Historical Methods: A Journal of Quantitative and Interdisciplinary History

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/vhim20

Do ut des (I Give So That You Give Back): Collaboratories as a New Method for Scholarly Communication and Cooperation for Global History

Tine de Moor [a] & Jan Luiten van Zanden [b]

[a] Research Institute for History and Culture, Utrecht University

[b] International Institute of Social History, Utrecht University

Version of record first published: 07 Aug 2010.

To cite this article: Tine de Moor & Jan Luiten van Zanden (2008): Do ut des (I Give So That You Give Back): Collaboratories as a New Method for Scholarly Communication and Cooperation for Global History, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 41:2, 67-80

To link to this article: http://dx.doi.org/10.3200/HMTS.41.2.67-80

HM
Data Developments
HISTORICAL METHODS, Spring 2008, Volume 41, Number 2

Do ut des (I Give So That You Give Back)

Collaboratories as a New Method for Scholarly Communication and Cooperation for Global History

TINE DE MOOR
Research Institute for History and Culture
Utrecht University

JAN LUITEN VAN ZANDEN
Utrecht University
International Institute of Social History

Abstract. With the growing need for large sets of data in historical science, especially now that global history and world history are the objects of increased attention, cooperation among historians has become more useful and necessary. An academic career is often not long enough to gather all the data necessary to support a hypothesis. But researchers are not always willing to share their data because developing data sets requires a great deal of time and labor. The result is that large, highly interesting data sets are often not accessible to interested colleagues. Even if colleagues are able to access the data sets, many data sets prove to be incompatible with other data sets: this is sometimes caused by incompatible data formats or database designs, other times, by a lack of metadata. The authors investigate new methods for scholarly communication and cooperation, paying special attention to collaboratories, or laboratories without walls, and what they can mean for preserving, sharing, and maintaining the quality of large data sets in the humanities and social sciences. Examples from the area of global history illustrate these points. The difficulties of setting up a collaboratory and interaction with other methods of data collection, such as data archives and data availability policy journals, and their benefit the historical sciences, are also discussed.

Keywords: collaboratories, cooperation, data sets, global history

Data Sharing and "Big History": An Introduction

In the historical sciences, there is a growing need for more internationally oriented data collection to answer the research questions that are increasingly posed in the expanding field of global and world history, or "big history."[1] Global questions require global data. To encourage efficient and effective use of data that have already been collected and to promote cooperation among scholars worldwide, new methods of data sharing and scholarly communication must be designed. Developments in information sciences have enhanced possibilities in data collection (through digitization and databases), thus allowing historians to process a growing amount of data. Data hubs have filled this need to a certain extent for the past few years, at least for economic historians.[2] More is needed to practice global history than raw data alone, however. In addition to indispensable metadata, working with data from sources that originated in different parts of the world requires other knowledge: the context in which the information was collected, the method of collection, and possible problems in the interpretation of the data (Allen et al. 2005). Intensive communication and interaction among researchers when setting up large-scale data infrastructures on specific topics is a must, because it is impossible to create and maintain databases that contain the knowledge necessary to fully understand context and origin. Developing such large data sets usually takes considerable time. To prevent the disintegration of initiatives such as economic history data hubs and the data themselves, a certain degree of institutionalization can help safeguard the future storage and accessibility of data.

Over the past two decades, the natural sciences have established *collaboratories*, that is, laboratories without walls, to fulfill these needs. Scientists are connected to one another, to instruments, and to data, independent of time and location, thereby creating a virtual community of peers. As will subsequently be explained, whereas sharing data can enhance

productivity, it also involves certain risks. When collecting and exchanging data among researchers—in a collaboratory or data archive—there are risks that can only be prevented in a limited way by technology. Although this is often overlooked, intensive cooperation cannot be durable without a solid institutional framework to anticipate problems that arise from collective action. One such problem is "free riding," or taking advantage of what is offered to the group without contributing to it. After providing some theoretical background about the possible risks of collaboratories, we will show how research on the functioning of what is now being described as an "information commons" may offer some guidance in the construction of such a framework. A transparent institutional design can create the right incentives for researchers to cooperate and contribute to the common good: to encourage high-quality scientific output.

Here, we introduce the concept of collaboratories in historical research and explain how this form of collaboration complements existing forms of data collection and distribution, such as centralized historical data archives. Traditionally, historical data archives have not had an institutional design that initiates and supports intensive interaction among researchers. For a researcher, participation in a data archive is usually a one-way trajectory. Typically, after submission of the data, a researcher has no further contact with the data archive or with those who use the deposited data. Where the information cycle for a depositor of data in a data archive stops, the advantages of a collaboratory begin: mutual exchange, peer review, and improvement of data quality. Not only does it become possible to amass data on the same theme, thus creating new opportunities for comparative research, but interaction with other researchers also offers more background information on the data and their possible uses. In this article, we will explain how, although data archives, collaboratories, and other new initiatives, such as journals with a data availability policy (DAP), which require authors to make all data referenced in an article available to any researcher, may have different objectives and outcomes, they can play complementary roles at the beginning and end of specific information cycles.

We use two examples of collaborative projects that are close to the ideal of a true collaboratory (although they do not entirely realize the ideal) to illustrate our ideas: the Global Price and Income History Group (http://gpih.ucdavis.edu/), in cooperation with the Historical Prices and Wages initiative at the International Institute of Social History (IISH) (http://www.iisg.nl/hpw/), and the Historical Sample of the Netherlands (HSN) (http://www.iisg.nl/~hsn/), also hosted by the IISH—an international collaboration among universities and archives to "preserve original population data, microdata, and their supporting documentation, and to improve access to these data in accordance with national confidentiality standards" (HSN 2006).

Where Do All the Data Go? The Need for More Cooperation among Scholars in the Social Sciences and Humanities

For the past few decades, digitizing data has become the main method of data collection and storage for historians and other social scientists. Considering the vast amounts of data that have been collected since the 1980s, the scientific community, especially in the social sciences and humanities, has only benefited from one of the most tempting advantages of digitization to a very limited degree: the possibility to exchange and reuse data and to pool and share large bodies of data among groups of scholars. Since the 1980s, data archives have been set up to collect and preserve those data sets as much as possible. There are still two issues that prevent data archives from becoming central players in this field, however. First, convincing those who have the data to make them accessible for other parties to use seems to be somewhat difficult. Researchers still wrestle with the idea that others may benefit from using their painstakingly gathered data sets and, perhaps even more important, they also fear that by making their data sets public, mistakes in the collection or processing of their data and their (to be) published results may be discovered. Releasing them gives the collector a sense of losing control over the carefully collected data.

