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 welcome. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors
The QUARTERLY is published four times per year. Authors are encouraged to submit papers as word processing files. 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. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". Journal of the American Society for Information Science 30(2):77-82, March 1979.

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

Production
Fadi Dagher
423 West 120th St., #107
New York, NY, 10027
USA
Email:fd2102@columbia.edu.

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

Title: Newsletter - International Association for Social Science Information Service and Technology
ISSN - United States: 0739-1137 © 1997 by IASSIST. All rights reserved.

FEATURES

4 Editor's Notes

5 Mashing Maps
Rachael E. Barlow

11 Reward and Punishment Mechanism for Research Data Sharing
Jinfang Niu

16 An Overview of Publicly Available Quantitative Cross-National Crime Data
P. Stamatel
Editor’s Notes

Welcome to the fourth issue of the IASSIST Quarterly, vol. 30.

The 34th IASSIST annual conference will take place at Stanford University, Palo Alto, California, USA, May 27-30, 2008 with the theme “Technology of Data: Collection, Communication, Access and Preservation”. The conference will examine the role of technology and tools in various aspects of the data life cycle. See the call for papers in this issue of the IQ.

At the IASSIST 2007 conference in Montreal - the best conference ever - one of the presentations in the session “Data Services mash-ups: Maps, Research and Everything!” was Rachael Barlow from Trinity College in Hartford, CT presenting “Maps that Mash: Daring, Dangerous, or Dumb?”. The article is now called “Mashing Maps”. Barlow introduced a class at Trinity College - a small, liberal-arts college - to the facilities of Google maps mash-ups; the idea of the class was to attempt to do something meaningful with data for people living in the local community Hartford, Connecticut. The students in the class created several mash-ups, for example one group of students mapped food resources in Hartford, everything from community gardens to grocery stores to food pantries. Some maps are shown in the article. As a sociologist Rachael Barlow investigated the effects of the Trinity College class, where the production of the mash-ups made a cooperative connection between the faculty and students inside the college’s walls and the general community members outside those walls. She found that the connection improved the image of the college and also that the creation of the mash-ups created a demand for similar work from other local organizations. Barlow concludes that “mash-ups succeeded not only in mixing up online content and tools, but also people; in this case the students, local organizations, faculty, administration, and media that participated in the project’s upstream and downstream processes”. Those were certainly very successful mix-ups.

At the same 2007 conference – as you remember “the best ever” - in the session “Data Beyond Numbers: Using Data Creatively for Research”, Jinfang Niu from University of Michigan presented what is now her paper on “Reward and Punishment Mechanism for Research Data Sharing”. The author is defining the problems in the concept of data as a public good: “In the data sharing case, data producers make efforts to prepare the data for deposit, but the benefit of the data preparation largely goes to secondary users. In addition, data producers are at risk of being harmed by the misuse and misinterpretation of data by unqualified users, or by being charged with misconduct. That makes freeriding even more attractive. To motivate data producers to prepare and share data, there must be some incentive mechanisms”. To add rigidity to the analysis Jinfang Niu presents some mathematical modeling of the issue. (A very bold reviewer accepted the challenge of the mathematics.) As many of the IQ readers are working in data libraries and data archives, it could be easy simply to demand that all research data should be made available for the public. And this can be enforced by not releasing the last portion of grant money until that requirement is met, as in “trust is good, but control is better”. However, you could argue that the backside is that data that were better forgotten are now carried on indefinitely in archives. However, don’t expect that the mathematics can make the definitive judgment concerning whether to keep or to reject/scrap data. But do expect to get enlightened on the issues of reward and punishment, including a proposal to enforce citation of datasets - an issue IASSIST has promoted since the foundation of the organization.

In the same session Janet Stamatel from University at Albany presented some very interesting thoughts on “The Importance of Data Visualization in Data Literacy”. However, approached by one of the other presenters at the same session (this editor), Janet Stamatel preferred to work a bit more on the issue before publishing her presented thoughts on data visualization. So we as readers can look forward to that. In our talks it turned out that Janet Stamatel also had presented at the 2006 IASSIST conference - presently the next-best ever - and she agreed to publish her paper as an article in the IQ with the title: “An Overview of Publicly Available Quantitative Cross-National Crime Data”. The article is a specialist article that “reviews the content, data collection methods, geographic and temporal coverage, and accessibility of three main sources of publicly available, quantitative cross-national crime data, with a particular emphasis on recent changes with respect to data availability. These sources are the International Police Organization (Interpol), the United Nations Crime Surveys, and the European Sourcebook”. If your subject area is not in crime data, then the article can be read as an important contribution to the issue of globalization, that also emphasizes the problems of validity in cross-national data when the author remarks “researchers have noted considerable inconsistencies across the sources”.

Remember to have a look at the website http://iassistdata.org and the IASSIST blog - the IASSIST Communiqué – at http://iassistblog.org.

Articles for the IASSIST Quarterly are very welcome. Articles can be papers from IASSIST conferences, from other conferences, from local presentations, discussion input, etc. Contact the editor via e-mail: kbr@sam.sdu.dk.

Karsten Boye Rasmussen, November 2007
Introduction
This is the story of a class at a small, liberal-arts college. The class attempted to do something meaningful with data for people living in the community nearby. The college is Trinity College. The nearby community is Hartford, Connecticut. The students in the class created five “Google maps mashups.” One group of students mapped food resources in Hartford, everything from community gardens to grocery stores to food pantries (see Figure 1). Another group mapped houses in disrepair in Hartford’s south end, along with the contact information of the absentee owners whose negligence had caused such deterioration (see Figure 2).

The question: why should those in the data world care about this rather small project at a rather small school? As a sociologist, I am inspired to bring in a little sociological perspective. Susan Leigh Star and James R. Griesemar talk about boundary objects, “objects which are both plastic enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity across sites.” Sociologists who study science often talk in terms of upstream and downstream processes. Upstream processes refer to what happens before the point at which a technological innovation is considered “done” (ready for the marketplace, for consumption, etc.). Downstream processes refer to what happens after this pivotal, and as sociologists will note, socially-constructed point.

