, because the survey is run every year, the data provide a way to track rapidly changing community trends and the opportunity to change data collection to respond to current events, including natural disasters like hurricanes and forest fires, and economic crises.

The Census Bureau believes that, despite having a smaller sample size, the ACS will actually provide more accurate data than the decennial long form for two reasons. First, because the ACS is being run constantly, a professional staff has been hired on a permanent basis to work in local areas. Instead of having to hire a huge number of temporary, non-professional staff who have to be trained in a very short period of time, this permanent staff will gain deeper experience and local knowledge over time that will improve data collection. For instance, issues such as reaching non-English speaking groups will become easier to address since these long-term staff will either be members of those communities themselves or able to develop relationships with leaders in those communities. This is also the reason the Census Bureau cites for the ACS saving money over the decennial long form, that it is more cost effective to maintain a smaller collection and processing staff throughout the decade than to hire and train a much larger number of workers once a decade.

Second, the non-response follow-up procedures for ACS are more extensive than those of the decennial long form, including telephone contacts as well as in-person visits (U.S. Census Bureau 2008b, 82). As an example, “a comparison between ACS and Census 2000 data for the Bronx showed that while the Census 2000 had a higher initial mail response rate than the ACS, it was less effective than the ACS during
follow-up phases, when information is collected from nonrespondents” (U.S. Census Bureau 2008c, 8).

The ACS is also capable of producing some data that the decennial census was not. The decennial long form asked people to answer questions based on their "usual residence" defined as "the place where the person lives and sleeps most of the time" (U.S. Census Bureau 2007, C-1). If someone received a form at an address where they did not live most of the time, the form would indicate they should only fill out a form for their usual address. Consequently the decennial census had no mechanism for counting temporary populations like people who live in Florida in the winter months or people who live in the northern states in the summer. The ACS, however, counts people at their "current residence," defined as "everyone who is currently living or staying at a sample address...except for those staying there for...less than two consecutive months" (U.S. Census Bureau 2009a, 6-1). Moreover, the counting goes on year-round instead of on one day, so the ACS is able to account for temporary residents regardless of season. Areas that have significant seasonal migrant worker populations will also notice higher ACS counts versus the decennial long form figures since the year-round data collection will better account for such groups. Some researchers have stated concern about the comparability of school enrollment data since the ACS will collect data in the summer months when children are not in school (Gage 2006, 247). However, at least as far back as 2005, the questionnaires have been worded to ask whether children have been enrolled "in the last three months." Consequently, the time of year when a respondent receives the survey should not matter for this variable.

Overall, researchers are beginning to appreciate the advantages in the ACS data over the decennial long form data. In a 2006 study, Gage found, after graphing multiple variables for two California counties:

In most cases, even when statistical tests identified differences [from decennial long form data] as significant, the ACS data generally appeared useful and usable. Simply observing a statistically significant difference provides no guidance as to which data are better.….For practical purposes it appears that most of the ACS data could, on an annual basis, be used in place of the census data and should provide a more current measurement, especially as the census count ages and remains static throughout the decade (247).

However, challenges still abound, particularly for new users.

**Challenges and How to Meet Them**

The degree of difference between the methodology of the ACS and the decennial long form survey results in a number of notable challenges for data users who want to do time series analysis. Essentially, the two surveys are not comparable. The simple cost of running the survey every year results in a significant compromise: the sample size of the ACS is decidedly smaller than that of the decennial long form. Griffin and Waite from the Census Bureau argue that the “estimates of sampling error for the five-year ACS estimates will be about one-third higher than those from decennial census estimates” (2006, 216), but they maintain that “this is acceptable given the reduction of bias due to timeliness and the potential for reductions in nonsampling errors because of factors such as the use of automated instruments and experienced interviewers” (2006, 216).

It must be emphasized that the Census Bureau's goal with the ACS is not to produce a population count but rather to produce an estimate of the characteristics of the population. That is, this data will be less useful than the Census long form for noting the absolute numbers of the population but very useful for studying trends over time. This is an important distinction because, with its smaller

![Figure 2: Ranking Table – Percent of People with a Disability (Note option on left to View as a Chart)](image-url)
sample size, the ACS is not very useful for pin-pointing exact numbers.

