# Do ut des.[1]

## Collaboratories as a "new" method for scholarly communication and cooperation for global and world history

**Tine De Moor & Jan Luiten van Zanden**

**Draft article/In process of publication**

Tine.Demoor@let.uu.nl

Janluiten.vanzanden@let.uu.nl

**August 2006**

[1] Latin for "I give so that you may give"

## *1. Introduction on data sharing and "big" history*

Within the historical sciences, there is a growing need for more internationally oriented data collection in order to answer the research questions that are being increasingly posed in the expanding field of global and world history, or "big history" in short.[2] Global questions require global data. In order to encourage efficient use of the already collected data and the cooperation between scholars world-wide, new methods of data sharing and scholarly communication need to be designed. Developments in information sciences have enhanced the possibilities in data collection (through digitisation and the creation of databases), thus allowing historians to process a growing amount of data. In the past years "data hubs" have filled this need to a certain extent, at least for economic historians.[3] However, there is more needed than just raw data, to practice global and world history on a scientific basis. In addition to the indispensable metadata, other knowledge is necessary to work with data from sources that originated in different parts of the world: knowledge about the context wherein the information was collected (e.g. the economic situation (such as a crisis) can have influenced the results), the method of collection (e.g. moment of the year censuses were performed) and the possible strains on the application of those data (e.g. the difference in consumption pattern - the composition of "food baskets" - between China and Europe when calculating living standard).[4] Intensive communication and interaction between 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 which is necessary to understand the data. Moreover creating such large datasets usually takes considerable time. In order to avoid that initiatives such as the economic history data hubs and the data itself disappear with their initiators, a certain degree of institutionalisation can help to safeguard the future of the data. Solid networks of collaboration need to be designed not only for the direct benefit of their participants but also for future researchers in these fields of expertise, for whom –due to expanding knowledge- the writing of global history will most likely become even more complex.

Over the past two decades, the natural sciences have established so-called "collaboratories" to fulfil these needs. A collaboratory is a "laboratory without walls", where scientists are connected to each other, to instruments, and to data independent of time and location, hereby creating a virtual community of peers. As will be explained, sharing data can enhance productivity but it also involves certain risks. When collecting and exchanging data among researchers –within a collaboratory or a data archive- one faces a set of risks that can only to a limited degree be prevented by technology. Although this is often overlooked, intensive cooperation cannot be durable without a solid institutional framework that anticipates problems of collective action –such as freeriding- rather than dysfunctional computer networks. After having given some theoretical background about the possible risks of collaboratories, we will show how research on the functioning of what is now being described as "information commons" may offer some guidance in the set up of such a framework. A clear and instructive institutional design can create the right incentives for researchers to cooperate and contribute to their common good, in casu high quality scientific output.

This article introduces the concept of collaboratories into historical research and explains how this form of distributed collaboration complements existing forms of data collection and distribution, such as in centralised historical data archives. Traditionally,

[2] World history is considered here as the historical research that concentrates on the specific histories of countries and continents and on the comparison between those different parts of the world. Global history stands for the historical research on the links and interactions between different parts of the world, and the emergence of global systems of change (e.g. World Systems Analysis).

[3] The work of emeritus professor Angus Maddison (RUG), whose work is also available on the website of the GGDC (www.ggdc.net) is the best example of such a hub on economic history. Recently this work has been further developed in two directions, one on contemporary measurement of output and productivity, and the other on historical national accounts. See herefore: [Referentie afmaken!]

[4] See the discription of these difficulties in the paper by Moll Murata, van Zanden, Ma, Allen

historical data archives do not have the institutional design to initiate and support intensive interaction between researchers. For a researcher, participation in a data archive is usually a one-way trajectory. Typically, after the submission of the data, a researcher usually has no more contacts with the data archive nor with the future users of his deposited work. There 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. Not only does it become possible to amass data on the same theme, thus creating new opportunities for comparative research, the interaction with other researchers also offers more background information on the data and its possible uses. In this article we will explain that although data archives, collaboratories and other new initiatives such as peer-reviewed journals with a Data Availability Policy have different objectives and outcomes, they can play complementary roles at the beginning and end of specific information cycles.

## *2. Where do all those data go? The need for more cooperation among scholars in the social sciences and humanities*

Since a few decades, digitising data has become the main method of data collection and storage for historians. Considering the vast amounts of data that have been collected since the 1980s, the scientific community and in particular that of the social sciences and humanites has only to a very limited degree benefited from one of the most tempting advantages of digitisation –namely the exchange and re-use by other parties and the pooling and sharing of very large bodies of data. Data-archives have over the past decades been set up to collect and preserve those data sets as much as possible. However, there are still two issues that prevent data-archives from becomes central players in scientific circles. First of all, convincing those with the data to submit them for other interested parties seems to be very cumbersome. Researchers still wrestle with the idea that others may benefit from using their painstakingly gathered datasets and –maybe even more importantly- fear that by making their datasets public, mistakes in the processing of their data and their (to be) published results might be discovered. Giving the data out of hands, gives the collector a feeling of loss of control over the so carefully collected data.

A second issue is the link between the data archive and other parties potentially interested in the already collected data. In some cases, data archives have managed to gather large amounts of data, but possible "second-hand users" do not always find their way easily to those virtual archives. Data archives offer potentially diverse but mostly no direct application possibilities as the collectors of data may not feel the need to point this out to the rest of the scholarly community. It is agreed upon that it is important to store data for the future, but how to make them also of high value for future researchers? Data quality plays important role in the relationships between the three involved parties - data collectors, secondary users and data archives-: the collectors fear that their data might be discovered as insufficient, the secondary users fear that the data might not be trustworthy. If the distance between these parties could be made smaller, higher quality of the data could also be obtained. Data –as much as their results-to-be-published can improve significantly by peer-reviewing. In the event of global comparisons of data, such peer review and "peer support" even becomes an essential part of the scientific process (see higher). As data archives are not topic-specific, it cannot be expected that they also offer the necessary expertise to comment on the data or even evaluate what is being offered for storage. That is a role for the scholarly community to play. The quality depends entirely on the self-criticism and good-will of the depositor of the data. The meta-data that come with the data sets or the publications on the basis of the data are essential to understand what has been done with the information in the databases but are often not sufficiently specific for other interested users.