A second issue is the link between the data archive and other parties who may potentially be interested in the collected data. In some cases, data archives contain large amounts of data, but potential secondary users cannot find their way easily to those virtual archives. Data archives may offer potentially diverse uses, but no direct application possibilities, a fact that the collectors of data may not feel the need to point out to the rest of the scholarly community. When the importance of storing the data for the future has been agreed on, the issue of how to realize the value for future researchers remains. Data quality plays an important role in the relationships among the three involved parties: data collectors, secondary users, and data archives. The collectors fear that the quality of their data may be questioned and the secondary users fear that the data may not be trustworthy. If the distance between these parties were reduced, higher quality standards could be obtained. Part of the solution is that the quality of data can improve significantly through peer review. In global comparisons of data, peer review and peer support can become an essential part of the scientific process. As data archives are not topic-specific, one cannot expect the archivist to offer the necessary expertise to comment on the data or adequately evaluate what is offered for storage. Here, the scholarly community should step in. The quality depends entirely on the self-criticism and goodwill of the depositor of the data. The metadata provided with the data sets, or the publications based on those data, are essential to understanding what has been done with the information in the databases but are often not sufficiently specific for other interested users.

Since 2000, some journals have integrated a DAP clause in their instructions for submissions—the *Journal of Political Economy* (2008) and the *American Economic Review* (2008) require that "prior to publication, the data, programs, and other details of the computations sufficient to permit replication" must be submitted to the journal to enhance external peer review by those not included in the prepublication review process.[3] This development, made possible by the Internet, is part of a new trend in the academic world toward more openness and freedom of information, which is evident in the increasing establishment of e-journals, some of which are also open access.[4] E-journals offer the possibility to link to other articles. In the future, journals may well want to offer the possibility of linking to the data on which the article is based.

Journals with a DAP are a new trend that has not yet caught on in the historical sciences. Attempts to convince some of the best journals on economic history to introduce similar policies have not yet been successful. A few journals offer data in the form of printed appendices, but as these no longer have the same structure and flexibility of the original database, third parties are probably not very likely to check the published results. Nor are the possibilities offered by state-of-the-art information technology used to the fullest. Although these new developments in scientific publishing are definitely positive, they do not guarantee access to all the data that have been gathered, solve problems such as divergent data formats, or guarantee the future of the original digital data sets. The questions of how long will they be available online, how well they are being stored, and in what format remain unanswered. Moreover, there are hardly any provisions for metadata as offered by historical data archives. For the sake of large-scale international comparisons, we need to go a step further: data should not only be protected against being lost but they should also be made public for peer review after they have been used (in data archives) or published (in journals). There is a need for more interaction and exchange of data and metadata among researchers before the publication stage to construct large high-quality international data sets. It is clear that, although archives and e-journals both perform absolutely necessary functions for distribution and preservation of data, lack some essential attributes to make this happen; collaboratories can help fill these gaps.

Creating a collaboratory entails more than simply contacting fellow researchers and deciding "to do something together," collecting data collectively, as is already done in data hubs. The goal of a collaboratory is to provide complete location-independent access to instruments, data collection, analysis resources, and other researchers in a particular field (Agarwal, Sachs, and Johnston 1998). Exchange of data and intensive cooperation among scientists are in many ways beneficial for all participants: those who participate can benefit from the data collection and from the intellectual advantages "collective thinking" may have. Potentially, in data collection and in the actual research process, the total outcome can be more than the sum of its parts. Nevertheless, informal agreements alone will never lead to a sustainable cooperation among researchers. An institutional design that offers a stable and sufficiently flexible environment for the collaboratory participants is necessary. Research on the functioning of an information commons can offer some guidance.

What Is a Collaboratory?

According to William A. Wulf (1993, 854–55), the term *collaboratory* first appeared in the scholarly community in the late 1980s as "a laboratory without walls, where scientists are connected to each other, to instruments, and to data independent of time and location." It can be regarded as "an organizational entity that spans distance, supports rich and recurring human interaction oriented to a common research area, and provides access to data sources, artefacts and tools required to accomplish research tasks" (Finholt 2002, 647). Collaboratories can provide communication environments and tools for scientists, serve as a communication tool for students,[5] allow the collection of data, give online access to data to members and, in some cases, nonmembers of the collaboratory, and create the means to share scientific instruments within research or learning communities.[6] The idea was originally developed for the joint use of instruments but has now been applied to many other forms of collective work.[7]

Notwithstanding the many variations in objectives of current collaboratories, they all have several common features (Lunsford and Bruce 2001):

- *Boundary crossing*: A collaboratory is first a tool to bridge the gaps and distances of: (1) geography, by providing international access through the Internet; (2) time, by supplying both synchronous and asynchronous communication technologies; (3) institutions, by allowing groups access to tools and materials of common interest; (4) disciplines, by enabling the participants to decide what resources are most relevant to a topic, without regard for traditional understandings of what constitutes a particular discipline.
- *Shared inquiry*: Participants not only share common goals in data collection but also a common set of problems or issues of interest, which they study in depth.
- *Intentionality*: A collaboratory is a joint venture; there is a shared consciousness of the status of its Web site as a mutual project.
- *Active participation and contribution*: The success of a collaboratory is to a large degree decided on by the extent that its members use and add to its resources.
- *Members only:* Although the data collected by a collaboratory can become freely accessible over time (usually after publication of the research results), participation in

a collaboratory is generally for members only. Membership is usually restricted to peers in the research field.

- *Access to shared resources*: Collaboratories provide unique information (data, links, research findings) and tools needed by participants.
- *Technologies*: Collaboratories involve technologies. These vary from scientific instruments shared by sophisticated communities, the unique symbol systems used among participants, or the information technologies necessary to communicate.[8]
- *Limited in time*: Collaboratories are set up to reach certain research goals (creating a data set, answering certain research questions). Once these goals have been attained, the collaboratory is dissolved, although in some cases their results remain available via the collaboratory's Web site.

Collaboratories differ in several ways from the wikis—Web sites that allow any user to contribute or modify content—that are increasingly being established on the Internet. Even though wikis can play a valuable role in offering free and unrestricted access to information, the user has no guarantee that the data are reliable. If the author's credentials and validity are unknown, the quality of the information cannot be judged. Although collaboratories can be as free as a wiki for the dissemination of data (if the authors of the data allow it), access to the input side is restricted to peers. In this way, the quality level is controlled, and collaborators can trust one another's data.