The smaller sample size is also the reason the Census Bureau is publishing the confidence intervals (CIs) for each estimate with the ACS, to demonstrate the accuracy of each figure. While the numbers which appeared in the decennial long form were also estimates, the sample size was sufficient that the Bureau didn't feel the need to emphasize the CIs. Unfortunately one of the results of this was users came to see the long form numbers as actual counts rather than the calculated figures they really were. With the ACS, particularly for smaller geographies and smaller groups of population (by race or income, etc.), the CIs can be quite large despite targeted over-sampling to off-set this problem. Hence, it's more important for the Bureau to highlight them and explain what they mean. The state ranking tables offer a visual display of the CIs that is very helpful once one understands how to interpret them (See Figure 2).

Figure 2 shows a ranking table for percent of people in each state with a disability. You will see on the left side of the page there is an option to View as a chart.

Figure 3 shows the chart. The red dots represent the estimates while the blue lines bracketing each dot represent the CIs. While the estimate will always be shown in these charts as the center of the CI, technically the definition of a confidence interval is that the true value may be anywhere within the CI. So when a user looks at the chart, anywhere the estimates' CIs overlap, technically the actual values for those states might be the same. So the order of those rankings might be considered ties, or even be reversed. This illustrates why the data should be used for tracing trends, not as absolute numbers.

Another effect of the very different methodology is that many ACS variables are not comparable to decennial long-form ones, even when they have the same name. Novice users will almost certainly be tempted to make direct comparisons without realizing they are trying to compare apples and oranges instead of apples to apples. For example, while the decennial census is taken on a single date, the ACS is a rolling survey, with responses collected every month of the year. Consequently rather than try to ask about respondents' income "last year" as the decennial does, respondents will be asked to provide their income during the twelve months prior to the date they receive the survey. This is likely to be a challenge for respondents to even answer. On the decennial census date, April 1st, most U.S. respondents are working on or have finished their Federal income tax returns (due April 15th) and can easily cite their previous calendar year's income. Citing the previous twelve months' income for the ACS, however, will require some figuring, especially if the given twelve-month period encompasses a change in rate of pay, or commission income that varies from month to month. Users of the final estimates are likely to think "last year's income" is essentially the same as the "income of the last twelve months" without realizing the significance of the different reference periods involved. The ACS's reference period is also an example of why this data should be used for trend analysis rather than point-in-time exact figures.
have estimates based on the average of five years of data. Because the distinction is based on population totals, large cities will have single-year estimates while small towns will have three- or five-year estimates—and the three different levels are not comparable to each other. Instead, in addition to the one-year estimates, the Census Bureau is making available averaged-year estimates for larger geographies that should be used for comparisons to smaller geographies.

For example, to compare the state of North Carolina and the city of Charlotte, one can use one-year estimates for both since the population of each exceeds 65,000 people. However if one were researching the city of Kannapolis, its population was 36,699 in the 2000 decennial census. Because this falls between 65,000 and 20,000 people, the ACS will only provide estimates based on the average of three years of data in order to have enough respondents in the sample to create accurate estimates for the size place it is. In this case, to compare Kannapolis with the state, one would need to use North Carolina's three-year averaged data instead of its one-year estimate. Likewise to compare North Carolina and Mount Airy, a town of 8,460 in the decennial census, one would need to use the state's five-year averaged data since Mount Airy will only have ACS estimates based on the average of five years of data.

Data for the smallest geographies, all those with less than 20,000 people (including all Census tracts and block groups), have not yet been released. The American Community Survey began full-scale data production in 2005 (with the exception of Group Quarters data which was added in 2006), so until it has had five full years of data collection, the pool of respondents will not be large enough to create the estimates for the smallest geographies. With an extra year for processing time, the Census Bureau will not release five-year averaged estimates until close to the end of the 2010 calendar year. Consequently for a while yet data users will be frustrated when trying to find data on small places or rural areas. However, this issue will disappear entirely once the first five-year estimates are released since five-year estimates will be available every year thereafter. Knowing which estimate to use for larger geographies and how to explain the use of different figures in context when writing a grant proposal, for instance, will be a particularly difficult issue for novice users.