I argue that mashups, although not objects in the ordinary sense (since they are digital, not material), are boundary objects. In the case of the Trinity-College class, mashups facilitated the cooperation of those inside the college’s walls (faculty and students) with those outside (community members). These insiders and outsiders of academe approached such mashups while entertaining very different ideas about mashups what were good for. Yet despite these two groups’ different perspectives and intentions, the mashups, “plastic enough to adapt to local needs,” had the potential to appease both groups.

Upstream processes: making the maps
But to buy such an argument about the mashups’ status as boundary objects, the reader needs to learn a little about the upstream processes of the mashups created by the students in this course. Long before the creation of these mashups, a Trinity-College professor and I composed a set of goals for the course we were going to co-teach. We wanted the students to learn something about cities generally and about Hartford specifically. We wanted students to develop technical skills for managing data, but also communication, networking, and problem-solving skills. And we wanted students to participate in the construction of the knowledge they gained in the course by working with and doing something for the local community.

Perhaps now is the time for a quick definition. In the Web 2.0 world, a mashup refers to the product one creates when mixing together the dynamic elements of preexisting websites. As Rich Gibson and Schuyler Erle put it, “in music, when you create a new song by taking the melody from one song and the lyrics from another, it is called a mashup. A lot of times things go poorly, but now and then the results are stunning.” The same, they explain, is true for web mashups. Remixing websites might produce something silly or extraneous, or something significant and revealing.

Hence, we titled our course Invisible Cities for a reason: so that students could experiment with rearranging data in ways that would allow them to reveal something about the city of Hartford that was otherwise invisible. Mashups seemed like a timely, if not faddish, means to this end. The whole idea of a mashup is to take what already exists, stir it around and create something not yet seen.

Months prior to the beginning of the semester, the professor and I asked the leaders of two community organizations whether they would like to work with our class. The leaders agreed, but not for the reasons that mattered to us. After all, why would these organizations care about what privileged Trinity students learned? Instead, the leaders of these groups were primarily concerned with furthering their own organizational goals: informing and mobilizing residents in the south end of Hartford.
The course began and the students learned that they were expected to work with and for community groups. They also learned that their work would take the form of maps: collecting data for them, arranging that data in spreadsheets, and “mashing” the data with Google maps using the plethora of online tools that make such mashing easy. During the early part of the course, we trained students on the array of skills they would need to make all of this possible: they needed to know what a Google mashup was and how to make one, what a spreadsheet of data looked like and how to manage it, and what concerns Hartford residents and how to work with them.

A month into the semester, we arranged for the students to meet with the community-organizations’ leaders to exchange ideas about what kind of data would make sense for a map mashup. This meeting allowed us to hear what issues had relevance for these organizations. The list was long. They imagined maps that plotted, among other things: banks with free checking (for residents who otherwise cannot afford a checking account), known places where buses idled (emitting pollution and sickening kids), voting stations throughout the city (that were otherwise not well advertised), the location of advocacy groups (like themselves), and places for city residents to access the Internet (to access these maps and the larger web). After the meeting, when the class reconvened, the students faced the task of deciding what maps they could make, wanted to make, and mattered most.

Two of the maps that “won” in this contest are the two I mentioned at the beginning of this article. I mentioned these two because they represent different ways students went about mashing. The students who created the food resources map inherited an Excel spreadsheet from a third-party organization in Hartford that had already collected its own data on locations to access food in the city. The students did not have to collect data anew; instead, their work involved rearranging the Excel file so that it was readable by the online tool (called “Zeemaps”) they had chosen to make their mashup. But the students who created the mashup of abandoned properties did have to collect data. One of the organizations with which we collaborated had given this latter group of students an initial list of properties it had identified as problematic. But the students had to locate a large amount of additional data on their own: data from Hartford’s assessor that confirmed whether the houses had recently changed ownership, data from the Connecticut Secretary of State about whose names were behind many of the Limited Liability Corporations listed in place of owners for some of these properties, and data from a recent “City Scan” project that described the character of these properties’ blight (broken windows, lawns in need of mowing). Finally, the students turned to the phonebook to get owners’ contact information, since the organization that requested the map wanted community members to call these owners and ask them to clean up their properties.

**Downstream processes: after the maps**

But let us put aside for a moment the upstream processes that led to the creation of these mashups and examine their downstream processes, what occurred after the mashups were online, accessible to the public. The mashups, once created, were both opportunity-makers and pressure-cookers. The attention they received from the local media and the college administration allowed those of us who taught the course and the students enrolled in it to receive more accolades than any of us perhaps deserved. The mashups made the college “look good” in front of a statewide audience that often has perceived Trinity as disengaged from its urban environs. But the attention simultaneously put pressure on the professor and me to make more maps. In the wake of the mashups’ online publication, other local organizations were soon knocking on our doors, as were other Trinity faculty asking us if we could help them enter the map-mashup world. In other words, the mashups never were finished, even when we pretended that they were by putting them online for official consumption.

The attention also put pressure on the students and community groups to keep the original maps updated. As the world the data described changed, the maps needed to change, too. Furthermore, almost immediately, a few individuals wrote emails complaining that the mashups contained misinformation. Responding to these issues was a challenge. Although mashup technologies make updating easy, fact-checking takes time and people, both of which—at least inside the ivory tower—tend to disappear quite quickly once a semester ends and winter/summer break begins.

Finally, I personally questioned whether all five maps matched equally well the desires of the organizations with which we worked. The fact that the organizations perceived some maps as more useful than others was reflected in which maps required further refinement after the semester’s conclusion. While we are still working with one organization on one map a year and a half after we started, some of the maps we never touched again. I suspect this was not because these latter maps were perfect: they were not. Instead, because the class did not perfectly read the organizations’ and community’s needs, these latter maps found no constituency to care about them, discover their faults, and ask for a better product.