The development of collaboratories according to the given common set of features stems from a trend that has been developing over the last half century toward large-scale projects, or so-called big science, which requires more collaboration among scientists, not only in the natural sciences but also in the social sciences and the humanities (Endersby 1996; Weinberg 1961). The natural sciences have an advantage: they also have a common scientific language, such as chemical or mathematical formulas, and most of the data used for research are newly created, allowing the researchers to start from a consensus about which data to collect and how to input them in the database (whose structure can be agreed on in advance). History data must be derived from what has been left by our predecessors. The Historical Prices and Wages database shows that this is not a straightforward task. Its information is derived from a multitude of very different sources, with varying accuracy and applicability. Some sources tend to underestimate or overestimate the situation, depending on the purposes of the archival document. Some data refer to hourly wages, other data to daily, weekly, or even annual wages, with varying numbers of hours attached to each time entity, and differences from place to place. Wages may also differ with differences in experience, education, gender, age, task, or function of the receiver, or they may be supplemented with wages in kind. And what about estimating the cost of labor in societies where wage labor is still a rare phenomenon? All in all, extensive metaknowledge is necessary before wages across countries and over time can be compared. The Global Price and Income History Group of University of California, Davis, in collaboration with the Historical Prices and Wages Initiative in Amsterdam and the help of many researchers associated with other universities, have managed to bring together a multitude of such data, along with working papers that contain extensive metaknowledge. Another example of such a collaboratory is the HSN, which "strives to construct life histories as completely as possible for a representative portion of the nineteenth- and twentieth-century population in the Netherlands" based on a sample from all persons born in the Netherlands from 1812 to 1922. The data were and still are being collected in the Netherlands by a large team of researchers, in a data format that was agreed to by an international team of researchers who coordinate the collection of similar data all over the world (International Microdata Access Group 2002; IISH 1935).

Both examples match the previously mentioned features and can be considered a type of collaboratory. The concept of a collaboratory has even more potential than these two initiatives show, however. They do not fully exploit the potential of the idea—the Historical Prices and Wages database could be improved if one paid more attention to metadata standards and to the future of the data set in the long term. These are issues the HSN project has taken into account. The two examples differ in the degree to which they offer third parties accessibility to their collected data. In principle, whereas a collaboratory is not specifically designed to do this, once the data have been used by scientists, there is no reason to keep the public from having access to the data sets (with the possible exception of privacy reasons in the case of HSN, but even those can be solved). Global Price and Income History Group publishes all its results on two Web sites (hosted by institutes at Davis and Amsterdam) (Global Price and Income History Group 2004). Opening these data collections, which were developed by leading researchers, can provide added value to society as a whole, and encouraging other researchers who may not yet be on a peer level in that scientific discipline (e.g., young researchers) to use the data and develop their own line of research or a new collaboratory.

The Institutional Design of a Collaboratory: Collaboratories as Commons

The "Science of Collaboratories" project hosted at the University of Michigan has so far identified more than 200 collaboratories that reflect the ambitious challenges of today's science, the extremely expensive instrumentation that they often require, and the availability of high-capacity networks and computing resources. Very few social scientists are included in this inventory.[9] Considering the

advantages of scale and other benefits that can be gained from collaboration, the large number of collaboratories should not be surprising. Although new collaboratories have been formed continuously since the 1990s, and many have realized high-quality scientific results, many of them no longer exist. Nevertheless, the fact that these collaboratories no longer exist does not mean the formula does not work. Collaboratories usually terminate their own existence when the immediate need for the researchers' cooperation is no longer present—that is, after the data are collected and the results of the specific research have been published. Collaboratories are topic-oriented research networks that often serve an ad hoc need within certain scientific circles. In the process of working to achieve their goals, however, researchers face several problems that may be solved if an appropriate institutional environment is created. Such a framework should be another issue: Should it be embedded in a larger steering institution (like a data archive) or should it be autonomous? Should it be open access, or should the membership be restricted? What other rules should be created to prevent free riding?

Given the features of a collaboratory, it becomes clear that this form of scholarly communication and data sharing is quite similar to a much older institution, the commons. Commons, often referred to in the social sciences as "Common Pool Resources" are "natural or man-made resources sufficiently large that it is costly to exclude users from obtaining subtractable-resource units" (Ostrom 1992, 259). The historical commons of Europe were formed in accordance with the local laws and powers from the bottom-up by villagers who needed pasture land for agricultural purposes. The villagers limited access to these agricultural areas to members of the commons, and designed use and management rules that tried to prevent overharvesting to secure a sustainable future. Although the way they functioned is now receiving increasing attention, their dissolution has drawn most interest as a potentially important factor in creating the destitute, cheap labor force necessary to make the Industrial Revolution possible. In the social sciences, the theory and practice of today's common pool resources have been inspirations to help understand new forms of virtual information on knowledge commons (each term has a slightly different meaning), such as the Internet.

Very little research has been done on the requirements necessary to build a successful collaboratory. Besides the advice to honor some very general rules such as "make sure your working community is ready" and "tackle big questions," no good recipes for setting up a successful collaboratory have been developed.[10] One can, however, find much inspiration for designing well-functioning institutional frameworks from the research on information/knowledge commons. The basis of this field of study is the question of how to organize and regulate the use of goods that are held in common. The common-property debate started in the 1970s as a reaction to the "Tragedy of the Commons" article by Garret Hardin (1968), and initially focused only on the management of natural resources. Hardin claimed that human greed made it impossible to manage a shared resource in a sustainable way. Because members of a common property arrangement think only of the short-term advantages obtainable from a common resource, common property arrangements are bound to end disastrously. When natural resources are concerned, overexploitation and eventually the disappearance of the resource are the usual results. Hence, the tragedy of the commons. According to Hardin, the only solution to prevent such an outcome was to privatize the good or make it public property, whereby the use of the resource would be arranged by the state. Over the past 25 years, researchers from the social sciences have countered this pessimistic view of common property by describing and analyzing a wide range of sustainable commons all over the world. They showed that common property can work if a specific institutional design is followed. Elinor Ostrom (1990) gave an initial overview of the characteristics of such a design in her seminal work "Governing the Commons"; her design principles can be used as guidelines to develop a framework for collaboratories. During the 1990s, the debate broadened to include different forms of common property beyond natural resources; researchers have started to apply the theoretical results of research on local natural resources to what has been termed the "global commons" (water, air, etc.) and the virtual commons. This attention to the virtual commons was a consequence of the growing popularity of the Internet, and other new forms of electronic communication in general, especially among scholars (e.g., e-journals).