Another issue related to sample size is the suppression of data. In the decennial census, the Census Bureau employs thresholds below which data for very specific occupations or population groups will not be published in order to protect confidentiality. Because of the Bureau's confidence in the sample size based on five years' worth of data, the ACS does not employ such thresholds. Instead, staff tests for the statistical reliability of the one- and three-year estimates and suppress tables when at least 50 percent of the included estimates (that is, cells within the table) fail the Coefficient of Variation test. The Bureau states that the five-year estimates will not be tested at all since the sample size based on five years of data will ensure viable estimates. An example of this might be detailed race breakdowns in a rural state, especially ones that tend to be more homogenous racially. For example, Montana might.

FIGURE 4: Base table, B02003. RACE - Universe: TOTAL POPULATION
be more likely to be suppressed for this variable than North Carolina. One way the Census Bureau handles this is by the production of Base versus Compressed tables. Base tables provide all the detail users are used to seeing in the decennial long form data. But for a table that is likely to be suppressed because more than half of its cells fail the statistical test, the Bureau may produce a Compressed, or C, table for the same subject. Figures 4 and 5 demonstrate the difference between the two.

Figure 4 is the Base (or B) table for Race from the 2008 1-Year Estimates. It was necessary to run this report for the country as a whole—even the most populous states were suppressed. This is understandable when one considers how detailed the categories are for "Population of two races," with fifteen different race combinations including, for instance, one for those respondents who indicated they were both American Indian/Alaska Native and Native Hawaiian/Other Pacific Islander. In Figure 5, the C table for the same variable, you can see that the Two or More Race categories have been severely compressed to the four most commonly chosen categories and one titled "All other two race combinations." Here it was possible to generate data for states at both ends of the population spectrum as well as for the nation. Users familiar with the P(opulation) and H(ousing) tables of the decennial data will easily translate to the ACS system of labeling tables B(ase) or C(ompressed) in the title, as noted in these Figures.

Another method the Bureau recommends5 for ameliorating estimates with very large margins of error (MOEs) is to combine several geographies or several variable categories. The method is straightforward: one simply sums the geographies or categories to create a larger “sample.” Of course, it can only be used for straight summed data like population, race, sex, etc.; it cannot be used with calculated values such as medians. Then to calculate the new MOE for this new “estimate,” one squares each original MOE

<table>
<thead>
<tr>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Year Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-Year Estimates (2005-2007)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25.9</td>
</tr>
<tr>
<td>3-Year Estimates (2007-2009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24.8</td>
</tr>
<tr>
<td>3-Year Estimates (2008-2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-Year Estimates (2006-2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Estimates for a Geography with More Than 65,000 People
and adds them together, then takes the square root of that sum, or $\sqrt{\text{MOE}^2 + \text{MOE}^2 + \text{MOE}^2}$. Of course, this method needs to be used with some care. Reliable data will not result from combining geographically distant geographies. Geographies at the same summary level (e.g., tracts combined with tracts or counties combined with counties) that border one another and have similar characteristics to the one under examination are to be strongly preferred.

To return to the issue of estimates based on the averaged data of several years, we can better understand the effect of averaging several years of data by considering two examples presented by Deborah Griffin and her colleagues at the State Data Center/Business and Industry Data Center Annual National Training Conference in 2004. In the first example, values are steadily increasing over time—the percentage of foreign-born population might be such a variable. Tables 1 through 3 show hypothetical values for such a variable. (For simplicity's sake in this example, the values across the geographies are the same, although this would probably seldom be true in actuality.)

Consider that an average is a measure of the middle. Consequently when one averages data to create an estimate, the estimate will tend more toward the middle of the figures averaged.

Figure 6 shows how the averaged estimates' values will follow the trend of the annual estimates in cases of steady
increases but will tend to lag slightly behind it. This would also be the case with steady decreases in values. However, what happens when the trend fluctuates? Figure 7 shows such an example, a hypothetical view of home ownership rates.