**Concluding remarks**

Time to take stock. I propose that there are a few things we can learn from our foray into map mashing. Mashups allowed our students to stand in a somewhat strange, but useful position relative to the local organizations they were trying to “serve” (a term I use with some trepidation). Mashups were easy: for the students to make and for the organizations to imagine and to use. In mashing, students were not so much producers as they were interpreters,
since they were not creating a product from scratch as much as they were using online, freely-available tools to provide organizations with a new perspective on what these organizations, in another form, already knew. For a small, liberal-arts college that lacks the resources larger schools often have, mashing allowed the students to do something with the community they probably could not have done any other way in the tight timeframe of an academic semester. Their newfound role as data massagers was a suitable one, given the institutional constraints. As Edward Maloney has noted, “what makes mash-ups interesting from a teaching and learning perspective is that they permit people with very little technical know-how to manage knowledge online, modeling solutions for others to see, collaborate on, and use in new ways.5

Furthermore, as boundary objects, mashups succeeded not only in mixing up online content and tools, but also people, in this case: the students, local organizations, faculty, administration, and media that participated in the project’s upstream and downstream processes. In this way, mashups were a means to “open data,” the idea behind the most recent IASSIST conference, where I first presented this paper. Now, as Gibson and Earle point out, mashing does not always have stunning results. As my reportage of the downstream processes makes clear, the mashing of people was not always as one might have hoped. However, this is all part of the mashing gamble: that the benefits of open data will outweigh the risks.

Acknowledgements
Special thanks to Dan Lloyd, my professor-collaborator on this project, David Tatem, for helping with some of the technical aspects, and Vincent Boisselle, for giving me permission to get involved.

References


Endnotes


4. Note here there are many different kinds of mashups, not just ones involving Google maps. Mashups might not even involve maps at all!


* Rachel E. Barlow presented this at the IASSIST 2007 conference in Montreal as “Maps that Mash: Daring, Dangerous, or Dumb?”. Contact information: Rachel E. Barlow, Trinity College Library, Raether Library and Information Technology Center, Trinity College, 300 Summit Street, Hartford, CT 06106. Rachael.Barlow@trincoll.edu, +1 (860) 297 – 4114
The 34th International Association for Social Science Information Services and Technology (IASSIST) annual conference will be held at the Stanford University, Palo Alto, California, USA, May 27-30, 2008. This year's conference, Technology of Data: Collection, Communication, Access and Preservation, examines the role of technology and tools in various aspects of the data life cycle.

The theme of this conference addresses how technology can affect aspects of data stewardship throughout the data lifecycle. The methods and media by which data are collected, shared, analyzed and saved are ever-changing, from punch cards and legal pads to online-surveys and tag clouds. There has been an explosion of data sources and topics; vast changes in compilation and dissemination methods; increasing awareness about access and associated licensing and privacy issues; and growing concern about the safeguarding and protection of valuable data resources for future use. The 2008 conference is an opportunity to discuss the role of technology – past, present, and future – in all of these arenas. We seek submissions of papers, poster/demonstration sessions, and panel sessions on the following topics:

- Issues and techniques for preserving “old” data as well as information “born digital”
- Methods, technology and questions surrounding data dissemination, including best practices and innovations
- Archival and preservation challenges presented by new processes
- Metadata
- Innovation in the use of data for teaching and research
- The legal issues surrounding new technologies
- Changes in resource discovery methods
- Data services in virtual spaces
- Providing services to users with different degrees of technical “savvy”
- Tools and spaces for research collaboration

Papers on other topics related to the conference theme will also be considered. The deadline for paper, session, and poster/demonstration proposals is December 17, 2007. The Conference Program Committee will send notification of the acceptance of proposals by February 8, 2008. Individual presentation proposals and session proposals are welcome. Proposals for complete sessions, typically a panel of three to four presentations within a 90-minute session, should provide information on the focus of the session, the organizer or moderator, and possible participants. The session organizer will be responsible for securing session participants. Organizers as well as panel participants are also welcome to submit additional paper proposals but please note that Conference Program Committee may need to limit the number of presentations per person.

Proposals for papers, sessions, and poster/demonstrations 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.

Proposals can be submitted via email to: iassist08@gmail.com
A conference website with on-line submission form will be available shortly. A separate call for workshops is also forthcoming.
Reward and Punishment Mechanism for Research Data Sharing

Abstract
Many funding agencies require grantees to deposit their data into an archive after they finish their research projects. The archive processes and disseminates the data for public use. These deposited data sets are public goods that benefit users and society. However, under voluntary contribution, public goods tend to be under-provided. For normal public goods, the contributors benefit from their own contributions as much as free-riders. Contributors are not harmed by their contributions. In the data sharing case, data producers make efforts to prepare the data for deposit, but the benefit of the data preparation largely goes to secondary users. In addition, data producers are at risk of being harmed by the misuse and misinterpretation of data by unqualified users, or by being charged with misconduct. That makes free-riding even more attractive. To motivate data producers to prepare and share data, there must be some incentive mechanisms. In this paper, I built a simple mathematic model to analyze the effects of punishment and reward. Hopefully it will help policy makers decide on incentive mechanisms for data sharing.

Background
Many funding agencies require grantees to deposit data into an archive after they finish their research projects, such as the National Institute of Justice (NIJ), the National Institutes of Health (NIH) in the United States, and the Medical Research Council (MRC) and the Economic and Social Research Council (ESRC) in the United Kingdom. The archive processes and disseminates data for public use. Data sharing benefits society in many ways. It saves funding and avoids repeated data collecting efforts, allows the verification and replication of research findings, facilitates scientific openness, deters scientific misconduct, and supports communication and progress.