Collaboratories can be considered another example of a virtual commons. The most characteristic problem faced by users of common property is scarcity: there is a limited amount of the good, and many people want to use it. In the case of the collaboratories, the scarcity is not so much the availability of the data but the professional benefits to be reaped from using the data, more precisely the publication of the results based on the data. If participants of the collaboratory publish data gathered by others, they can be considered free riders. The information common (as is so for any other common) will experience a collapse, breakdown, or tragedy if free-riding behavior starts to dominate cooperative behavior. Free riding takes place when someone seeks to gain an advantage at the expense of others by not (or insufficiently) contributing to a joint effort; it exists when a person benefits from the contributions of others without making any corresponding contribution. It leads to a decrease in the total value of the good. The advantage of collaboration is also its biggest disadvantage: sharing information also means losing control over one's work and the creation of the opportunity for many to profit from the contributions of some. It can be said that the collaboratories as presented here are "in between" formal and informal communication. They are created from the informal exchange of ideas by scholars who have acknowledged one another as

peers in their field of study. By allowing another peer in the network, they create their own "information common."[11]

The processes that take place within the collaboratory as knowledge commons are related to data in their unfinished format, whereas the data they make available to the public are finished results, results that all peers consider good enough to be disseminated in wider circles. As soon as the result has become a public good, participants withdraw their (copy)rights as creators, also withdrawing their responsibility for anything that happens to the data thereafter.[12] In practice, if the data are used by others (e.g., in publications), the members of the collaboratory cannot be held responsible for possible errors in publications that have used their data.

One of the potential threats to every common—whether it consists of natural resources or virtual information—is a rise in the number of members. A larger number of participants in a collaboratory can have benefits, however. The more people contribute information, the more data are available to others. There are also several potential disadvantages to more participants. The participants can become so numerous that efficient management of the group becomes impossible or very difficult at the least. Moreover, it is unlikely that all participants in the collaboratory will benefit as much as they have contributed. Once the data are publicly available, however, the situation changes: at that point, the members of the collaboratory have already reaped the benefits of their efforts. Afterward, it can only be beneficial if as many people as possible use the data.

Often, it is not the number of people that becomes a problem to the functioning of the collaboratory, but the extent of their participation. Although being part of a group of experts is always interesting for the experts themselves, problems may arise if many participants do not contribute but nonetheless have access to the results obtained by others. This situation was noticed by the "Upper Atmospheric Research Collaboratory" that was initiated in 1993 to serve the needs of a distributed community of space physicists at an observatory above the Arctic Circle. The pool of participants in data-gathering sessions was expanded, but new participants tended to be relatively passive (Finholt 2002).[13] Some participants observed only what took place in the collaboratory. In 1999, that collaboratory was dissolved. It is in this context that one should consider Ostrom's advice to limit access to the common by clearly defining the boundaries (Ostrom 1990, 90). When one sets up a collaboratory, eligibility for membership, and rights and responsibilities, must be defined. This stage in the information cycle is necessary because most scientists will not collaborate in the precollection or the postcollection phases if they cannot be sure their data will not be widely distributed if they have not yet published their results. It all comes down to gaining trust via the principle of *do ut des*, or "I give so that you give back"—data for data. The extra advantage of a collaboratory is that this kind of exchange is not a one-to-one agreement but rather a group agreement, by which social control within the group can contribute to the success of the collaboratory. This condition is essential to make the exchange of data useful in comparability of the data (kind of data) and data format (digital format, software used, etc.).

An essential aspect of the day-to-day workings of a common is the existence of mutual trust: a collaboratory cannot function if trust does not exist among the participants (nor can any other common). To achieve trust, the right institutional design is needed and the peers must be willing to participate and follow the rules. It is important to have a specific framework modelled on the specific needs of participants, including requirements to be a member of the collaboratory, the operating procedure, and the instruments to "punish" those who break the rules. Thus, an information commons, like a collaboratory, differs from a digital archive. A digital archive is a pool of information in an institutional framework, but those who donate the data are no longer involved after they contribute the data and do not obtain rights (via this submission) to the data that others have submitted, as is the case in a collaboratory. Their participation stops after they have pressed the button to submit or download. Although the submitter can set conditions for downloading and using the data, he or she is not necessarily involved in what use is eventually made of the data. Also, the data archive is not necessarily focused on the users but on the data that have been donated. The researcher depends on personal creativity to connect his or her research to what might be available on the hard disk of the data archive. Besides, those who manage the collected data enter the picture only at the very end of the flow of information, at the point when the data already exist. That said, it is not surprising that researchers are sometimes reluctant to deposit data. For many scholars, it is not a natural reflex to deposit their data (or even to consider doing so). Some reluctance to submit data can also be explained by the fact that researchers have to supply documentation (metadata) for using their data. Here, again, it is the stage of the information cycle when the data are submitted that is the stumbling block. Many data archives now advise researchers on how to set up their databases and how to document them, but this is not sufficient for achieving comparable data sets. To develop comparable data sets is certainly not a simple matter of adequately documenting what has been collected, it is also a matter of fine-tuning in advance what and how those data will be collected.

Management of a common requires a set of rules to organize interaction between the created goods and the users of those goods. In a collaboratory the rules relate to several aspects: (1) the way the goods should be created (the standards that should be followed), (2) the way they should be deposited (agreements on standards), (3) the way they can be used (use rules—e.g., not for commercial use), and (4) who can use them (access rules—e.g., only for use by

recognized peers). The members of the collaboratory can decide jointly on the eventual end use of the created goods. Regulation will, however, have little chance of success if there are no sanctions when rules are broken (Ostrom 1990, 90). One possible sanction might be exclusion from further use of the collaboratories' work and data. If people are not sufficiently committed to strive for an optimal outcome, the quality of the data entered or delivered may not be high enough. Commitment can be encouraged by assuring that benefits depend on participation/contribution to the common good of the collaboratory (preventive method) and by assuring that those who do not respect the rules (e.g., by entering low-quality data) are "punished." This policy can be enforced by random checks of the data and by more general social control among peers. Adding a "reputation component" (or peer review of the data) may be helpful to sustain a high-quality collaboratory.

As soon as the collaboratory claims its property rights in the form of a publication, and members have received academic credit for them, the collaboratory can choose to turn that information into the private property of the association of scholars and sell the data to third parties. Or it can decide to make the data public property. Thomas Finholt (2002) refers to several collaboratories in which agreements were signed to protect one another's rights. In a community of brain researchers, a formal covenant was drawn up that was signed by scientists as a condition of use of the collaboratory. It specified how community data were to be used, paying particular attention to the protection of younger researchers' interests. Elsewhere "rules of the road" described how public data were to be used, including rights of first publication and mechanisms for sharing credits. There are quite a few other instances when good regulation has led to the establishment of long-term intensive and highly productive cooperation among scientists.[14]

Collaboratories and the Information Cycle

We have already mentioned that some of the advantages collaboratories have over data archives are because of the early interventions in the information cycle that go with setting up a collaboratory: such interventions are established right at the beginning of the cycle, after peers have exchanged ideas about a potential data collection project, but before the actual collection. So goes the theory, but in practice researchers may well have already started collecting data, as in the case of the Historical Prices and Wages project: several researchers had a data set on a particular country and period and brought them together in a data hub. A comparison of the information cycles in a "normal" situation may clarify the actual differences: there, the researchers do not set up a collaboratory or a collaboratory information cycle. In the first case (see figure 1), the researcher transforms an idea into research questions and designs a database that can contain the data necessary to answer specific questions. Increasingly, data archives offer researchers assistance in this stage, but overall, researchers largely work on their own. In subsequent phases, researchers collect the data and process them, after which the results may be published. The right side of the figure shows that in the course of data collection researchers sometimes use data hubs for retrieving data. When the results are published in a DAP journal, the researchers are also requested to deposit their data. All journals with such a policy do not use internationally accepted data-archiving methods or metadata protocols, but simply put the data on an accessible Web site. In an increasing number of cases, the individual researcher also finds his or her way to the data archive. At the end of the information cycle, the researcher sometimes submits the entire data collection to the data archives, allowing third parties to use it for new research, which may lead to a new start of an information cycle.

In the ideal collaboratory information cycle, illustrated in figure 2, researchers discuss and fine-tune their ideas about research questions from the start. They exchange ideas about the necessary data and data format and set up a database format that can include data in different but comparable formats. Thereafter, data collection can start. Data processing is also a collective process, because this may raise new questions about the data and is usually the basis of collective publications in which the members figure as coauthors.

The Complementarity of Collaboratories, Data Archives, and DAP Journals

The collaboratory formula has tremendous potential to solve some difficulties experienced by data archives to the particular stages in the information cycle in which collaboratories, data archives, and DAP journals participate. There is no point in trying to merge all the functions of these three initiatives. On the contrary, although there is always a certain overlap, researchers (in collaboratories), archivists (in data archives), and editors (of DAP journals) are different stakeholders within academia, and it would not be wise to confuse or conflate these different positions. On the path to an optimal and durable exchange of data and ideas, data archives, DAP journals, and collaboratories can combine strengths by reciprocally offering the services in which they are best qualified.

Complementarity should be the key word on the path to optimal exchange of data and ideas. Each initiative faces three similar problems: they must (1) convince authors to make data available, (2) ensure that the data are of high quality, and (3) solve the problem of data storage. Each form of data exchange scores differently. Collaboratories can solve these problems to a certain degree, but in many ways the two other initiatives in data collection and scholarly communication can play a complementary and essential role.

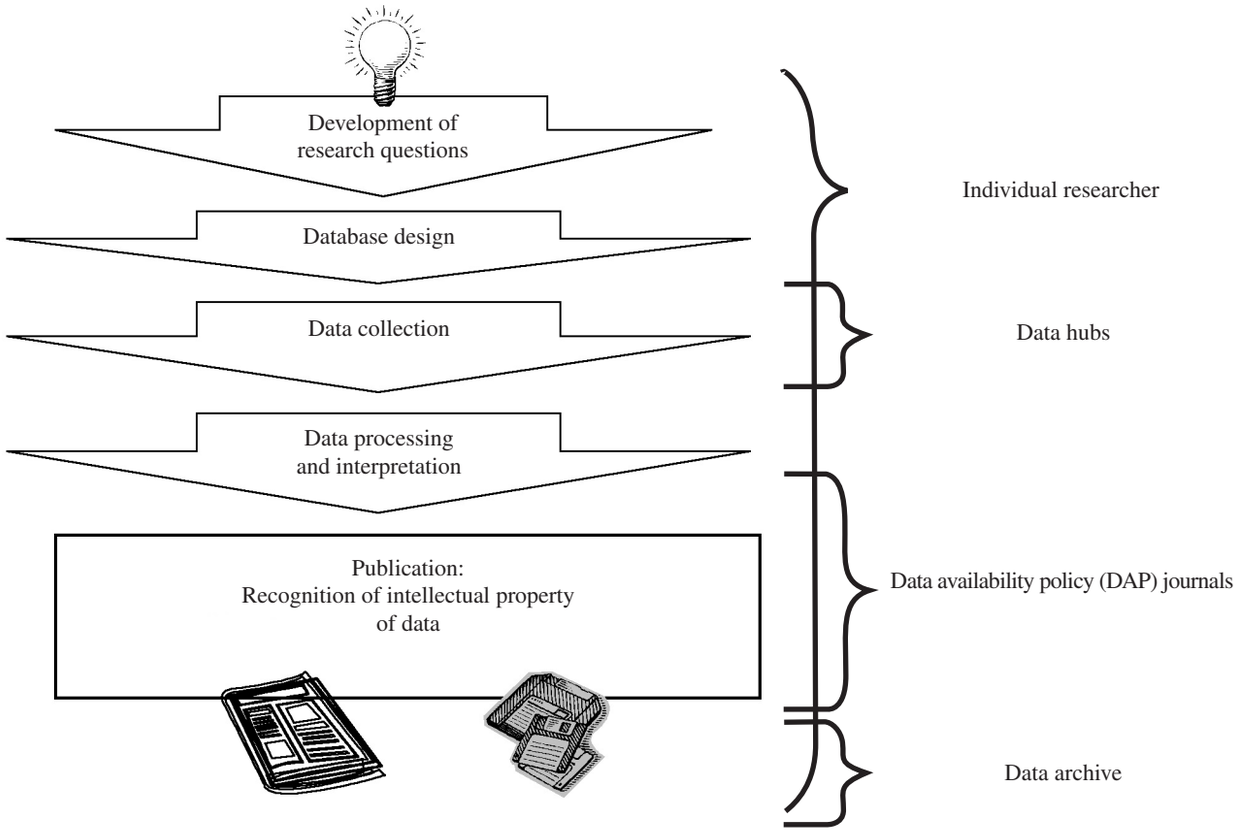

FIGURE 1. Normal (individual) information cycle from idea to publication of data set.

Each initiative has its own strengths; in that sense, it can contribute to the composition of useful global data sets by offering solutions to the three problems. In solving the problem of enhancing availability of data for secondary users (exchange of data), we can assume that DAP journals offer the most powerful solution: if data availability is set as a requirement for publication, authors are very likely to be willing to provide their data, considering the importance of academic publications. As the data are made available to the wider public only after publication, the authors can still benefit in full from their data-collection work. Collaboratories have the advantage in that if an author participates, he or she also obtains other researchers' data (before publication). "I give so that you give back" is still a good incentive for cooperation among researchers. But apart from the promise that a participant will be able to use another researcher's data before publication, there is little incentive to make data available to others, as in the case of DAP journals. Submitting data to a data archive is still not a natural reflex for researchers, even though some funding bodies make it a requirement after the termination of a research project. The extent to which researchers are willing to share largely depends on their good will. Given what has already been said about these three approaches, figure 3 is self-explanatory.