Recently, some journals have integrated so-called "Data Availability Policy"-clauses in their policies which should allow readers to check the validity of the authors' claims on the

basis of the data sets that were used. For example the "Journal of Political economy"[5] and the "American Economic Review" both require in their Data Availability Policy that "prior to publication, the data, programs, and other details of the computations sufficient to permit replication" have to be submitted to the journal in order to enhance external peer review, of those who are not included in the prior-to-publication reviewing process. This development – which is clearly a consequence of the possibilities and popularity of the internet- is part of a new trend in the academic world towards more openness and freedom of information, as is for example also visible in the increasing establishment of e-journals, of which some are also open-access.[6] Moreover, e-journals offer the possibility to link-up content (the article) with "click-on-demand proof", readily available at the journal's website.

Admittedly, journals with a Data Availability Policy are a novel trend, and as such it should not surprise that it does yet not seem to catch on in the historical sciences: none of the present day publishing devices in history include the possibility to check the data linked to an article, probably often to the relief of the author. Some journals offer data in the form of printed appendices but as these do no longer have the same structure and flexibility of the original database, third parties are probably not very prone in checking the published results. Although these new developments in scientific publishing are definitely positive, they do not guarantee access to all the data that were gathered, nor do they solve problems like diverging data formats, or is the future of the original digital data sets secured. How long will they be available online? And: how well are they being stored and in what format? Moreover, there are mostly no provisions for metadata as offered by historical data archives. For the sake of large-scale international comparisons that global and world history require we need to go a step further: data should not only be prevented from being lost or made public for peer review after their usage (data archives) or publication (journals), there is a need for more interaction and exchange of data between researchers before the publication stadium if one wants to construct large high-quality international datasets. It is clear that, although they both perform absolutely necessary functions for the distribution and preservation of data, archives and e-journals both miss some essential attributes to make this happen. Here the so-called collaboratories come into the picture.

Before explaining the extra advantages collaboratories can offer in comparison to data archives and other data initiatives, a clearer description of "collaboratories" will be given. Creating a collaboratory is more than getting in touch with fellow researchers and decide "to do something together", to collect data collectively as is done already in data-hubs. The goal of a collaboratory is to provide complete location-independent access to instruments, data collection and analysis resources, as well as to other researchers in a certain field.[7] The idea behind this is that exchange of data and intensive cooperation between scientists is in many ways beneficial for those who participate: those who participate can benefit from the data collection and from the intellectual (and for academics thus also professional) advantages "collective thinking" may have. As well in data collection as in the actual (collective) thought processes the total outcome will be more than the sum of its parts. One can assume that -as more knowledge becomes available over time- the validity of this assumption also becomes greater over time too. In this sense society as a whole may benefit from the enhanced output of science to the investment it makes in science. However,

---

[5] The Data Availability Policy of the *Journal of Economic Policy* 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." For more details see, http://www.journals.uchicago.edu/JPE/datapolicy.html. The Data Availability Policy of the AER has a similar content. See http://www.aeaweb.org/aer/data_availability_policy.html [Deze voetnoot is grotendeels identiek aan noot 25]

[6] 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 feasibly for scholars to distribute their knowledge at high-speed and low costs.

[7] Agarwal, Sachs, and Johnston, 'The reality of collaboratories'

informal agreements only will not lead to a sustainable cooperation between researchers– not now nor in the future. An institutional design that offers a stable but nevertheless sufficiently flexible environment for the collaboratories' participants is necessary. Research on the functioning of the so-called information commons can offer some guidance here.

### *3. What is a collaboratory?*[8]

The term "collaboratory" made its appearance in the scholarly community in the late 1980s. A collaboratory is "a laboratory without walls, where scientists are connected to each other, to instruments, and to data independent of time and location".[9] It can be considered 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, artifacts and tools required to accomplish research tasks". The collaboratory-triangle (see underneath) shows the interaction between information, researchers and facilities as the core-business of collaboratories. Depending on these interactions, different objectives can obtain the upper hand. Collaboratories can provide communication environments and tools for scientists, can serve as a communication tool for students,[10] allow the collection of data, give online access to data to members and –in some cases- non-members of the collaboratory, and create the means to share scientific instruments within research or learning communities.[11] It is clear that social scientists, and in particular historians, will benefit more from data collection than from joint use of instruments.[12]

Notwithstanding the many variations in objectives the currently existing collaboratories have, they all have several common features:[13]

- *Boundary-crossing*: a collaboratory is first and foremost a tool to bridge gaps and distances of (a) geography, by providing international access through the Internet; (b) time, by supplying both synchronous and asynchronous communication technologies; (c) institutions, by allowing groups access to tools and materials of common interest; and (d) disciplines, by enabling the participants to decide what resources are most relevant to a topic, without regard to traditional understandings of what constitutes a particular discipline.
- *Shared inquiry*: participants do not only share common goals in e.g. data collection but also a common set of problems or issues that interest them and that they study in depth
- *Intentionality*: a collaboratory is a joint venture; there is a shared consciousness of the status of it's website as a mutual project. There is a "tipping point" which leads to the

[8] Wulf, 'The collaboratory opportunity.': 854-855
[9] Finholt, 'Collaboratories': 647
[10] e.g. pupils of schools have virtual access to a virtual microscope that allows them to examine the growth of small creatures, see Bugscope, http://bugscope.beckman.uiuc.edu/. 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 using a web browser.
[11] The Science of collaboratories project has come up with several different names for different types of collaboratories. There are about six different types according to their classification: distributed research center (like Biocore: http://www.ks.uiuc.edu/Research/biocore/), virtual learning community, virtual community of practice, shared instrument (like Earthscope, http://www.earthscope.org , expert consultation (like TeleMedicine, http://www.telemedicine.arizona.edu, community data systems (see e.g. http://research.umbc.edu
[12] 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, as in the Human Genome Project, or the ATLAS Project at the European Organization for Nuclear Research (better known as CERN). ATLAS coordinates 1,800 particle physicists in 34 countries. They also can marry formerly separate threads of research. For instance, the Space Physics and Aeronomy Research Collaboratory, based at the University of Michigan, gives researchers simultaneous access to both observations and predictive models, so they can predict "space weather" (such as the geomagnetic storms that produce aurora borealis events) and then see what actually happens.
[13] These features are partly based on Lunsford and Bruce, 'Collaboratories: Working together on the Web.'

critical mass awareness needed before a collective site is perceived by its members as a collaboratory.

- *Active participation and contribution*: the success of a collaboratory is to a large degree decided upon by the extent that its members use and add to its resources
- *Members-only:* although the data as collected by a collaboratory can become freely accessible in due time (mostly this means after publication of the research results), participation in a collaboratory is usually 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 its 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
- *Limited in time*: collaboratories are set-up to reach certain research goals (creating a dataset, answering certain research question….). Once these goals have been reached, they are dissolved, though in some cases their results remain available via the collab-website.

The development of collaboratories towards this common set of features stems from a trend that has developed over the last half century towards large-scale projects or the so-called "big science" which leads to a higher need for collaboration between scientists, not only by beta-sciences but also by the social sciences, including the humanities (cf. the increasing importance of global/world history).[14] This need for collaboration stems from the increasing amount of information that leads scientists every day a bit further away from the traditional "Humanist Ideal". The specificity of each country, region or place –in terms of politics, culture, language, etc.- makes it however impossible for a single historian to collect data on a world wide basis with the same accuracy. The beta-sciences have the advantage that they also have a common scientific language, be it –to put in simple- a chemical or a mathematical formula. The trend of the past decennia towards global or world history, necessitates the need for more collaboration among historians, and other members of the humanities at large.

### *4. The institutional design of a collaboratory: collaboratories as commons*

The project "Science of collaboratories", hosted at the University of Michigan, has so far identified more than 200 collaboratories (in which very few social scientists are participating)[15] reflecting the ambitious challenges of today's science, the extremely expensive instrumentation that it often requires, and the availability of very-high-capacity networks and computing resources. Often the collaboration enables research that due to its magnitude and scale just could not be done otherwise. Considering the advantages of scale and other benefits that are to be reaped from collaboration, this high number of collaboratories should not be surprising. However, although new collaboratories have been formed continuously since the 1990s and many have resulted in high quality scientific results, many have by now also ceased to exist. The end of these collaboratories however does not mean that the formula does not work. Collaboratories usually terminate their own existence when the immediate need for the researchers' cooperation is no longer present: the data were collected, the specific research questions they had to answer had been published and the researchers proceeded to follow their individual interests. Collabs are topic-oriented research networks that often serve an ad hoc need within certain scientific circles. In the process towards achieving their goals they face however several problems, which may be solved by setting-up an appropriate institutional environment. How this framework should look like is however an other issue: should it be imbedded in a larger, steering institution, like a

[14] Weinberg, 'Impact of large-scale science on the United States'. For big science and higher need of cooperation in social sciences: Endersby, 'Collaborative research in the social sciences: Multiple authorship and publication credit'
[15] See the "Science of collaboratories"-website: http://www.scienceofcollaboratories.org/

data-archive or should it be autonomous? Should it be open access or should the membership be restricted? What other rules should be created in order to prevent freeriding?

Considering the features of a collaboratory, as described in the previous part of this article, it becomes clear that this form of scholarly communication and data sharing, is quite similar to a much older institution, the commons. Commons are –in the social sciences jargon also referred to as Common Pool Resources (CPRs)- are "natural or man-made resources sufficiently large that it is costly to exclude users from obtaining substractable-resource units".[16] The historical commons of Europe were formed from the bottom-up, by the villagers who needed the pasture land to make their agricultural system running, however in accordance with the local powers. The commoners limited access to members-only, designed use and management rules that tried to prevent overharvesting and secured a sustainable future. Though their actual functioning is receiving increasing attention, it is in particular their dissolution that has drawn most attention, as a potentially important factor in creating the destitute, cheap labour force that was necessary to make the Industrial Revolution possible. Within non-historical circles the theory and practice of today's common pool resources has been the inspiration for understanding new forms of virtual or information or knowledge commons (each term has a slightly different term), such as the internet. By considering the risks that collaboratories run, it is shown in this article that collabs can be considered as knowledge commons. By establishing this link, the knowledge that has been gathered in circles of commons-researchers might become of use to set-up adequate institutional structures to make collaboratories work.

Very little research has been done on the requirements to build a successful collaboratory. Besides the advice to honour some very general rules such as "Make sure your working community is ready" and "tackle big questions" there are so far no recipes for setting up a successful collaboratory.[17] We can however find much inspiration for designing well-functioning institutional frameworks from the research on information/knowledge commons. At the basis of this field of study is the question how one can organise/regulate the use of goods that are held in common. The common-property-debate started of in the 1970s as a reaction to the "Tragedy of the commons"-article by Garret Hardin (1968, Science) and focussed the first few decades only on the management of natural resources. Hardin claimed that due to the greedy nature of human beings it was simply impossible to manage a good in common in a sustainable way. Since members of a common property regime only think of the short term advantages they can obtain from a common good, common property regimes are bound to end in ruin. In the case of natural resources this comes down to overexploitation and eventually the disappearance of the resource. Hence, a tragedy of the commons. The only solution to prevent such a tragedy was according to Hardin by privatising the good or making it public property, whereby the use of the resources would be arranged by the state. Over the past 25 years, researchers from all corners of science have countered this pessimistic view on common property regimes by describing and analysing a wide range of commons all over the world. They showed that common property can work, if a certain institutional design is followed. Elinor Ostrom gave a first overview of the characteristics of such a "design" in her seminal work "Governing the commons" in 1990; Ostrom's design principles can be used as a guide when developing a framework for collaboratories. We will use some of those principles here to explain what should be taken into account when designing such an institutional framework. During the 1990s the debate has also broadened to other forms of common property than natural resources; researchers have started to apply the theoretical results of research on local natural resources on what has been termed the "global commons" (water, air, etc.) and on the virtual commons. This attention for the virtual commons was a consequence of the growing popularity of the World Wide Web, and of other new forms of electronic communication in general and among scholars in particular (e.g. e-journals).

---

[16] E. Ostrom in Bromley and Feeny, *Making the Commons Work: Theory, Practice and Policy*.

[17] Other advice is "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, 'Rules of the collaboratory game'
http://www.technologyreview.com/articles/04/11/wo_bender112304.asp?trk=nl

Collaboratories can be considered as another example of virtual commons. We will briefly touch upon the characteristics of collaboratories as virtual commons before explaining how this approach can contribute to preventing problems. The most characteristic problem users of common property face is scarcity: there is a limited amount of the good and multiple people want to use these goods. In the case of the collaboratories, the scarcity lies not exactly within the availability of the data, but within the professional benefits that are to be reaped from using these data, more precisely the publication of the results based on these data. If participants of the collaboratory publish data gathered by others, they can be considered as free-riders. The information common will –like any other common- experience a collapse, breakdown, or tragedy, if freeriding behaviour starts dominating cooperative behaviour. Freeriding can be considered as a situation whereby one person seeks its self-interest at the expense of others by not or insufficiently contributing to a joint effort when the person will benefit from the contributions of others. Thus, and this will turn out to be important later on, it leads to a decrease of the total value of the good. The advantage of collaboration is thus at the same time also the biggest disadvantage: sharing information also means loosing control over ones work and the creation of the opportunity for many to profit from contributions of some. One can say that the collaboratories as presented here are "in between" formal and informal communication.[18] They are being formed by the informal exchange of ideas by scholars who have acknowledged eachother as peers in their field of study. By allowing another peer to the network they create their own "information common".[19]

It should be clear that the processes that happen within the collaboratory as knowledge commons are related to data in their unfinished format whereas the data they make available to the public are the finished goods, those goods that all peers consider good enough to be disseminated in wider circles. As soon as the good has become a public good, they withdraw their (copy)rights as creators, hereby however also withdrawing their responsibility in relation to anything that happens with the data thereafter.[20] In practice: if the data are used by other's (in e.g. their publications), the members of the collaboratory can not be held responsible for possible mistakes that are being made in publications using their data.

One of the important potential threats to every common –whether it consists out of natural resources or virtual information- is a population rise. An increasing number of participants to a collaboratory can have benefits: the more people contribute the more data that become available to the others. There are several downsides to a rise in participants too however. Firstly, the number of participants can rise to such a number that efficient management of the group becomes impossible or very difficult. Moreover, it becomes rather unlikely that those who become part of the collab will also benefit as much as they have contributed. As soon as the data become publicly available however, this situation changes: at that point the members of the collab have already reaped the benefits of their efforts. Afterwards, it can only be beneficial if as many people as possible use the data.

However it is mostly not the number of people that forms a problem to the functioning of the collaboratory, but their degree of participation. Being part of a group of experts is for experts of the subject always interesting. However, if a large number of people does not contribute but does manage to gain access to the results obtained by others, problems may arise. This problem was for example noticed by a collaboratory called the "Upper Atmosperic Research Collaboratory", that was initiated in 1993 to serve the needs of a distributed community of space physicists at an observatory above the Artic Circle.[21] The pool of

[18] Hess and Ostrom give a definition of "Scholarly communication": .. is the system through which research and other scholarly writings are created, evaluated for quality, disseminated to the scholarly community, and preserved for future use. The system includes both formal means of communication, such as publication in peer-reviewed journals, and informal channels, such as electronic list servers."

[19] In literature the term "knowledge common" is also used. However, considering the goods that are exchanged in a collaboratory, the term information common seems to be more appropriate.

[20] Although initatives such as the "Creative Commons" also offer licences that allow an itnermediate solution, whereby only a number of rights are reserved. see: www.creativecommons.org

[21] Finholt, 'Collaboratories'

participants in data gathering sessions was expanded but the new participants tended to be relatively passive.[22] Some participants only observed what was going on in the collaboratory, thus stressing the educational use of the collaboratory, although this might not have been an objective of the collaboratory. In 1999 this collaboratory was dissolved. It is in this context of possible problems arising due to the number of participants that Ostrom advises to limit access to the common by clearly defining boundaries.[23] When setting up a collaboratory, it has to be defined who can actually be part of the collaboratory. Again, this situation changes if the collaboratories data move into the public domain. Although the data supllied by the collab in principle do not change (they remain digital data), their status changes. Instead of being accessible by everybody, a collaboratory keeps these data temporarily as the property of those who have collected it. This stadium in the information cycle is necessary because of a very simple, but essential, reason: most scientist will not collaborate in the pre-collection nor in the post-collection phase if they can not be assured that their data will be kept from wide distribution. It all comes down to gaining trust via the principle of "do ut des" or data for data. The extra advantage of a cçollaboratory here is that this kind of exchange is not a one-to-one agreement but a group-agreement, and this is important, in the pre-data-collection phase. This last aspect is essential to make the exchange of data useful, in terms of comparability of the data (kind of data) and data format (digital format, software used, etc.).

An essential aspect of the day-to-day working of a common is the existence of mutual trust, as a form of social capital: a collaboratory –nor any other common- will not function if no trust exists between the participants. In order to achieve such trust one needs the right institutional design and the willingness of the peers to participate, thereby following certain rules. An institutional framework, consisting of the necessary do's and don'ts, can help to enhance mutual trust among the peers participating in the collaboratory. Not mere institutional frameworks –like incorporation in a larger, already well-established institution, but a specific framework that is modelled after the specific needs of the participants, including requirements to be a member of the collaboratory, the necessary do's and don'ts and the instruments to "punish" those who breach the rules. This is one of the points where an information commons like a collaboratory differs from a digital archive: a digital archive is a pool of information within an institutional framework, however those who donate the data are not involved any longer after having donated, nor will they –via the submission of their data- obtain rights on the data that others have submitted. Their participation basically stops when they have pushed either the submit- or the download-button. Although of course the submittor can set conditions for the downloading and using of the data, he or she is not necessarily involved in what happens with those data. This also means that the data archive as an institution are not necessarily modelled on the users but on the data they have donated, being an archive not the funder of researchers. The researcher is dependent on his own creativity to connect his research with what might be available on the hard disk of the data archive. Besides, those who are managing the collected data only come into the picture at the very end of the flow of information, at the point when the data already exist. This being the case, it is not surprising that there is insufficient trust from the side of the researchers to deposit the data. Insufficient here means: it is not a natural reflex yet among researchers –certainly not of the historical kind- to deposit their data (or even to consider it). Part of the reluctance for data submission might also be explained by the documentation in terms of metadata that reserachers have to supply. Here again, it is the stage of the information cycle wherein the data are submitted that is spoiling the fun. Many data archives are now advising researcher how to set-up their databases and how to document them, but again for achieving comparable datasets this is insufficient. Making comparable datasets is certainly but not only a matter of sufficently documenting what has been collected –which is largely a post-collection-procedure, it is also a matter of fine-tuning in advance what will be collected.

[22] Finholt, 'Collaboratories' refers to McDaniel, SE, Olson, GM, Olson JS (1994), Methods in search of methodology: Combining HCI and object orientation, in Proceedings of CHI '94 (pp. 145-151), New York, ACM press.
[23] Ostrom, *Governing the Commons: the Evolution of Institutions for Collective Action*: 90, design principle 1

The management of a common requires a set of rules in order to organise the interaction between the created goods and the users of those goods. In a collaboratory the rules concern several aspects: the way the goods should be created (in practice: which are the standards that should be followed), the way they should be deposited (agreements on standards), the way they can be used (use rules; e.g. not for commercial use) and by whom they can be used (access rules; e.g. only by recognised 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 is no sanctioning if rules are broken.[24] One of the possible sanctions could be exclusion from further use of the collaboratories' work and data. Entering the wrong data: if people are unsufficiently commited to strive to an optimal outcome, the quality of the data entered or delivered may not be high enough. This can be prevented by assuring that their benefits will depend upon their 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 wrong data) are "punished". This can be achieved by random checks of the data and by more general social control among peers. Adding a "reputation-component" (or peer review of the data) to this might be helpful to sustain a high quality collaboratory.

As soon as the collaboratory claims its property rights in the form of a publication, and has thus received the academic credits for them, the collaboratory can choose to turn that information into private property of the association of collaborators and sell the dtaa to third parties. Or they can decide to make the data public property by denouncing their property rights. At that point freeriding is no longer possible since they do not contribute in the way the members of the collaboratory did, namely as peers. As the peers do not know who is using their data, they cannot acknowledge the third parties. However, not recognizing those users as peers also frees them of the need to recognize the data as correct. Basically, the collaborators can also decide not to change the property situation, they can continue as a common just like before. They can decide to include new participants on the basis of their access rules. However, in doing so they keep the information to themselves.

Finholt refers to several collaboratories whereby agreements were signed to protect each others rights. In a community of brain researchers a formal covenant was set up, that was signed by scientists as a condition of use of the collaboratory, and specified how community data were to be used, thereby 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.[25] There are quite a few other examples where good regulation has led to the establishment of long-term intensive and highly productive cooperation among scientists.

### *5. Collaboratories and the information cycle*

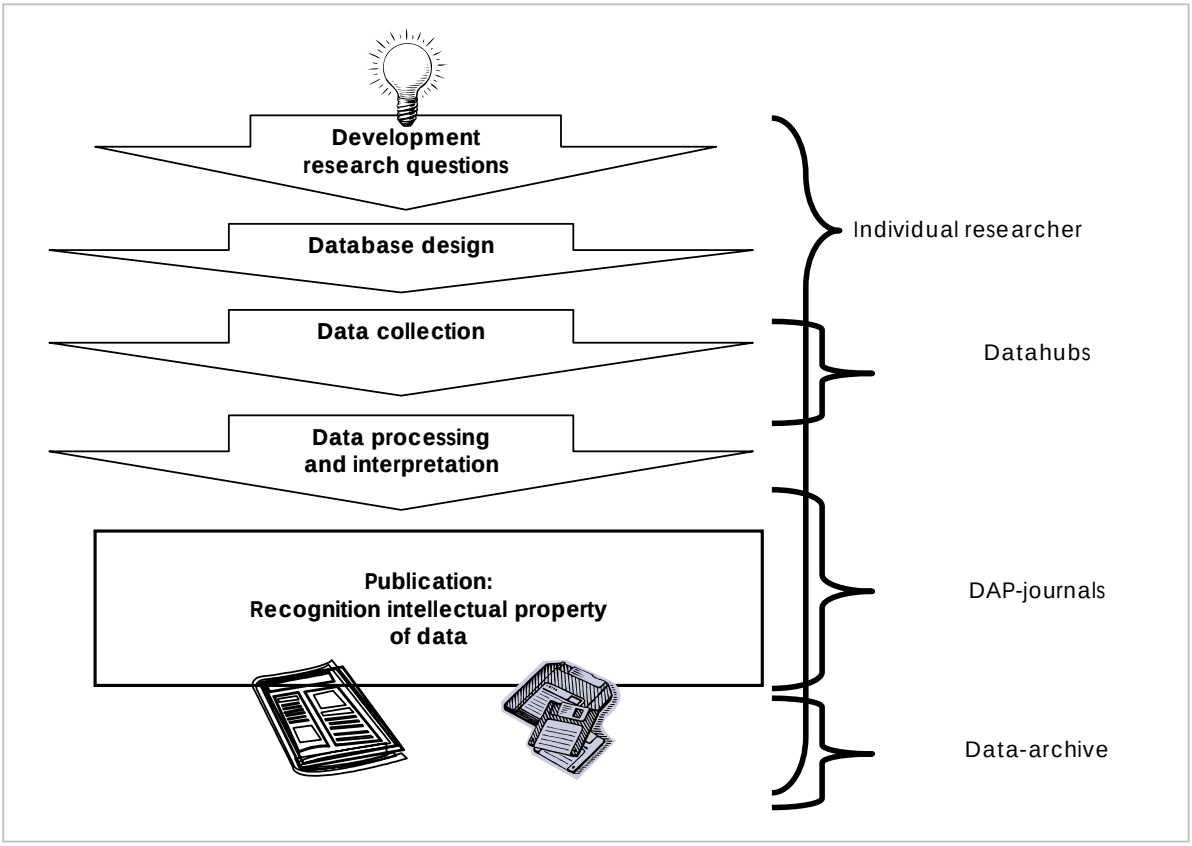
It has been mentioned several times so far that the advantages collaboratories have over data-archives, and also over DAP-journals though this has not been explained at full yet, is due to their position within the information cycle: right at the beginning of that cycle, after peers have exchanged ideas about a potential data collection project, but before the actual collection. Of course, this is what happens in theory. In practice the researchers might have already started to collected data. The fact however that assuming their exchange is based on the intention to compare the data sets, assumes also that they will fine-tune their further data collection in order to make them comparable.

A comparison of the information cycles in a normal situation whereby researchers do not collaborate and the collaboratory-information-cycle may clarfiy what the differences actually are. In the first case (see Figure 1), the researcher transforms an idea into research questions and designs a database that can contain the data necessarily to answer those questions. Increasingly, data archives are offering researchers assistance in this stage but

[24] See among others Ostrom, *Governing the Commons: the Evolution of Institutions for Collective Action*: 90
[25] Finholt, 'Collaboratories' He refers to the UARC and SPARC collaboratories . See SPARC: http://www.si.umich.edu/sparc

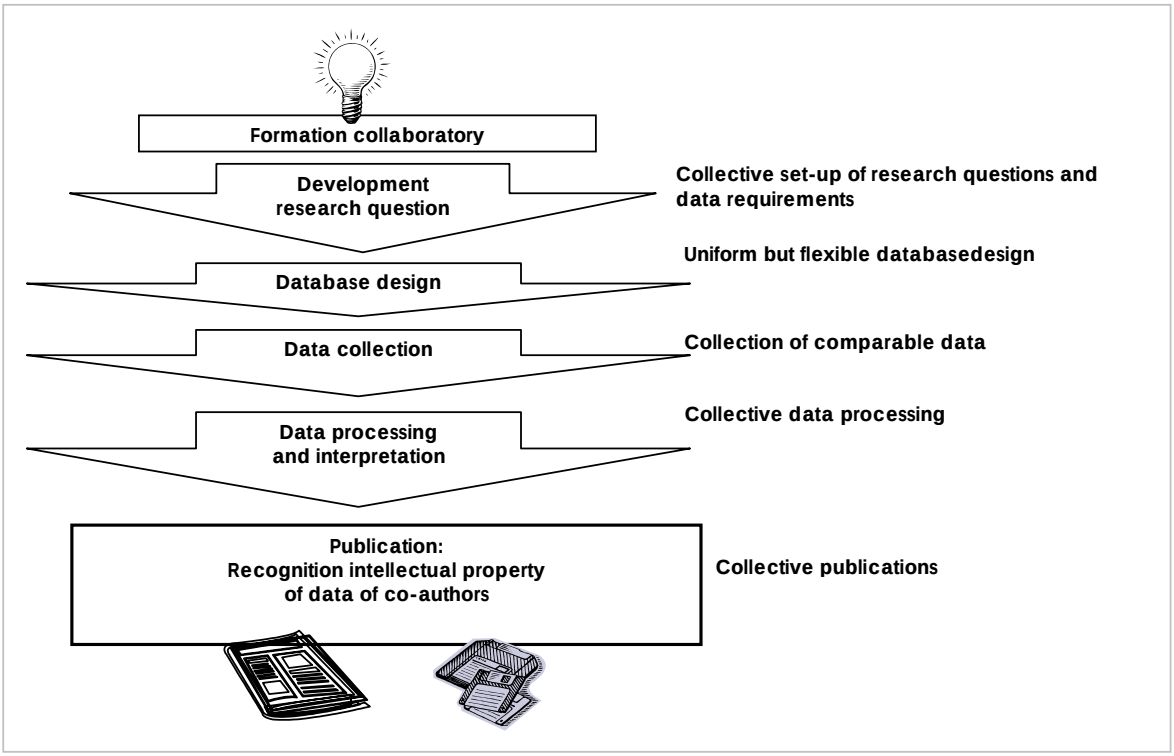
overall, the researchers are mostly doing this on their own. In subsequent phases the researcher collects the data and processes these, after which the interpretation of the results hopefully end up published. On the right hand side of the figure it is indicated that researchers sometimes, during the data colelction make use of datahubs for retrieving data. This is for example done by the historians using the prices and wages hub of the IISH. When the results are published in a DAP-journal, the researchers are also demanded to deposit their data. So far, all journal that have such a policy, do not use recognised data archiving methods but simply put the data on an accessible website. In an increasing number of cases, the individual researcher also find his way to the data archive. At the end of the information cycle the researcher sometimes submits the whole data collection to the data archives, thus allowing third parties to become inspired for new research, which may lead to a new start of an information cycle.

**Figure 1: Normal (individual) information cycle from idea to publication and dataset**




In the –it has to be admitted, ideal- collaboratory information cycle, researchers fine-tune their ideas about research questions from the start. The exchange their ideas about the necessary data and data format, and set-up a data base format that can contain data in different but comparable data formats. Thereafter the data collection can start. The data processing is also a collective procedure, since this may raise new questions about eachothers data and usually also is the basis of collective publications, whereby the members figure as co-authors.

**Figure 2: Collaboratory information cycle**




### *6. The complementarity of collaboratories, data archives and DAP-journals*

In this article we have frequently referred to the potential the collaboratory-formula has in solving some of the difficulties data archives are experiencing. It has been pointed out that this has to do with the different spaces within the information cycle that collaboratories, data archives and also DAP-journal are situated in. There is no point in trying to merge all functions of these three initiatives. On the contrary: although there is always a certain overlap, the researchers (in collaboratories), archivalists (in data archives) and journal editors (of DAP-journals) are different stakeholders within academia, and it would not be wise to mingle these stakeholders different positions. On the path towards an optimal and durable exchange of data and ideas, data archives, DAP-journals and collaboratories can unite forces by reciprocal offering of the services they are best in.

Complementarity should be the key-word on the path towards optimal exchange of data and idea. Each of these initiatives face three problems: first of all convincing the authors to make available the data, secondly making sure that the data are of a high quality and thirdly the problem of data storage. Each of the three forms of data exchange score differently when it comes to solving these problems. Collaboratories can solve these problems to a certain degree, but in many ways two other initiatives in data collection and scholarly communication can hereby play a complementary and essential role.

**Figure 3: Overview of problems of data archives, DAP-journals and collabs**

Overview of problems collaboratories, data archives and DAP-journals face on the path towards exchange of data, with the three initiatives in order of successfulness 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 well-defined circles)
3. Data archives (rely upon external institutes to convince researchers to deposit data)

**PROBLEM 2. MAKING SURE IT IS HIGH QUALITY DATA: 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 somebody finds it worthwhile to check

**PROBLEM 3. KEEPING THE DATA ACCESSIBLE, USEABLE AND USEFUL: storage of data for further use**

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

Note: 1 = best solution, 3 = weakest solution

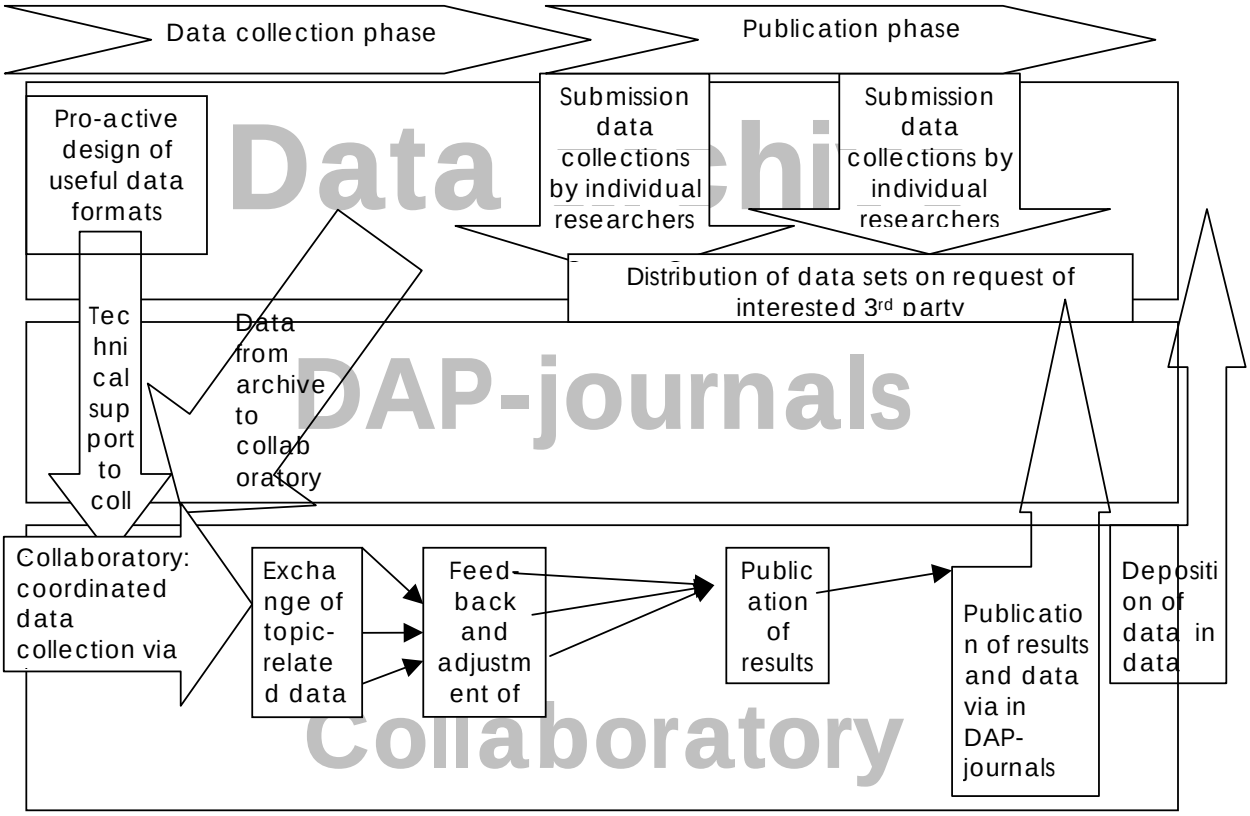
Each of the initiatives has its own points of strength, and in that sense they can each contribute to the composition of useful global datasets by offering their solutions to those three problems. In solving the problem of enhancing the availability of data for secondary users (exchange of data), we can assume that DAP-journals offer the most powerful solution: if data availability is set as a requirement for publication, it is very likely that authors are willing to provide their data considering the importance of publications for once's curriculum. As the data are made available to the wider public only after publication, the authors can still benefit at full from their data collection work. Collaboratories have the advantage that by participating one can also obtain –prior to publication- other researchers' data. "Give and though shall be given" is still a good incentive for cooperation among researchers. But apart from the promise that one will be able to us another researchers data prior to publication, there are not as many stimuli to make ones data avilable to others, as in the case of DAP-journals. Submitting your data to a data archive is however still not a natural reflex of researchers, even under the pressure of funding bodies that sometimes demand this after the termination of a research project. The degree to which researchers are willing to do this, remains thus largely dependent on their good will.

The degree to which the three initiatives can solve the second problem depends on the stage and intensity where others –peers- can check the data. The sooner in the information cycle and the more intense this happens, the better for the quality of the research results. In the case of the DAP-journals, a select group of peers can look at the data right before publication, when an author offers an article for publication. And soon after publication the data become available to whoever 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 prior to publication, and if the author decides so, this could be years after the collecting of the data has actually been finished. Both data archives and DAP-journals thus offer an insufficient solution to the peer-reviewing of the data prior to publication. Collaboratories however start right at the beginning of the research process to exchange ideas about the data collection method. The building of data hubs that are virtually accessible is a collective effort of peers that are linked to a collaboratory, who can correct eachother during the whole research process, until publication, and if necessary even after that.

The third problem of long-term storage is one wherein data archives score the best. As this is a novel process, it is still unclear how DAP-journals will deal with data archiving. They put data on their websites but for how long and in what sort of data system, nobody –including

usually the editors of the DAP-journals…- knows. If these journals intend to make the data available for a long time, it is unlikely that they have the expertise to keep them actually available in the long run. In this sense it is in the long-run no solution to effective data storage. The same goes for data collaboratories : these are concentrating on creating good data sets that can answer particular research questions, hopefully also resulting in top publications, but what happens after that is still unclear. In this sense DAP-journals and collaboratories are probably both equally insufficient in solving the third problem. The real expertise for data storage is available within the data archive institutions. It would therefore be optimal if collaboratories can at the end of the information cycle offer the information gathered within the hub for storage. The same goes for DAP-journals.

**Figure 4: Possible forms of cooperation among data archives, DAP-journals and collaboratories**




The optimal solution would however be a close cooperation between data archives and the other forms of scholarly communication right from the start of the research projects. Optimal efficiency of (human) resources could be achieved when collaboratory-participants apply the expertise of data archivalists when setting up data formats for their data hubs. This early cooperation between data archives and collaboratories could result in a win –win-situation for both parties: collaboratories can benefit from the expertise of data archives and data archives can expand their data collection with large amounts of high quality data and can anticipate on the needs of historians via close collaboration with the collaboratories, without hampering the scientific process itself. In the same sense, data archives could set-up intense collaboration with DAP-journals, by offering support for the data storage of the submitted article.

It is clear that every form of scholarly communication and exchange has its advantages and disadvantages. Linking and to a certain degree merging the activities of data archives, collaboratories and DAP-journals can lead to an optimal use of the available

financial resources, in terms of infrastructure and labour. Close interaction between the three initiatives could ease the way towards more accessible data and more quality. Via the collaboratories, top researchers can convince journals in their fields of research to implement a data availability clause in their policy. This will increase open access to scientific data in general, and enhance the awareness of the need for peer review of data –and not only results- among researchers.

## *7. Conclusion*

In this article, we have pleaded for the set up of new collaboratories in the social sciences/humanities, and in particular in history. This alternative concept of collaboration and exchange between scientists is still largely unknown to this scientific community although it can prove to be a rewarding method of overcoming the data collection problem that is necessary for the "big history" of future historiography. The increasing importance of cross-boundary global work, like in economic history, makes it impossible for one single researcher 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 can – for several reasons- not solve this problem adequately.

In terms of benefits, the main difference between collaboratories and data archives is that contrary to a data archive a collaboratory enhances –at least if it turns out to be successful- communication between the participants. Because interaction is possible, comparison of the data and if necessary adjustment/correction is possible before publication, hereby making the data collection more efficient for the collectors and the second hand users.

In terms of functioning, the main difference is the degree of participation between all the parties involved. Participation enables an intensive process of peer-interaction and –review that would be impossible within the context of a data archive. But the risks that such degree of data sharing and communication entail need to be taken care of via specific institutional frameworks. Data archives avoid this risks of exchange for a great deal since the exchange only starts after the author of the data is convinced that he has had all possible benefits for himself. This condition however prolongs the moment of making the data public and it makes pre-publication peer-review impossible. Both these outcomes are not advantageous for the advancement of science in general. A combination and cooperation between historical data archives, different collaboratories and other forms of interactive peer review like DAP-journals could however be extremely beneficial for the academic research in general.

REFERENCES

Agarwal, D. A., Sachs, S. R., and Johnston, W. E. The reality of collaboratories. 110 (1-3), 134-141. 1998.

Bender, E., 'Rules of the collaboratory game', *Technologyreview.com*, November 22 (2004),

Bromley, D. W., Feeny, D., *Making the commons work: Theory, practice and policy*, 1992).

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

Finholt, A., 'Collaboratories', *Annual Review of Information Science and Technology*, 36 (2002), pp. 465-91.

Kouzes, R. T., J. D. Myers, W. A. Wulf, ' Collaboratories: Doing science on the internet', *IEEE Computer*, 29 (1996), pp. 40-6.

Lunsford, K. J. and Bruce, B. C. Collaboratories: Working together on the Web. 45 (1), 52, 7p. 2001.

Ostrom, E., *Governing the commons: the evolution of institutions for collective action*, 1990).

Weinberg, A., 'Impact of large-scale science on the United States', *Science*, 134 (1961), pp. 161-4.

Wulf, WM. A. The collaboratory opportunity. Science 261 (5123), 854, 2p. 1993.