Before deposit, data depositors need to prepare their data according to the requirements of data archives. The purpose of the preparation is to help secondary use of data and protect the privacy of human research subjects. Data preparation includes three kinds of work: preparing data, creating documentation and processing confidential information. Data preparation includes checking the integrity\(^1\) and consistency of data, careful naming of variables and choice of variable labels that will be easy for secondary users to understand, organizing the variables such as grouping them to enable secondary analysts to get an overview of the data quickly, etc. (ICPSR, 2005). Data documentation provides metadata about the data sets and research projects, such as the principal investigator of the project, when and where the data were collected, the methodology and procedures used to collect the data, details about codes, definitions of variables, frequencies, and the like (NIH, 2003). Even data collection instruments, such as questionnaires and interview guides are required parts of documentation. Documentation is indispensable for the searching, managing, preserving and re-using of data. In other words, without adequate documentation, secondary users of a data set will not be able to find the data, nor will they be able to interpret and analyze the data. As a result, the goal of data sharing will not be achieved. In addition, insufficient documentation might lead to the misuse of data or incorrect conclusions. To protect confidential information in data, all direct identifiers, such as names, addresses, telephone numbers, and Social Security Numbers, have to be removed. In addition, indirect identifiers and other information that could lead to “deductive disclosure” of participants’ identities should also be removed or processed before the data are made public.

Data preparation involves a lot of work, and a fair amount of it is done only for secondary users. For example, data producers do not have to process the confidential information if they keep the data for their own use. Many data producers do document data for their own use. However, documentation created for the producer’s own use are informal and biased toward short-term needs. To share with others, data producers have to take extra effort to shape the “public face” of their documentation (Markus, 2001). In addition, data producers should take the main responsibility for preparing data. Zimmerman (2003) found that both secondary data users and data managers (intermediaries or data archivists) agree that no one understands the data better than the scientists who gathered them, and that it is the data producers who must document data.
When publicly funded research data are disseminated to the public through the website of a data archive, no one is excluded from using them, and one individual’s use of the data and documentation does not reduce the amount available for other people. Those data sets are public goods by definition (Mas-Colell, et al., 1995). Since no one is excluded from the online data archive whether or not they have deposited data, as with other public goods people have strong incentives to free ride in preparing and depositing data. From the game theory perspective, free-riding in voluntary contribution to public goods tends to be the dominant strategy in a non-cooperative game (Bergstrom, et al., 1986; Cornes & Sandler, 1986).

For normal public goods, the contributors benefit from their own contributions in the same way as free-riders, and they are not harmed by their contributions. For example, once a bridge is built, the contributors and free-riders get the same benefit. However, in the data sharing case, the depositor of a data set does not benefit from the data he deposited in the same way as secondary users. A data depositor is unlikely to use his own data deposited into a data archive, either because he has used it before, or because he keeps his own data for future use. People mostly benefit from others’ contributions. The benefit of depositors’ effort in data preparation largely goes to the users. In addition, data producers are at risk of being harmed by the misuse and misinterpretation of data by unqualified users, or by being charged with misconduct. That makes free-riding even more attractive. To change this situation, there must be some incentive mechanisms to motivate researchers to prepare and deposit data.

The incentive mechanisms that some funding agencies have implemented focus on punishment for non-compliance with data sharing requirements, and pay less attention to rewards. According to the policy of the National Institutes of Health (NIH), in the case of noncompliance (depending on its severity and duration), NIH can take various actions to protect the Federal Government’s interests. In some instances, for example, the NIH may make data sharing an explicit term and condition of subsequent awards (NIH, 2003). Under the policy of the ESRC, “The final payment of an award will be withheld until data has been deposited in accordance with the requirements. The requirements of the data sharing policy are now a condition of ESRC research funding.” (ESRC, 2000). The data sharing policies of NIH and ESRC do not mandate that users cite the data they use, and they are against the idea that data producers require co-authorship as a condition for sharing the data. NIH explicitly stated that they do not offer rewards for doing well in data-sharing. Data from a survey of the grantees of a funding agency showed that some grantees expect rewards for data deposit. For example, one grantee said he would be more likely to deposit data if there were some sort of acknowledgment that he had deposited data, such as a certificate. Some other grantees claimed that some sort of punishment would make them more likely to deposit data, for example, if data deposit were mandatory to receive new funding from NIJ, or a prerequisite for publishing a paper derived from the data. One grantee was strongly against a punishment mechanism. He said: “Do you really want a system where archiving data prevents people from publishing or from doing new work? This would be a triumph of bureaucracy over common sense. If the funding agency becomes obsessed with bureaucratic requirements, they will drive away talented researchers.”

I believe that either punishments or rewards would provide incentives for data producers to take more effort in data preparation. But when decide the punishment or reward mechanisms, the level of the punishment and reward should be carefully chosen. Otherwise, unintended consequences might occur. To help illustrate this, I have built a very simple mathematical model

**The model**

Three parties are involved in the model. They are the data producer, the data user and the funding agency. In reality, there are many data producers and users. To make the problem simple, I only consider one data producer and one user. The data producer has total fund P. He chooses 0 and e to spend on research and data preparation respectively \( P = θ + e \). He benefits \( Ω(θ) \) from spending \( θ \) on research. I assume that \( Ω(θ) \) is concave, differentiable and \( Ω(θ) >0 \), meaning that the more the data producer spends on research, the more he will benefit, but the increase rate of the benefit decreases. See Figure 1 for the graph of \( Ω(θ) \). I also assume that the data producer always tries to maximize his benefit when making decisions. I analyzed and compared the social benefits generated from three scenarios: no reward & no punishment, punishment only and reward only. Social benefit is defined as the sum of the benefit gained by the data producer, the user and the funding agency in each scenario.

**Figure 1**
Scenario 1: No reward & no punishment
In the no reward & no punishment scenario, the data producer does not benefit from spending effort on data preparation, and he loses nothing if he does not spend any effort on data preparation. To state this formally, the utility function of the data producer is \( \Omega(\theta) = \Omega(P) \), \( 0 < \theta \leq P \). Since the more the data depositor spends on research, the more he benefits, the data depositor would spend \( \theta = 0 \) on data preparation to maximize his utility. His maximized utility is \( \Omega(\theta) = \Omega(P) \). For the user, since the data depositor did not spend any effort on data preparation for deposit there is no data to use, so the user’s utility is 0. The social benefit is the sum of the benefit of the depositor and the user: \( \Omega(P) \).