Here you can see that the averaged estimates tend toward the middle of the varying values, describing a trend that smooths the highs and lows to a flatter line. This is perhaps the biggest disadvantage of the ACS methodology and it is not so much an issue of accuracy as of precision. Researchers familiar with the decennial long-form data will miss that survey's ability to provide (essentially) one-year estimates for all geographies. Some users of the ACS data have even indicated that one should not use overlapping estimates; in other words if one uses a 2004-2006 three-year estimate, it would be better to wait for the 2007-2009 data with which to compare the same geography. However, the Census Bureau would again assert that the ACS data is best used to understand trends, and that the 2005-2007 three-year estimate will provide an update to the 2004-2006 one, even if much of the pool of respondents remains the same.

Geographic boundaries of the most recent year in multi-year averaged estimates apply. To do this, ACS staff re-create the earlier years' estimates with the current year's geographic boundaries in order to include respondents for the new geography for all years of the average. Also, dollar values for earlier years of an average are inflation-adjusted to the most recent year. (Griffin, et al., 2004) In a year that a small-sized geography crosses the threshold to the next size (i.e., from less than 20,000 to between 20,000 and 65,000) it will begin to have three-year averaged estimates produced as well as five-year estimates. Likewise, in a year when a medium-sized geography crosses the threshold to the large size (i.e., from between 20,000 and 65,000 to over 65,000) it will begin to have single-year estimates produced as well as three- and five-year. The reverse is also true. If a geography loses population and drops below the threshold, it will lose the estimates of the larger category—a place dropping below 65,000 would lose the single-year estimates and a place dropping below 20,000 would lose the three-year estimates.

The Future
How specifically the surveys are able to describe a community has always been at the forefront of the challenges the Census Bureau faces. Protecting confidentiality is of paramount importance, punishable by fines and imprisonment. Yet the American public demands the smallest level of geography possible for both political and economic planning reasons. Officials at the Bureau recognize that striving for this level of detail is costly. At a hearing of the Congressional Joint Economic Committee, former Census Bureau Directors Louis Kincannon and Kenneth Prewitt both testified that even for the decennial census, data at the block level is unnecessary for the purposes of redistricting and dropping the smaller geographic levels would significantly cut costs. (2009, timestamp 98:18) With follow-up questioning, Prewitt stated that data at the census tract level would provide sufficient detail (2009, timestamp 98:18).

Conclusions
The best preparation for understanding a community’s ACS figures is to know the community very well. Local knowledge will help researchers identify when the ACS data are incorrect or insufficient. Where researchers are not familiar with local communities, they must carefully attend to the MOEs and decide when the data are sufficient to the research purpose at hand and when they are not. For novice users, guidance on using the ACS is critical. Librarians need to be on-hand in academic and public libraries to assist users with both navigating the American FactFinder interface and understanding the ACS data.

The Census Bureau is fully aware of how difficult the ACS is to use, particularly for novices, and it works constantly to make tools available to assist with it, including extensive technical documentation, guidance on making comparisons between different editions of ACS data, and Compass handbooks customized for different audiences. There is also an e-tutorial to assist novice ACS data users. This suite of tools is available on the ACS’s How to Use the Data web site at http://www.census.gov/acs. While learning to use the ACS will take some effort, it is imperative to do so. The long form on the decennial census will not return and the ACS will remain the best data available.

References


Notes:
1 Authors are Michele Hayslett, University of North Carolina at Chapel Hill, michele_hayslett@unc.edu, and Lynda Kellam, University of North Carolina at Greensboro, lmkellam@uncg.edu.


3 Personal communication (by telephone and email) with Bob Coats, North Carolina's Liaison to the Governor for the Census on August 6, 2009. The Bureau has described decennial census-taking as the "largest peace-time mobilization of personnel in U.S. history" (U.S. Census Bureau 2009b, 2).


5 Personal communication with Kelly Karres at the North Carolina State Data Center annual meeting, Raleigh, NC, August 23, 2010.

6 http://www.census.gov/acs/www/guidance_for_data_users/e_tutorial/
IASSIST 2011 - CALL FOR PAPERS

The 37th International Association for Social Science Information Services and Technology (IASSIST) annual conference will be hosted by Simon Fraser University and University of British Columbia and will be held in Vancouver, Canada, May 31 - June 3, 2011.

The theme of this year's conference is Data Science Professionals: a Global Community of Sharing. Social science benefits from professional practices that enable sharing of data, information, and knowledge with a global community. This theme is intended to stimulate discussions about ways in which sharing data, information, and knowledge can contribute to research and to professional practices that enable scientific progress. Submissions are encouraged that offer improvements for creating, documenting, submitting, describing, disseminating, and preserving scientific research data.

We seek submissions on the theme outlined above, and encourage conference participants to propose papers and sessions that would be of interest to themselves and other attendees. Below is a sample of possible topics that may be considered: * open data and the development of knowledge communities

- Data sharing, access and management in the future
- Identifying and reducing barriers to data sharing - Issues of confidentiality in sharing
- Sharing professional data science skills, knowledge, & techniques within & across discipline
- Citation of research data and persistent identifiers
- Metadata facilitating data sharing
- Emerging research infrastructures and data sharing
- New data partnerships in knowledge communities
- Sharing resources and data through social networks
- Identifying user needs and customizing data services to meet the needs
- The evolving data librarian profession
- Data science practices that support global use and understanding of research data
- Open (linked) data and digital repositories
- Preservation for sharing, recovering data for contemporary use

Proposals on other topics related to the conference theme will be considered too. Papers will be selected from a wide range of subjects to ensure a broad balance of topics.

The Program Committee welcomes proposals for:
- Individual presentations (typically 15-20 minutes)
- Sessions, which could take a variety of formats (e.g. a set of three or four presentations, a discussion panel, a discussion with the audience, etc.)
- Posters/demonstrations for the poster session
- Workshops (pre-conference workshops that blend lecture and hands-on instruction). [Note: A separate call for workshops is forthcoming.]
Again this year we would be interested in receiving submissions for presentations in formats successfully introduced in last year's conference, in particular
- Pecha Kucha (a presentation of 20 slides shown for 20 seconds each, with a heavy emphasis on visual content).
- Round table discussions (as these are likely to have limited spaces, an explanation of how the discussion will be shared with the wider group should form part of the proposal).

Session formats are not limited to these ideas and session organizers are welcome to suggest other formats. Proposals for complete sessions should list the organizer or moderator and possible participants; the session organizer will be responsible for securing both session participants and a chair.

All submissions should include the proposed title and an abstract no longer than 200 words. Longer abstracts will be returned to be shortened before being considered. Please note that all presenters are required to register and pay the registration fee for the conference; registration for individual days will be available.

A web form for submission of proposals will be available on the conference web site on October 18, 2010.

Deadline for submission: November 29, 2010.

Please note that the Conference Program Committee may not be able to accept all proposals.

Conference papers will be considered for articles in the IASSIST Quarterly. This applies to individual papers as well as selections of papers from sessions that could form special issues of the IASSIST Quarterly.

For more information about the conference, including travel and accommodation, see the conference web site at:
http://www.rdl.sfu.ca/IASSIST/

Online conference registration is scheduled to open in early February, 2011.

Make plans to come to Vancouver for IASSIST 2011: 31 May - 3 June 2011!

Questions may be sent to the Program Planning Co-Chairs, Bob Downs, Ernie Boyko and Tuomas J. Ålateräi at iassist2011@gmail.com
The International Association for Social Science Information Service and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the IASSIST QUARTERLY. They also benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST.

Membership fees are:
- Regular Membership: $50.00 per calendar year.
- Student Membership: $25.00 per calendar year.
- Institutional Subscription: $75.00 per calendar year (includes one volume of the Quarterly)

Please make checks payable, in US funds, to IASSIST and Mail to:

IASSIST, Assistant Treasurer US
Shelagh Mackay
University of California
Davis, CA, USA
95616-8617

I would like to become a member of IASSIST. Please see my choice below:
Options for payment in Canadian Dollars and by Major Credit Card are available. See the following web site for details:
http://datalib.library.ualberta.ca/membership/membership.html
- $50 (US) Regular Member
- $25 Student Member
- $75 Subscription (payment must be made in US$)
- List me in the membership directory
- Add me to the IASSIST listserv