The extent to which the three initiatives can solve the second problem depends on the stage attained and the intensity peers can use to check the data. The earlier in the information cycle and the more effective the checks are, the better the quality of the research results. In the case of the DAP journals, a select group of peers can look at the data immediately before publication, at the stage when an author offers an article for publication. Soon after publication, the data become available to anyone who wants to check the published results. Data archives offer this opportunity only after the author has decided that the data can be used by others. Logically, this would not be before publication, and if the author so decides, it could be years after collecting the data has been completed. Data archives and DAP journals thus offer an inadequate solution to peer review of data before publication. Collaboratories, however, start at the beginning of the research process to exchange ideas about the data-collection method. Building data hubs that are accessible virtually is a collective peer effort linked to a collaboratory. Peers can correct one another during the entire research process, up to publication, and if necessary, after that.

The third problem of long-term storage is where data archives score the best. As this is a new process, it is still

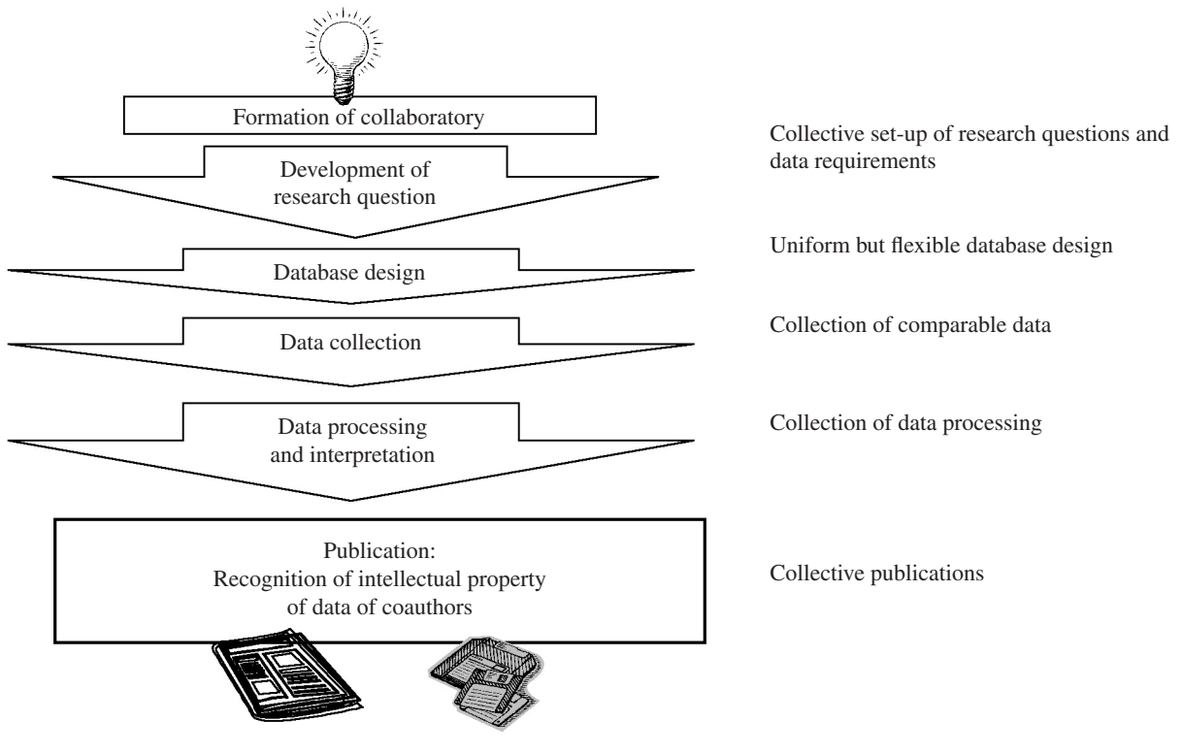

FIGURE 2. Collaboratory information cycle.

Overview of problems faced by collaboratories, data archives, and DAP journals on the path to exchanging data, ranked in order of their success in solving the problems

PROBLEM 1. ENCOURAGING WILLINGNESS TO SHARE DATA:
Convincing collectors of data to make their data available to others

1. Data collected via DAP journals (no sharing, no publication)
2. Collaboratories (sharing within a well-defined group)
3. Data archives (relying on external institutes to convince researchers to deposit data)

PROBLEM 2. ENSURING THAT THE DATA ARE HIGH QUALITY:
Quality control of data via peer review

1. Collaboratories (pre-collection agreements on data collection)
2. Data collected via DAP journals (peer review during publication process)
3. Data archives (post-collection, if someone finds it worthwhile to check)

PROBLEM 3. KEEPING THE DATA ACCESSIBLE, USABLE, AND USEFUL:
Storage of data for further use

1. Data archives (their core business)
2. Data collected via DAP journals (temporary storage in Web site journal)
3. Collaboratories (cease to exist after a while, data can be lost thereafter)

FIGURE 3. Overview of problems of data archives, DAP journals, and collaboratories.
Note. 1 = best solution, 3 = weakest solution.

unclear how DAP journals will deal with data archiving. They place data on their Web sites, but no one, including the editors of the DAP journals, knows how long it will be there and what kind of data system will be used. If these journals intend to make the data available for a long time, it is unlikely the archivist will have the expertise to keep the data available in the long run. In this sense it is not a long-term solution to effective data storage. The same is

true for data collaboratories: they concentrate on creating good data sets able to answer particular research questions, which they hope will also result in major publications. But what happens to the data afterward is unclear. In this sense, neither DAP journals nor collaboratories are adequate to solve the third problem. The best expertise for data storage is available in data archive institutions. It would therefore be optimal if, at the end of their information cycle, collaboratories could offer the information they have gathered in their hubs for storage. The same goes for DAP journals.

The optimal solution would be close cooperation among data archives and the other forms of scholarly communication from the start of research projects. Optimal efficiency of (human) resources could be achieved if collaboratory participants applied the expertise of data archivists when setting up data formats for their data hubs. This early cooperation between data archives and collaboratories could be advantageous for both parties: collaboratories can benefit from the expertise of data archives, and data archives can expand their data collections with large amounts of high-quality data. This also allows them to anticipate the needs of historians via close collaboration with collaboratories, without hampering the scientific process. Similarly, data archives could set up a cooperative arrangement with DAP journals by offering support for the data storage of a submitted article. Figure 4 illustrates how these different parties can interact at different phases of publication.

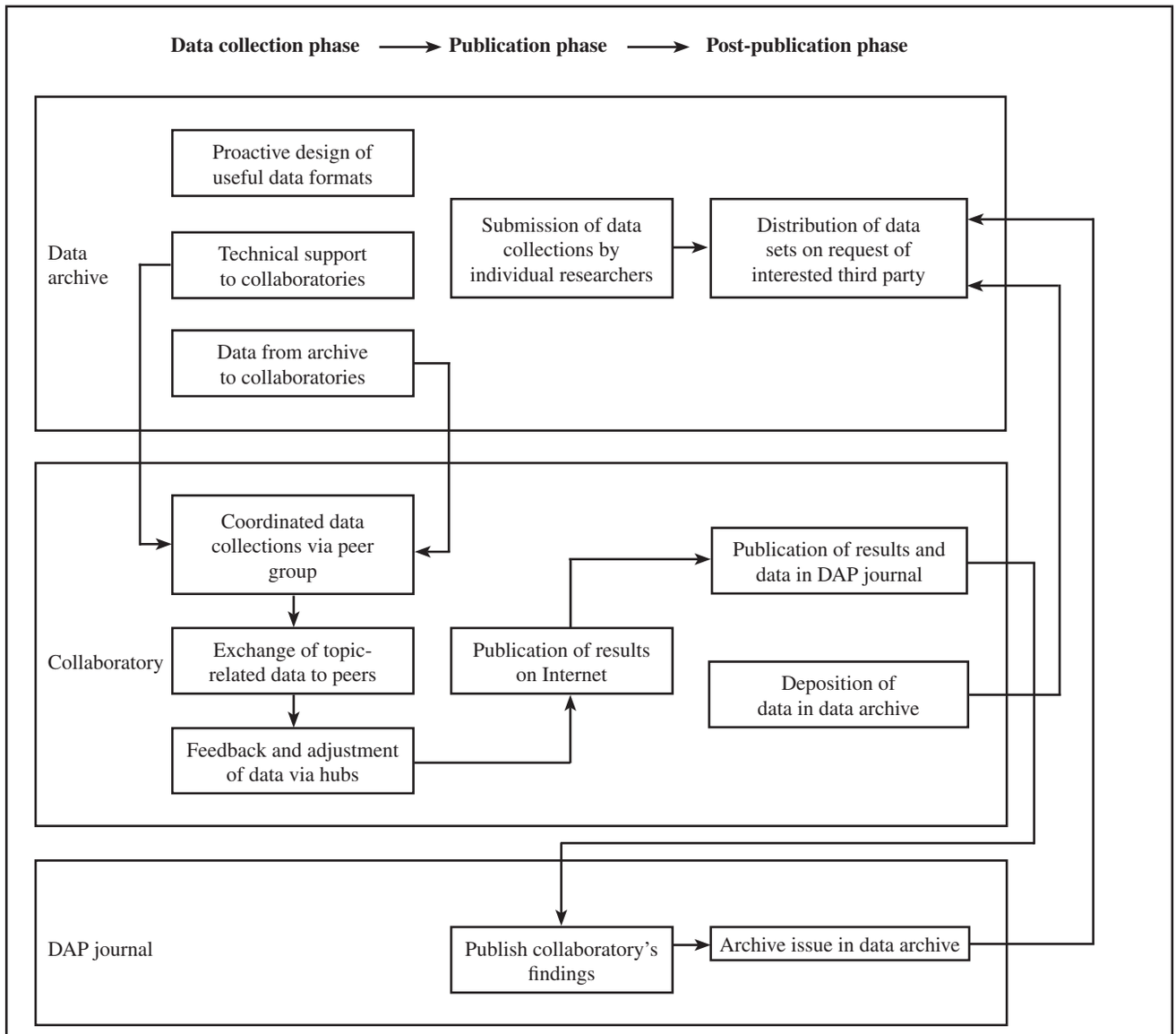

FIGURE 4. Possible forms of cooperation within and among data archives, collaboratories, and DAP journals.

It is clear that the various forms of scholarly communication and exchange have advantages and disadvantages. Linking, and to a certain extent, merging the activities of data archives, collaboratories, and DAP journals can lead to optimal use of available resources in terms of infrastructure and labor. Via the collaboratories, major researchers can convince journals in their fields of research to implement a data-availability clause in their policy. Such a clause will increase open access to scientific data in general and enhance awareness of the need for peer review of data, as well as for publications, among researchers.

Conclusion

In this article, we advocate establishing new collaboratories in the social sciences and humanities, especially in history. This new model of cooperation and exchange among scientists is largely unknown to this sector of the academic community, although it can prove to be a rewarding method for surmounting the data-collection problem required when engaging in the big history of the future. The increasing importance of global research, as in economic history, makes it impossible for a single researcher alone to collect all the necessary data. Comparative research questions require so much data that the input of several researchers has become absolutely necessary. Data archives alone cannot adequately solve this problem.

As to the benefits of a collaborative approach, the main difference between collaboratories and data archives is that unlike data archives, a collaboratory enhances communication among participants. Because interaction is possible, comparison of the data and if necessary adjustment/correction are possible before publication, enabling data collection to be more efficient for the collectors and the later users.

The main difference in functioning is the extent of participation among the parties involved. Participation enables an intensive process of peer interaction and review that would be impossible in the context of a data archive. But the risks that such data sharing and communication entail need to be addressed by developing specific institutional frameworks. Data archives avoid many of the risks of exchange because exchange starts only after the author of the data is convinced he or she has already obtained all the desired benefits. But at the same time, the time until the data is made public is lengthened, making prepublication peer review impossible. Both these outcomes do not help to advance science. Nevertheless, cooperation among (historical) data archives, collaboratories, and other forms of interactive peer review such as DAP journals could be extremely beneficial for academic research. In some fields of history, especially the more quantitative niches that rely heavily on large data sets for statistical and econometric analysis, collaboratories are becoming more popular. Large areas of historical research can benefit from this new approach to data collecting, especially now that global history seems to be the wave of the future.

NOTES

1. "Big history" is used here to refer to historical research that concentrates on the specific histories of countries and continents and on comparisons between those different parts of the world. Global history is used to designate historical research on the links and interactions among different parts of the world and the emergence of global systems of change (e.g., World Systems Analysis).

2. The work of Angus Maddison, a well-known professor of historical economic analysis formerly at the University of Groningen (available on the Web site of the Groningen Growth and Development Centre [http://www.ggdc.net/]), is the best example of such a hub on economic history. This kind of international comparative work directed at establishing global data sets has been further developed in different directions, such as the history of prices and wages, historical national accounting, and the study of the biological standard of living.

3. The data availability policy (DAP) of the *Journal of Economic Policy* (2008) states: "It is the policy of the *Journal of Political Economy* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Authors of accepted papers that contain empirical work, simulations, or experimental work must provide to the *Journal*, prior to publication, the data, programs, and other details of the computations sufficient to permit replication. These will be posted on the *JPE* Web site. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the requirements above cannot be met. After acceptance, authors are expected to send their data, programs, and sufficient details to permit replication, in electronic form, to the *JPE* office. Complete instructions will be provided to the author with the acceptance letter." See http://www.journals.uchicago.edu/JPE/datapolicy.html. The DAP of the *American Economic Review* has a similar content. See http://www.aeaweb.org/aer/data_availability_policy.html.

4. The rapid development of high-quality and easily applicable software as in the Open Journal Systems (http://pkp.sfu.ca/?q=ojs) make it increasingly feasible for scholars to distribute their knowledge at high speed and low cost.

5. Pupils could have virtual access to a virtual microscope that allows them to examine the growth of small creatures (Bugscope 2007). The Bugscope project provides a resource to classrooms so that they may remotely operate a scanning electron microscope to image "bugs" at high magnification. The microscope is remotely controlled in real time from a classroom computer over the Internet by means of a Web browser.

6. The Science of Collaboratories Project has several different names for different types of collaboratories, according to their classification: distributed research center (Biocore, http://www.ks.uiuc.edu/Research/biocore/), virtual learning community, virtual community of practice, shared instrument (Earthscope, http://www.earthscope.org), expert consultation (TeleMedicine, http://www.telemedicine.arizona.edu).

7. Besides their diverging objectives, the scale of collaboratories can also vary significantly. Most collaboratories that manage to survive for at least a few years are on a grand scale, such as the Human Genome Project.

8. We will not go into depth here about the types of software that can be used for setting up a collaboratory, because there are many different options as more are increasingly becoming available. At the moment, one of the commercial software packages that is often used to facilitate exchange and cooperation between researchers is Microsoft Sharepoint.

9. See the Science of Collaboratories Web site: http://www.scienceofcollaboratories.org/.

10. Other advice is to "get each individual participant on board . . . gear up for major technical challenges . . . put enough resources into project management . . . talk the same talk . . . hold your course." (Bender 2004).

11. In literature, the term *knowledge common* is also used. However, considering the goods that are exchanged in a collaboratory, information common seems more appropriate.

12. This takes place although initiatives such as the "Creative Commons" (http://www.creativecommons.org) also offer licences that allow an intermediate solution whereby only a number of rights are reserved.

13. Here, Thomas Finholt referred to the UARC and Space Physics and Aeronomy Research Collaboratory (http://www.si.umich.edu/sparc).

14. See http://www.iisg.nl/hpw/ for more information.

REFERENCES

American Economic Review. 2008. "Data availability policy." http://www.aeaweb.org/aer/data_availability_policy.html (accessed May 19, 2008).

Agarwal, D. A., S. R. Sachs, and W. E. Johnston. 1998. The reality of collaboratories. *Computer Physics Communications* 110:134–41.

Allen, R. C. et al. 2005. Wages, prices, and living standards in China, Japan, and Europe, 1738–1925. International Institute of Social History. http://www.iisg.nl/research/jvz-wages_prices.pdf (accessed May 15, 2008).

Bender, E. 2004. Rules of the collaboratory game. *Technologyreview.com*, November 22. http://www.technologyreview.com/Biztech/13899/ (accessed May 15, 2008).

Bugscope. 2007. http://bugscope.beckman.uiuc.edu/ (accessed May 15, 2008).

Endersby, J. W. 1996. Collaborative research in the social sciences: Multiple authorship and publication credit. *Social Science Quarterly* 77:375–92.

Finholt, T. 2002. Collaboratories. *Annual Review of Information Science and Technology* 36:465–91.

Global Price and Income History Group. 2004. http://gpih.ucdavis.edu/ (accessed May 15, 2008).

Hardin, G. 1968. The tragedy of the commons. *Science* 16:1243–48.

Historical Sample of the Netherlands. 2006. HSN goal. http://www.iisg.nl/~hsn/abouthsn/goal.html (accessed May 19, 2008).

International Microdata Access Group. 2002. http://www.prdh.umontreal.ca/IMAG/index.html (accessed May 15, 2008).

International Institute of Social History. http://www.iisg.nl/iish.php (accessed May 15, 2008).

Journal of Political Economy. 2008. Data policy. http://www.journals.uchicago.edu/page/JPE/datapolicy.html?cookieSet=1 (accessed May 19, 2008).

Lunsford, K. J., and B. C. Bruce. 2001. Collaboratories: Working together on the Web. *Reading Online* 45:52–58.

McDaniel, S. E., G. M. Olson, and J. S. Olson. 1994. Methods in search of methodology: Combining HCI and object orientation. In *Proceedings of CHI '94*, S. E. McDaniel, G. M. Olson, and J. S. Olson, ed., 145–51. New York: ACM Press.

Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action*. New York: Cambridge University Press.

———. 1992. The rudiments of a theory of the origins, survival, and performance of common-property institutions. In *Making the commons work: Theory, practice and policy*, D. W. Bromley and D. Feeny, ed., 293–318. San Francisco: ICS Press.

Weinberg, A. 1961. Impact of large-scale science on the United States. *Science* 134:161–64.

Wulf, W. A. 1993. The collaboratory opportunity. *Science* 261:854–55.

ADVERTISING SPACE NOW AVAILABLE

Public Affairs & History

Advertise in Heldref Publications' Public Affairs & History Package in 2008 and save up to **40%.** With these publications you will reach a select, highly targeted audience of college and university educators. The package includes the seven titles below with a combined circulation of over 13,000.

- **World Affairs:** A Journal of Ideas and Debate
- **Historical Methods:** A Journal of Quantitative and Interdisciplinary History
- **Demokratizatsiya:** The Journal of Post-Soviet Democratization
- **Perspectives on Political Science**
- **Asian Affairs:** An American Review
- **History:** Reviews of New Books
- **Current:** Required Reading Recommended by Leading Opinion Makers

Advertising Package Rates:

Advertise in all 7 publications and save 40% (limited time offer)
Full page black and white: $1995
1/2 page black and white: $1200
Deadline for summer issues: September 10

Size Requirements:

Full page: 8 inches wide x 11 inches high
1/2 page: 7 inches wide x 4 1/2 inches high

For more information contact:
Sally Wright at 540-592-3634 or swright@heldref.org