Scenario 2: Punishment only
In this scenario, the data producer will be punished if the effort he spends on data preparation is lower than a threshold. To state this formally, the benefit function is: if \( \theta \geq e' \), the data depositor’s benefit is \( \Omega(\theta) = \Omega(P-e) \). Since \( \Omega'(\theta) > 0, \Omega(0) = 0 \), the data depositor’s benefit is: \( \Omega(\theta) - f \), \( f > 0 \). \( f \) is a fine that the data depositor has to pay to the funding agency if he is punished. \( e' \) is the threshold for punishment.

In this case, to maximize his benefit, the data producer needs to choose the highest possible benefit he could get if he passes the threshold versus if he does not. Mathematically, he needs to maximize a benefit function of two parts, and pick the one that is larger. When \( \theta \geq e' \), the data depositor’s benefit is \( \Omega(\theta) = \Omega(P-e) \). Since \( \Omega'(\theta) > 0, \Omega(0) = 0 \), we need to minimize \( \theta \), the smallest value of \( \theta \) is \( e' \), and the maximized benefit of the data depositor is \( \Omega(P-e) \). When \( \theta < e' \), the data depositor’s benefit is \( \Omega(\theta) - f = \Omega(P-e) - f \), again we need to minimize \( \theta \) to maximize the data depositor’s benefit. The smallest value of \( \theta \) is 0, so the maximized utility of the data depositor is \( \Omega(P) - f \).

Now compare \( \Omega(P-e) \) and \( \Omega(P) - f \).

If \( \Omega(P-e) > \Omega(P) - f \), the data depositor will be punished by the threshold for punishment. There are two explanations for this. First, if the threshold is fixed, it means the data depositor would choose to pass the threshold to avoid punishment. There are two explanations for this. First, if the threshold is fixed, this means that the data depositor will benefit more by passing the threshold. So the data depositor would choose to pass the threshold to avoid punishment. There are two explanations for this. First, if the threshold is fixed, this means that the data depositor will benefit more by passing the threshold. So the data depositor would choose to pass the threshold to avoid punishment. Second, if the threshold is fixed, this means the threshold \( e' \) is too costly to meet, so the data depositor would rather be punished than meet the threshold.

If \( \Omega(P-e) = \Omega(P) - f \), the data depositor is indifferent between preparing data for deposit and getting punished.

If \( \Omega(P-e) < \Omega(P) - f \), the data producer benefits more from being punished than from preparing and depositing data. To maximize his benefit, the data depositor will choose to spend nothing on data preparation and be punished. There are two explanations for this. First, if the threshold is fixed, it means the punishment is not severe enough to deter non-compliance behaviors. Second, if the punishment level is fixed, it means the threshold \( e' \) is too costly to meet, so the data depositor would rather be punished than meet the threshold.

Based on the analysis above, we can see that the data depositor’s benefit in this scenario is max \( \{\Omega(P-e), \Omega(P) - f\} \).

For the user, when the data depositor would prefer to be punished than deposit data \( (\Omega(P-e) < \Omega(P) - f) \), there is no data to use. So the user’s benefit is 0. If Max \( \{\Omega(P-e), \Omega(P) - f\} = \Omega(P) - f \), the data depositor loses f, but the funding agency gets f. The social benefit is the sum of the benefits of the data depositors, the users and the funding agency. So the social benefit \( = \Omega(P) - f + f = \Omega(P) \). This is equal to the social benefit in the no punishment & no reward scenario. We can see that too weak a punishment or too high a standard for data preparation is not effective.

The data depositor is punished, yet there is no gain in social benefits. This actually confirms the findings of existing literature that punishment is effective only when it is relatively harsh (Trevino & Ball, 1992).

When the data depositor chooses to meet the threshold for data preparation, there is data available to use. But deposited data sets are not always used. In reality, there are various reasons. For example, a user does not use a data set because it does not fit his research purpose, or because the documentation of the data is not sufficient. Here, I assume that the probability that the data is used depends on the fund that the data producer spends on data preparation. The more fund the data producer spends on data preparation, the more likely the data is used by the user. To state this formally, there is a probability \( \pi(e') (\pi'(e) > 0, \pi(0) = 0) \) that the user will use the data. If he uses the data, the user will benefit v, so the user’s expected benefit of using data is \( v * \pi(e') \). The social benefit is: \( \Omega(P-e') + v * \pi(e') \). Remember the social benefit in the no punishment & no reward is \( \Omega(P) \).

So when \( \Omega(P) < \Omega(P-e') + v * \pi(e') \), it means that an appropriate punishment and a carefully selected threshold causes higher social benefit than no punishment & no reward. When \( \Omega(P) > \Omega(P-e') + v * \pi(e') \), it means the reverse.

Scenario 3: Reward only
In this scenario, when the deposited data is used, the producer of the data gets a reward r, and the user of the data set benefits v. The deposited data has a probability \( \pi(e) (\pi'(e) > 0, \pi(0) = 0) \) of being used. So the depositor’s expected benefit from the reward is \( r * \pi(e) \), the user’s expected benefit \( v * \pi(e) \). The producer’s total benefit is
the expected benefit from the reward plus the benefit from doing research: \( \Omega(0) + r \pi(e) \leftarrow \Omega(P-e) + r \pi(e) \).

To maximize the benefit of the data producer, we need to check the first order condition of \( \Omega(P-e) + r \pi(e) \).

If there is a value of “c” which makes \([\Omega(P-e) + r \pi(e)] = 0 \leftrightarrow r \pi(e) = \Omega(P-e)\), then the benefit of a data producer is maximized when the marginal benefit of spending an additional amount of funding on research is equal to the product of reward and the marginal probability of being used.

If the reward “r” is so big that no matter how small the marginal probability of the data being used (\( \pi(e) \)) is, the product (r * \( \pi(e) \)) is always greater than the marginal benefit of doing research \([r \pi(e) > \Omega(P-e)]\), it means that the function \( \Omega(P-e) + r \pi(e) \) is monotonically increasing in the interval \( e \in [0, P] \). In this case, utility is maximized when \( e = P \), which means that to maximize his benefit, the data depositor should spend all funding available on data preparation. If the reward is so small that no matter how big the marginal probability of the data being used, the product (r * \( \pi(e) \)) is always smaller than the marginal benefit of doing research \([r \pi(e) < \Omega(P-e)]\), the function \( \Omega(P-e) + r \pi(e) \) is monotonically decreasing in the interval \( e \in [0, P] \). Here the benefit is maximized when \( e = 0 \), which means that to maximize his benefit, the data depositor should spend all funding on research. Then the data producer will not deposit data and there is no data to use. In this case, the data producer is not rewarded because he did not deposit data. His benefit is \( \Omega(P) \). The user does not benefit because there is no data to use. The funding agency does not need to pay any reward to the producer. The social benefit = sum of benefit (producer, user and funding agency) = \( \Omega(P) \). It is exactly the same as the case with no reward & no punishment. In this case, the small reward is not effective at all. This confirms the findings of other literature that rewards should be of sufficient value, as rewards of insufficient value are the same as no reward at all (Buhler, 1992). Neither of these two cases are what we want. So we need to be careful not to make the reward too big or too small.

Suppose the data depositor’s benefit is maximized at \( e = e^*\), and the data user’s benefit is \( v^*\pi(e^*)\). The data depositor’s benefit is \( \Omega(P-e^*) + r \pi(e^*)\), but the \( r \pi(e^*) \) is from the funding agency. In other words, the funding agency loses \( r \pi(e^*) \) in rewarding the data depositor. So social benefit = sum of the benefit of (depositor, user and funding agency) = \( \Omega(P-e^*) + v^*\pi(e^*)\).

\( \Omega(P) \) is the special point for \( \Omega(P-e^*) + v^*\pi(e^*) \) where \( e = 0 \). \( e^* \) is the maximized point, so \( \Omega(P) \) cannot be greater than \( \Omega(P-e^*) + v^*\pi(e^*) \). So reward causes at least as much social benefit as no reward and no punishment. But we need to find an appropriate reward to make sure that the data depositor does not choose \( e = 0 \) or \( e = P \).

This simple model reveals the importance of choosing an appropriate level of punishment and reward, and an appropriate threshold for punishment. It does not deal with specific kinds of punishment or reward. For example, we do not consider whether we should punish non-compliers by withholding 10% of their final grant, or by factoring the quality of deposited data into consideration of future grants. I propose the following reward mechanism for data sharing policies: make the citation of data sets or the acknowledgement of data providers a mandatory requirement of publishing, the violation of which is treated in the same way as using but not citing published papers. Treat the citation of data the same as the citation of published papers in the performance evaluation of researchers.

As a complement for the model, here is a qualitative analysis of the punishment and reward mechanisms. Effective punishments force all data producers without plausible excuses to prepare and deposit data, which would make all data collected under public funding accessible to the public. This gives users chances to verify the research findings of data producers, which would deter scientific fraud and misconduct. On the other hand, not all data sets will be used heavily (Niu and Hedstrom, 2007). Under the punishment scenario, even if the data is very unlikely to be used in the future the data producer still needs to prepare and deposit data to avoid punishment. Also, the archive needs to process, disseminate and preserve the data. Enforcing uniform strong punishment on all data sets would cause the waste of resources. Unlike the coercive and uniform nature of punishments, rewards are inductive and selective. Rather than forcing researchers, rewards induce researchers to prepare and deposit data. Researchers who expect their data to be used by other people will be motivated to do better in data preparation. Data depositors who do not expect their data to be used will not deposit data, which may be a good choice. In this case, not all federally funded data sets will be made available to the public. The chance to verify some research is lost. Also, data producers decide their effort in data preparation based on the expected future use of their data, which might be hard to anticipate.

Acknowledgments
This research is funded by the National Science Foundation, Award # IIS 0456022 as part of the project entitled “Incentives for Data Producers to Create Archive-Ready Data Sets.”

I gratefully thank Linhong Chen’s help and instructions in building the model.

References


Endnotes:
1. Integrity means no wild codes or impossible values. For example, a respondent has 99 rather than 9 children. Consistency means the variable values are consistent, for example, a respondent doesn’t work but reports earnings.

2. That survey was done in 2006 by the team of the NSF project “Incentives for Data Producers to Create Archive-Ready Data Sets.”

3. P: the total fund available for the research project. θ: the amount of fund the data producer spent on data preparation. Ω(θ): the first derivative of Ω(θ).

4. The fine is paid by the data depositor to the funding agency. So when the data producer pays f to the funding agency, the data depositor loses f, and the funding agency gets f.

*Jinfang Niu is a PhD candidate at the School of Information, University of Michigan. She has a Master degree in Library Science and 3 years working experiences in Tsinghua University Library, China. Her research interests include data sharing, metadata, digital preservation and digital libraries. The paper was presented at the IASSIST 2007 conference in May in Montreal, Canada. Contact: niujf@umich.edu.
An Overview of Publicly Available Quantitative Cross-National Crime Data

Introduction
Although the media regularly reports on crime and violence worldwide, there is not a large body of academic research systematically analyzing cross-national crime patterns and trends or developing rigorous explanations of international variations in crime occurrences. Because of quantitative data limitations, the little knowledge that we have on the subject of cross-national crime variation tends to focus primarily on developed countries and often uses dated information (Stamatel, 2006).

However, there has been a renewed interest in comparative criminology over the past decade due to globalization and technological advancements that have improved the ability to conduct cross-national crime research (Howard, et al., 2000). In particular, there has generally been an increase in the amount and the quality of quantitative cross-national crime data available to researchers. This paper reviews the content, data collection methods, geographic and temporal coverage, and accessibility of three main sources of publicly available, quantitative cross-national crime data, with a particular emphasis on recent changes with respect to data availability. These sources are the International Police Organization (Interpol), the United Nations Crime Surveys, and the European Sourcebook. There are also a number of reliability and validity issues to consider when analyzing quantitative cross-national crime data, but these issues are beyond the scope of this paper, and they have been discussed elsewhere in the academic literature (see for example, Neapolitan 1997; Howard et al., 2000; Howard and Smith 2003; Rubin 2006).

Interpol
Interpol data collected by the International Police Organization (http://www.interpol.int/) are the oldest quantitative cross-national data source. They have been collected annually from Interpol member nations since 1950. Police representatives are requested to complete a multilingual, one-page tabular form recording aggregate counts of offenses known to the police for the entire country for 14 offenses: murder, sex offenses, rape, serious assault, all kinds of theft, aggravated theft, robbery, breaking and entering, motor vehicle theft, other thefts, fraud, counterfeiting, drug offenses, and the total number of recorded crimes. Additionally, the police are asked to identify the percentage of these offenses that were attempts and the percentage of cases solved. Lastly, respondents are asked to provide information on the total number of offenders, and the percentage of whom are female, minors, and aliens.

The data collection process is voluntary. Representatives are given a set of definitions for each of the offense categories and requested to report the numbers for their countries according to these guidelines. It is important to note that Interpol does not provide any quality control measures on this data collection process. Accordingly, the organization provides a disclaimer in their publications stating that “the information given is in no way intended for use as a basis for comparisons between different countries” and that “the figures must be interpreted with caution” (Interpol 1999). However, given the limited number of quantitative cross-national crime data sources, researchers, the media, and others have regularly used Interpol data for cross-national comparisons. Interpol became concerned about the potential misuse of these data and stopped making them publicly available as of 2000, to the dismay of many comparative researchers. According to the Interpol website, the crime statistics are currently only available to “authorized police users” (http://www.interpol.int/public/Statistics/ICS/downloadList.asp).

The first published Interpol statistics contained data from 36 nations in 1954, and the number has increased since then. In 1998, 116 countries reported data to Interpol, and 93 did so in 1999. However, the composition of countries varies from one year to the next. For example, only 82 countries reported data to Interpol in both 1998 and 1999. This represents approximately one-third of all countries in the world. According to Neapolitan (1997), 150 countries have reported data at least once to Interpol. Figure 1 illustrates which countries reported to Interpol in 1998 and 1999. The large and populous countries of China, India, and the United States are noticeably absent from the collection during this time, and coverage for the Middle East and Africa is uneven.

Interpol data are officially available only in hard copy. Until 1992 the reports, called International Crime Statistics, were published biennially, and then annually from 1993 until 1999 when they stopped releasing the data publicly.
It is not clear at this time whether Interpol will allow researchers to petition for access to their data. Some of the Interpol data are also available electronically in the Correlates of Crime: A Study of 52 Nations, 1960-1984 dataset (ICPSR 9258) compiled by Richard Bennett (1990), who is currently updating the collection.

**United Nations Crime Surveys**

The United Nations Crime Surveys (UNCS) (http://www.unodc.org/unodc/en/crime_cicp_surveys.html), officially called the United Nations Surveys of Crime and Operations of Criminal Justice Systems, have been collecting quantitative cross-national crime data from United Nations member states since 1970. The stated goal of this program is

“to collect data on the incidence of reported crime and the operations of criminal justice systems with a view to improving the analysis and dissemination of that information globally. The survey results will provide an overview of trends and inter-relationships between various parts of the criminal justice system to promote informed decision-making in administration, nationally and internationally” (United Nations Office on Drugs and Crime, 2007).

The UNCS sends multi-lingual questionnaires to coordinating officers in United Nations member countries. The coordinating officers are typically United Nations correspondents who compile the data with assistance from government employees from a variety of relevant departments, such as police and corrections. The questionnaires are designed to record aggregate, national-level figures about crime and criminal justice systems in four areas: police, prosecution, courts, and corrections. The police section includes the number of offenses reported to the police annually for 18 offenses: intentional committed homicide, intentional attempted homicide, intentional homicide committed with a firearm, non-intentional homicide, major assault, total assault, rape, robbery, major theft, total theft, automobile theft, burglary, fraud, embezzlement, drug-related crime, bribery and/or corruption, kidnapping, and total recorded crimes. Other data collected include personnel figures and budgets for different components of the criminal justice system, as well as the number of suspects by age and sex who encounter different stages of the criminal justice system. The United Nations provides a modest level of quality control on the collected data. The data are considered official statements by national governments about the extent of crime and the operations of criminal justice systems in their countries and, therefore, these data are considered more valid than the Interpol data.

The UNCS collects data in multi-year waves (see Table 1). The first five waves were administered every five or six years, then subsequent waves were issued every three years, and most recently, every two years. In response to the demand for more recent cross-national crime data, the UNCS has increased the frequency of administration and dissemination. Geographic coverage has varied by wave, but on average about 80 countries participated in
any given wave. Countries’ participation varies by wave, similar to Interpol, thereby making it difficult to construct longitudinal data series for a large number of countries. Figure 2 illustrates the geographic coverage of the UNCS for wave 7, covering 1998 to 2000. While coverage for North America and Europe is quite good, it is sparse or inconsistent for other regions of the world.

### Table 1 Geographic and Temporal Coverage in the UNCS

<table>
<thead>
<tr>
<th>Wave</th>
<th>Years</th>
<th># of Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1970-1975</td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td>1975-1980</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>1980-1985</td>
<td>78</td>
</tr>
<tr>
<td>4</td>
<td>1985-1990</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>1990-1994</td>
<td>92</td>
</tr>
<tr>
<td>6</td>
<td>1995-1997</td>
<td>75</td>
</tr>
<tr>
<td>7</td>
<td>1998-2000</td>
<td>92</td>
</tr>
<tr>
<td>8</td>
<td>2001-2002</td>
<td>65</td>
</tr>
<tr>
<td>9</td>
<td>2003-2004</td>
<td>71</td>
</tr>
</tbody>
</table>

The UNCS datasets are available electronically through the website of the United Nations Office on Drugs and Crime, although the file format available depends upon the wave of data collection (see Table 2). The first five waves are also available through ICPSR, including a harmonized longitudinal file for 1970 to 1994, which can be downloaded or analyzed online (Burnham and Burnham 1999).

### European Sourcebook

The third, and newest, source of quantitative cross-national crime data is the European Sourcebook of Crime and Criminal Justice Statistics (http://www.europeansourcebook.org/), which was modeled after the Sourcebook of Criminal Justice Statistics (http://www.albany.edu/sourcebook/) produced by the United States Bureau of Justice Statistics. This data collection effort began in 1990 in response to the growing demand for accurate and timely crime data, particularly in the context of the growing Council of Europe, and the concern over the limitations of the other two quantitative cross-national crime data sources. The content of the data collected by the European Sourcebook is similar to that of the other sources. In particular, the European Sourcebook includes annual information on total number offenses reported to the police in each country for 14 types of crimes: total intentional homicide, completed intentional homicide, assault, rape, total robbery, armed robbery, total theft, theft of motor vehicle, bicycle theft, total burglary, domestic burglary, total drug offenses, total drug trafficking, and serious drug trafficking. Additionally, homicide victimization data from the World Health Organization and select measures from the International Crime Victimization Surveys are provided as points of comparison against the official records data reported by the police.

Like the UNCS, the European Sourcebook also collects information about the number of offenders by crime type, the percentage of offenders who are female, minors, and aliens, as well as information about caseloads, staffing, and dispositions related to prosecution, conviction, and corrections. However, the data about the criminal justice system are considered secondary indicators and are generally collected every five years, rather than annually.

Although the data for the European Sourcebook come from official records like the other two sources, the...
European Sourcebook differs in the way the data are collected and the high level of quality control imposed on the data collection effort. Each country participating in the effort has a national correspondent who is an expert in crime and criminal justice statistics and who is responsible for collecting and checking the data. These experts are typically either Ministry of Justice employees or academics. In addition to using standard classifications schemes for collecting data across countries, the national correspondents also agree upon certain quality control measures to ensure the accuracy and reliability of the data. Documentation for this collection is also more detailed than for Interpol and the UNCS data sources. For example, the European Sourcebook not only provides descriptions of crime definitions, but also detailed explanations of why some countries do not conform to these definitions. The European Sourcebook also provides a thorough discussion of the methodological limitations of this data collection.

The European Sourcebook began collecting data in 1990 and, like the UNCS, it is administered in waves. The first wave collected data from 1990 to 1996 and included 36 European countries. The second wave gathered data from 1995 to 2000 from 40 European countries. The third wave covered the years 2000 to 2003 for 37 countries, but it was a limited edition and not all of the tables were updated. In general the coverage for Europe is thorough and consistent, although some countries do not report data to this source, such as Serbia, Montenegro, and Bosnia-Herzegovina. Data from the European Sourcebook are available electronically from the Internet (http://www.europesourcebook.org/) in a variety of file formats (see Table 2).

Summary
The three main sources of quantitative cross-national crime data share the same goal to provide reliable, annual counts of the frequency of occurrence of conventional crimes across countries and, for the UNCS and the European Sourcebook, measures of the operations of criminal justice systems worldwide. Although the crime count data for all of these sources come from police reports, the sources differ in terms of the way the data are collected and the amount of quality control exercised. Additionally, since reporting crime data to these international sources is voluntary and depends upon membership to the broader organization, the sources vary greatly in terms of geographic and temporal coverage. It may be tempting to simply combine data from these sources to maximize sample sizes, but researchers have noted considerable inconsistencies across the sources, particularly for certain offenses and time periods (e.g., Bennett and Lynch, 1990; Howard and Smith, 2003; Gottschalk, et al., 2007). Quantitative cross-national crime data collections have improved recently with respect to the frequency of collection, greater electronic availability, and to some extent, improved quality control. Nonetheless, there are still numerous methodological concerns regarding these data and researchers should use them carefully and responsibly.

References


Endnotes:
1. A previous version of this paper was presented at the IASSIST meeting in Ann Arbor, Michigan, 2006. Janet P. Stamatel is an Assistant Professor in the School of Criminal Justice and the Department of Informatics at the University at Albany, 135 Western Ave., Albany, NY 12209, jstamatel@albany.edu.

2. The three sources discussed in this paper are all official records data. Quantitative cross-national crime data are also collected from victimization surveys (see the International Crime Victims Surveys at http://www.unicri.it/wwd/analysis/icvs/index.php) or self-report delinquency surveys (International Self-Report Delinquency Study). Although these are important sources of data, they have limited geographic and temporal coverage and, therefore, they are not used as often for cross-national crime analyses. Some researchers also use mortality data from the World Health Organization (http://www.who.int/whosis/en/) to analyze aggregate homicide victimizations. This is an important source of cross-national homicide data, although it is not as easily accessible as the sources discussed in this paper and it only provides a measure of one criminal offense.

3. Dr. Rosemary Barberet, John Jay College of Criminal Justice, will present a paper at the 2007 meeting of the American Society of Criminology titled “The Contribution of Interpol Crime Data to Cross-National Criminology” that will examine the impact of the loss of access to these data on comparative crime research.

4. Information obtained from personal communication with the author
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 subscriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

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://www.iassistdata.org/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

Please fill out the following information:

Name:

Job Title:

Organization:

Address:

City:    State/Province:

Postal Code:    Country:

Phone:    FAX:

E-mail:    URL: