Scientific Journal Publications: On the Role of Electronic Preprint Exchange in the Distribution of Scientific Literature

Article *in* Social Studies of Science · August 2005

DOI: 10.1177/0306312705052358

Kristrún Thorlacius Gunnarsdóttir
City of Reykjavik

Scientific Journal Publications:

On the Role of Electronic Preprint Exchange in the Distribution of Scientific Literature[1]

Kristrún Gunnarsdóttir

k.gunnarsdottir@surrey.ac.uk

Gunnarsdóttir, K (2005). On the Role of Electronic Preprint Exchange in the Distribution of Scientific Literature. *Social Studies of Science* 35(4):549-579.

ABSTRACT: The scientific community has begun using new information and communication technologies to increase the efficiency with which publications are disseminated. The trend is most marked in some areas of physics, where research papers are first circulated in the form of electronic unrefereed preprints through a service known as arXiv. In the first half of this paper, I explain how arXiv works, and describe the conceptual backstage and its growing influence. I will look at the motives behind the developing technologies and focus on the views of promoters and makers of the system. In the second half of the paper, I look at the eventual fate of papers initially circulated with arXiv. While it is argued that preprints are sufficient for the everyday scientific practice, nearly every paper in some specialities finds its way into formally peer-reviewed journals and proceedings. I argue that the continuation of traditional publication practices, in spite of their costs and inefficiencies when compared with arXiv, suggests that formally certified publication still has important roles. Certified publication verifies the relevance of scientific work and establishes professional credentials in the outer rings of the community, whose members are not sufficiently embedded in esoteric networks to make appropriate judgements on the basis of reading papers in isolation, or even through consultation.

Keywords academic reward, digital library, electronic publishing, experts, information technology, journal publication, peer review, subject classification

New communication and information technologies are changing the academic workplace, yet the impact is not well known. Networked computer systems for instant electronic exchange were introduced to the community in 1969, with the first email application of the Arpanet.[2] Nowadays, communication and database management technologies offer academic researchers instant global access to both refereed and unrefereed repositories of research results, some free of charge, with extended citation services, data harvesting and search functions. Nevertheless, older communication channels continue to have currency. These include: (1) the telephone, which now comes with call waiting, conference call options and call transfers; (2) conferences, workshops and seminars, which sometimes offer exotic foods, wines and opportunities to chat with colleagues; (3) book publications, often intended for a wider audience; and (4) conventional subscription journals, with their hierarchy and complex systems of refereeing.

Information technologies are widely considered to offer much faster, cheaper and more efficient solutions for publication and distribution. Why then, do all the old, slow and expensive channels continue to exist alongside the fast, efficient and cheap channels? To give a partial answer to this question, I will look at a case study of arXiv, an unrefereed preprint dissemination system originating in the high-energy physics (HEP) community.[3] Preprint distribution among physicists has been going on for the past 40 years, long before the advent of networked computing, but more recently the advancing technologies used to communicate research results have brought to the fore some puzzling aspects of how publication is regulated.

The present study examines the arXiv system, focusing on how that technological system serves the communicative needs of a mainstream science. This paper covers the history of the system and examines arguments, provided by its advocates, that explain the nature of its success, but also challenges the most puzzling aspects of its role in scientific publishing. My treatment of this system is influenced by several lines of work in science and technology studies (STS). Ethnographic studies of HEP offer insight on cultural and communicative aspects of the field

(Traweek, 1988; Knorr-Cetina, 1999). Studies of the history of the scientific paper also provide helpful information on how novel modes of publication influence scientific communication and community. In addition, studies of technological systems can shed light on the relationships between technological designs and the activities and identities of users.[4] Finally, borrowing from Collins' (1985; 1999) studies of scientific controversy, I will apply the notion of a 'core-group' to users of the arXiv system, focusing on its outer social rings, and I will ask what this context signifies with respect to publication trends, media formats, and policing, when successful translation and mediation of relevant scientific contributions is at stake.[5]

The Problem of Epistemological and Logistical Controls

One of the supposed disadvantages of an open online preprint exchange of academic publications, as compared with formally refereed journals, is that any person with institutional affiliations can upload their work into the system. The risk is that the repository will be flooded with erroneous results or irrelevant content. There are two parts to this objection. The epistemological objection is that the archives will become hosts for poor science and readers will be scientifically misinformed. The logistical objection is that a flood of irrelevant papers will hinder scientists in their search for information. I will argue that in the case of esoteric scientific practice, the epistemological objection is of no importance, because core-group readers are self-policing with regard to the quality of scientific claims. And, when research topics cross the boundaries of core-group practice, core-group members have ready access to other communication channels for consulting with colleagues in other specialities. I will argue that the logistical problem is essentially a problem of membership definition, of submitter authentication and subject classification, rather than of detailed refereeing, and this requires entirely different kinds of solutions. In the case of arXiv, authentication is automated, but human input currently solves the problem of inspection and subject classification. This is to say that authors classify their own papers as part of the uploading process, and system administrators and moderators correct the classification when necessary, before opening them for online dissemination.

The fact that an archive of unrefereed preprints can operate without presenting an epistemological problem only sharpens the question of why scientists still publish nearly all of their research through the refereed channels. Advocates of arXiv are proud to claim that most of the papers disseminated through the system are eventually published in conventional scientific journals and reputable conference proceedings. They cite this fact as an indication of the quality of the preprint service. Formal publication is still a necessary gatekeeper to assure outsiders that a piece of work has significance when they need to make decisions about appointments, promotions and funding (Bohlin, 2004). Conventional journal publications have a symbolic role for these outsiders, whereas the preprint dissemination bears the burden of information exchange in the scientific workplace.

Achieving Automation: Aiming for Autonomy

ArXiv began operation at Los Alamos National Laboratories (LANL) in 1991. Initially, it was a simple email reflector for exchanging papers among theorists in HEP. It grew quickly to implement the File Transfer Protocol (FTP) and, in 1993, the HyperText Transfer Protocol (HTTP) to enable the migration of the rapidly growing system to the newly invented World Wide Web. Its use spread in the larger HEP research community, and other core-groups who were comfortable with the infrastructure that the system provides, gradually became involved. Papers posted on arXiv have always been accessible for free to all people with Internet access, and no charge is required from contributing scientists who take responsibility for uploading their papers, using the automated author self-archival module.[6] The founder of arXiv explains the rapid acceptance of this model in his research community by saying that it was `facilitated by a pre-existing "preprint culture", in which the irrelevance of refereed journals to ongoing research has long been recognised' (Ginsparg, 1994:1)[7]

The library of the Stanford Linear Accelerator Center (SLAC) began actively collecting preprints of physics papers and maintaining bibliographic records as early as 1962.[8] In 1968 the bibliographic card catalogue of these preprints was used as test subject for the computer database system at Stanford University, and the development of the Stanford Physics (later Public) Information REtrieval System (SPIRES) began. In 1969 SPIRES, in collaboration with the American Physical Society, started a weekly publication of preprint titles (SPIRES-HEP), known as Preprints in Particles and Fields (PPF) and its sister ‘Anti-Preprints’ recorded the preprint titles that had been formally published (PPA).[9] It was also in 1962 that the Deutsches Elektronen-SYnchrotron (DESY) in Hamburg began publishing a list of all the published an unpublished HEP research literature they received, and they hired physicists to assign long lists of keywords. These two institutes began collaborating, and by 1974 high energy physicists and their collaborators on both sides of the Atlantic were subscribing to regular listings of new preprints in the HEP community, generated from a single database with keywords and citation overviews. They circulated copies of unrefereed research papers using ordinary mail when responding to requests from other scientists. These same papers would typically take anywhere from 6 months to more than 1 year to be published in the relevant scientific journal, and scientists preferred not to wait so long. Apparently, scientific practice required much faster exchange than conventional publications could offer, and the demand for preprints was well recognised in the community. In 1985 online access to the bibliographic database was made possible through email and other bit-net protocols, and its use spread quickly to more than forty countries. Subscriptions became electronic, but it still was impossible to get electronic access to the actual preprints or published papers. They were still sent by ordinary mail service upon request from subscriber.

At the end of the 1980s,TeX, a scientific typesetting software program, had become available to members of the community free of charge.[10] Suddenly it was straightforward to produce a preprint of equal presentational quality to a published paper. When arXiv was founded in 1991, it introduced a facility for authors to upload the TeX source of their papers and to place them on the network, thus making the files instantly available for downloading. However, users of the system still had to compile the papers and apply postscript interpreters on their local machines in order to read them. The construction of this new facility was accomplished by implementing a simple program design around an already existing bit-net technology. Uploaded paper source packages automatically went into a single repository and instant access was available by using basic terminal commands. The design of this system included an implicit assumption about would-be patrons: that they were trained in computer programming and system/software configurations, and thus would be able to understand the mechanisms behind the system and the software they were using. [11]

ArXiv as *a pure dissemination system* is one key concept introduced by its advocates (Ginsparg, 2003). The anticipated outcome was that control over the dissemination of the literature would be placed solely in the hands of the practising scientists. Technical developments have helped to give the impression that such an autonomy is possible. Current features include: (1) the TeX auto-compiler that compiles the TeX paper source on upload and tests the source code for integrity. Consequently, postscript documents are made available 'on the fly', and the original TeX source also is available for users who wish to do their own compiling. The auto-compiler project has been going on since 1995, and its development been for the most part in the hands of physicists and arXiv programmers at Los Alamos; (2) dynamic conversion of postscript outputs to a Portable Document Format (PDF) is available on demand; (3) automated citation analysis and automated uploads of journal references are carried out in collaboration with the library at SLAC (SPIRES-HEP) and with CiteBase, which provides Open Archives services of autonomous citation navigation and analysis;[12] (4) compliance to the Open Archive Initiative (OAI) protocol for metadata harvesting has been implemented since 2001. This protocol provides an application-independent interoperability framework, which is to say that independent web-based repositories scattered all over in different formats only need to make sure that their software and metadata is compliant with the protocol; robot harvesters on the Web, ploughing for data sources using OAI, will extract the data and index it for search engines.[13]

ArXiv has become a large technological system that has strong associations with basic research on networked computing and with broader developments in practices of document rendering and format conversion - all of which support its growth and add to its strength.[14] Physicists trained in HEP and related areas, and mathematicians, are mainly responsible for the design and development of the arXiv system. Those specialities provide much of the technological frame - the whole range of practices, ideas, and values that evolve around the technology.[15]

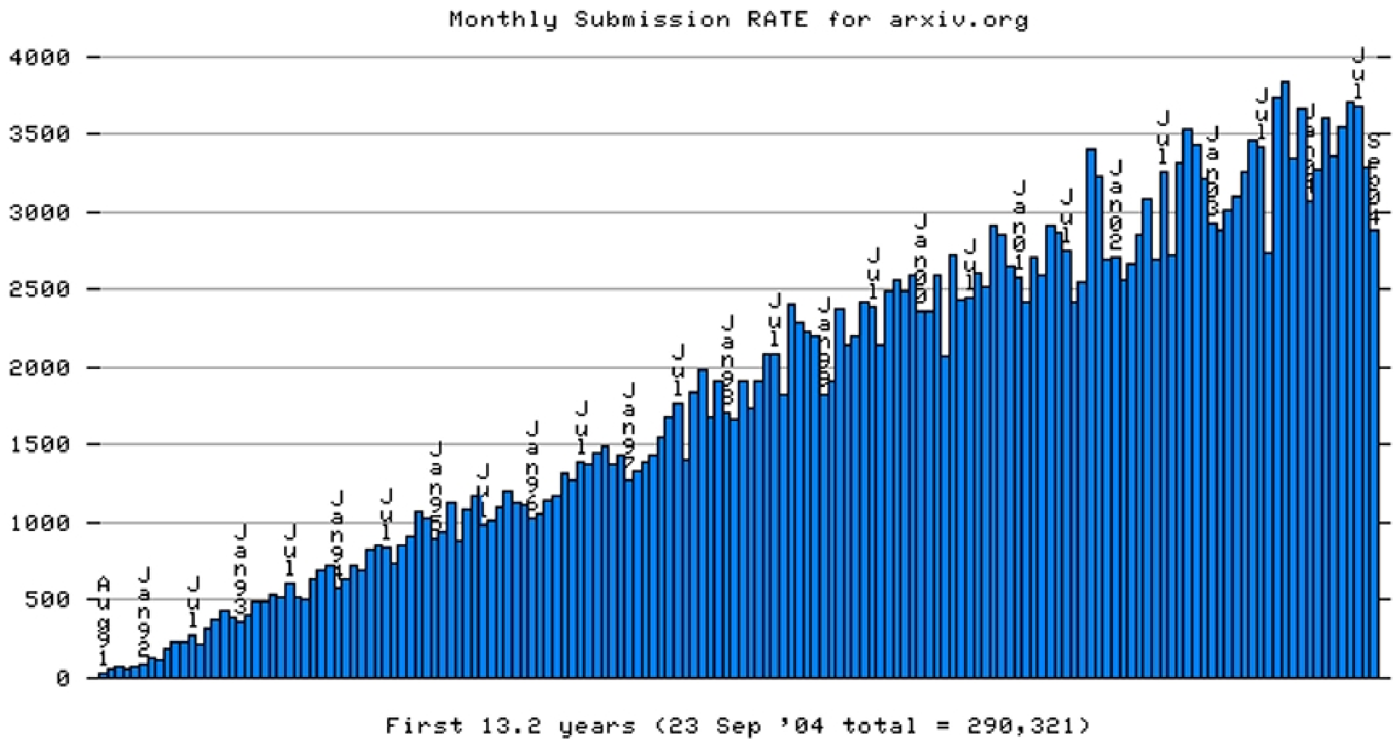

FIGURE 1
Monthly submission rates for arXiv in the first 13.2 years (total *N* at 23 September 2004: 290,321). From arXiv online submission statistics at <http://arxiv.org/show_monthly_submissions>. Courtesy of Cornell University Library.

Participation has grown steadily over time in specialities that find the arXiv infrastructure adequate for their purposes, despite the lack of conventional peer review. Recently, scientists in quantitative biology adopted the system.[16] As can be seen in Figure 1, arXiv has been successful as measured by a growing submission rate**.**

Advocates of the system claim that the emerging new technologies have increased the efficiency in their everyday practice and they have argued strongly that adoption of the system has led to economic gains. They point to figures that compare the costs and revenue associated with different publishing systems. They argue that peer review systems, although the formal reviewing of papers is done with voluntary work, entails unavoidable costs covering editorial expenses, secretarial work, administration, overhead, and so forth.[17] They contrast the enormous revenues generated by trade publishers and the costs associated with the publishing activities of science societies and university presses, with the cost of arXiv operations of only a few dollars per paper.

ArXiv is promoted as a feasible alternative to conventional practices for other reasons as well. Advocates of the electronic preprint exchange have repeatedly argued that the conventional peer review system does not verify research results. They claim, for example, that referees in the anonymous refereeing model have been known to accept papers based on their recognition of authors rather than their evaluation of the content of the papers.[18] They point out that the legitimacy of a paper's topic is the most important condition for it to be accepted for publication.

> Outsiders to the system are sometimes surprised to learn that peer-reviewed journals do not certify correctness of research results. Their somewhat weaker evaluation is that an paper is a) not obviously wrong or incomplete, and b) is potentially of interest to readers in the field. The peer review process is also not designed to detect fraud, or plagiarism, nor a number of associated problems -those are all left to posterity to correct. (Ginsparg, 2003)[19]

Ginsparg and other advocates of arXiv argue further that conventional peer review processes do little more than certifying publications by a scientific elite, and add little value to the routine validation of research results that already takes place in the course of the daily practice. They add that peer review serves the sole purpose of job and grant allocation, and that such certification through peer review can be greatly simplified while other processes are automated.[20] They see the problem as technical - a matter of economy and efficiency, rather than an epistemological problem in scientific practice.

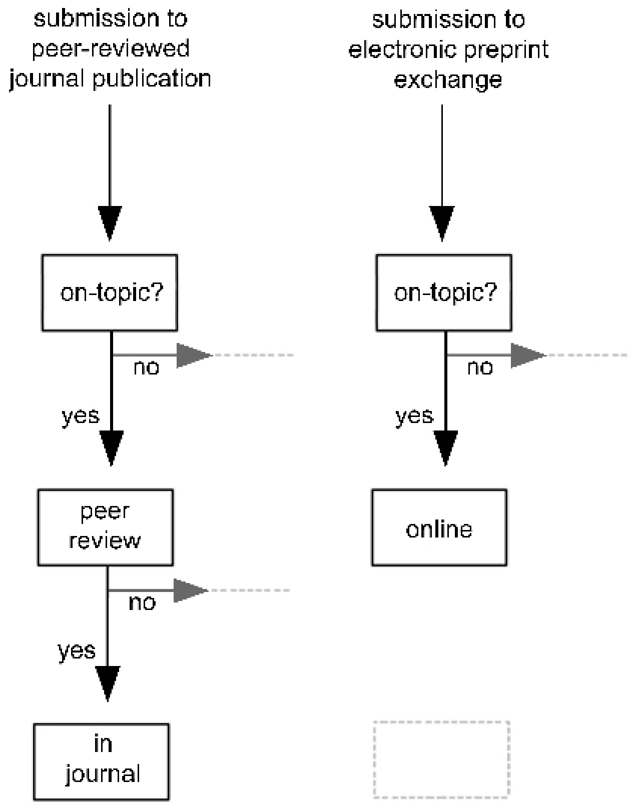

FIGURE 2
Comparison of conventional publishing procedures with the preprint exchange. Advocates of the preprint exchange argue that little value is added with the formal peer review process. (This figure is originally drafted collectively at a research group meeting at Cardiff University)

The ambition embodied in the arXiv system is to automate the dissemination process in a way that allows members of the core-groups to have little fear that 'non-certified' intruders will get into the system. The general rule is that legitimate members speak the same language and write 'on-topic' papers. Everyone else is excluded.[21] It has also been argued that legitimate researchers are published in the elite journals as a matter of course, and that journals also include research reports that are later refuted. The sorting mechanisms are already in place in post-graduate and post-doctoral training. Overall, scientists either become members of the core groups or they do not, and if they do get into such groups they produce mainstream science.[22] This means that the relevant threshold for entry can be implemented by a preprint server that does no more than make an 'on-topic' or 'off-topic' decision about paper content; in effect, arXiv's proponents argue, the journals operate according to the same basic standard.

History Seen in a Broader Perspective: ArXiv, the Unique Case

Since the early 1990s there have been larger and deeper crises in all areas of scholarly research about how to disseminate the conventional academic journal literature. ArXiv is seen as a pioneering response to this crisis. It is the first concrete example of scientists taking the dissemination process and the ownership of papers into their own hands, and scholars in other disciplines have wondered how they also can take advantage of such a system. I argue that it is important to recognize that arXiv is a unique case that does not easily compare with other initiatives. To clarify this point I will look briefly at two key arguments in the broader debate.

The first regards subscription charges, which have been a source of increasing concern for academic libraries and individual researchers. The question is: 'Who owns the scientific journal literature?'[23] Scientists and scholars usually receive no direct compensation for publishing journal papers or for reviewing and evaluating submissions. However, they and their institutions are welcome to pay heavy tolls for access to these same publications. This criticism expresses an increased resistance to profit-making trade publishers of scholarly literature, who operate in a particularly non-elastic market.[24]

Criticism is also aimed at not-for-profit societies and university presses that have high production costs. This criticism is largely based on the assumption that the digital and networked age should have made journal publication much cheaper, and printing them even entirely unnecessary. By the mid-1990s it had become impressively cheap for any individual to typeset a document with a simple hypertext mark-up language (html) editor, which soon became an export format option embedded in some word processors. It also had become possible for almost anyone to post documents online, from where they could be viewed and downloaded by anyone else with access to the relevant online network. It follows that digital processing and digital publishing should make formal and certified publication much less expensive. Demonstration of the arXiv model's very low production costs clearly supports this arguments, but predictions in the early 1990s that all academic journal publications would become very cheap, or even free of cost, and that journals would go immediately online as soon as journal administrators learned to implement and manage new technologies, did not come to pass. Conventional journal publications have gradually come online, but most of them have access constraints and the price paid for access is often high.[25] Nevertheless, pressures increase for publishers to allow free online access:

> Just as there is no longer any need for research or researchers to be constrained by the access-blocking restrictions of paper distribution, there is no longer any need to be constrained by the impact-blocking financial fire-walls of Subscription/Site-License/Pay-Per-View (S/L/P) tolls for this give-away literature. (Harnad, 1998, 2001)

It has been a source of pride for the arXiv operation that access to papers is free of charge and globally available. Even the final versions of papers that are formally published in journals are often posted with arXiv under a 'free-for-fair-use-by-all' principle. In this respect, the arXiv operation set an example that supported a much broader movement. It remains in firm opposition to profit-making trade publishing, and the new trend of posting for free has also put pressure on scientific societies and university presses, motivating them rethink their publishing practices.

The second argument centres on the role of peer review to uphold what is traditionally seen as the necessary certification and quality control. Apparently, no alternative publishing scheme of academic literature has renounced altogether the role of publication with elite journals for the formal certification process. However, arXiv differs sharply from other publication systems by maintaining an unrefereed preprints service, a large-scale self-moderating and self-policing scheme that openly disapproves of the certification process as inefficient and costly.[26]

Puzzling Aspects of Control, Configuration and Influence: ArXiv's Success

Publication is a significant product and an integral component of scientific research, whether distributed in printed or electronic form, refereed or unrefereed. Scientific communities use distinctive forms of internal communication, but little is known about how the patterns of communication differ from one field to another or how they are embedded in larger mechanisms of knowledge production. STS studies have drawn attention to cultural issues such as solidarity and discord, and to controversial situations such as priority disputes, uncertainties about replication, intellectual property disputes, and managing plagiarism, fraud and hoaxes.[27]

Studies of the culture of HEP shed light on the relevance of arXiv operations for that field. They depict a rather homogeneous research community, although consisting of many research units, which is an example of a 'communitarian' culture with 'object-centred' research and experimentation (Knorr-Cetina, 1999). There are only a few HEP laboratories worldwide, the experiments are enormous in size, and many published papers already are reviewed by large numbers of peers even before becoming preprints in arXiv. These papers can have hundreds of researchers listed as contributors.[28] Because authors are listed in alphabetical order they do not overtly indicate who is a leader, who is an inspiration, and so on.[29] Even when high-energy physicists work in very small groups, there appears to be little stress over priority and intellectual property. Sharing research results is a common practice, and the ground rules for data use are pretty clear. Extensive collaboration is vital and the community cannot help but depend on moral integrity and solidarity among its members.

This depiction of HEP is not that of a scientific field faced with controversy or a priority race. It is a description of a normal science that includes all of its members in large-scale collaborations that can include many different core-groups. This does not mean that there is no competition and no struggle required for gaining entry into the field and establishing a reputation. It only supports the view that, ”[a] core-group is much more likely to be solidaristic than a core-set” (Collins, 1999: 164).

ArXiv assumes no direct role in formal refereeing, but it stores papers that clearly undergo peer evaluation, whether or not they are later certified for formal publication. The operation depends on an overall solidarity among practising physicists and mathematicians. It offers no formal quality control mechanism to sanction plagiarism, fraud or marginal content. The most that arXiv enables is to revoke submitter status from ‘norm-breaking’ individuals and to establish priority by automatically stamping date and time on submissions when they are uploaded. Occasionally, ‘off-topic’ papers slip through, but this also happens with conventional publishing. Members of the participating groups are nevertheless confident that they can evaluate the posted papers. If they are uncertain about the validity of particular research findings, they can consult with other colleagues.[30] In other words, arXiv's very existence seems to demonstrate the irrelevance for practising physicists of the old slow process of conventionally certified publication.

But there is more to consider. One unique aspect of technological inventions is how they ignore disciplinary boundaries. Physicists become computer system engineers and are confident that their new ideas and choices are on the right track. Their tasks are problem oriented rather than discipline oriented, and may also involve budgetary and managerial considerations, as well as considerations of designing and developing the invention. In a historical account of invention, the connections should be followed wherever they lead.[31] Science and technology are both said to be socially constructed, but the boundaries between them are unclear, as is their relationship with economics. Technology is needed to do science and scientists produce new technologies.[32] Older technology also shapes new technology in the sense that already existing solutions serve as resources for new innovations and could actually be said to pre-configure some aspects of them.[33] This is particularly true of innovative uses of the Internet infrastructure, where an already existing grid facilitates linkages between an enormous number of new host types. ArXiv is built on a set of social and technological settlements, designed over decades by high energy physicists and their resource service units. They are the relevant social group included in the many design phases, but at the same time the boundaries of this group have become unclear.

Although with hindsight it is easier to see why some solutions succeed and become popular while others do not, none of the players in the current picture know the future. There is plenty of dust in the air that has not yet settled. Competing standards for hardware configurations, for document types and document handling, for security and access restrictions, and so forth, have come and gone, although some continue to prevail.[34] The arXiv system is also a service to scientists to *self-archive*, and when examined as a meaningful artefact designed with this particular usability in mind, it shows how that arrangement competes with other forms (artefacts) of online publications (Bohlin 2004). These are scholarly communication regimes whose symbolic social and cultural interpretation and meaning have not yet been stabilised. The cooperation between users and the system also calls for a 'configuration' or 'scripting' of both.[35] Technical proficiency is one important aspect of the history of arXiv's design and use. The technological frame incorporates basic research into information systems engineering, design, and management. Members of the user groups are expected to have an understanding of the underlying mechanisms, but not all potential users can be expected to put in the time and the effort to understand the underlying mechanisms, and so 'user friendly' simplifications are regularly suggested.[36] On the one hand, arXiv appears to be successful with the particular expert groups who directly apply its technological frame. Representatives of these academic disciplines contribute to the technical development of the system, because they can comfortably cross boundaries between their academic disciplines and the system's engineering.[37] On the other hand, the system's success thus far also indicates that behind the design is a relatively independent research group with a narrow disciplinary focus and control over enough technological and social resources to build a highly significant storehouse of scientific works. This suggests that a somewhat isolated group of designers and managers with vested interests has emerged.[38]

While the dust continues to settle, I will argue that the most significant aspect of the arXiv operation is the continued posting of unrefereed preprints, thus creating 'facts on the ground' supporting the claim that conventionally certified publications are irrelevant to daily scientific practice. One can then ask what influence arXiv has had on the role of conventional journal publications. Some indication of an answer may be found by consulting statistical data from the system taken at 3-month intervals (see Appendix: Tables A and B)[39]

The figures show for each archive in the database, how many papers are submitted per month for three years (2000, 2001 and 2002), and how many existing citation references have been added to the database, referred to at arXiv as the 'Journal-Ref'; in Tables A and B these are also presented in terms of percentages ('in print', '%').[40]

Tables A and B are two snapshots in time, showing the number of submissions for each archive and how many of them are already accepted for publication in some journal, while the comparison of Tables A and B shown in Table 1 shows the traffic of new journal citations over these three months. At first glance, an obvious diversity appears in Tables A and B with regard to the average percentage of papers that have been provided with citation references: the proportion is much greater for the 'high energy physics', 'nuclear physics' and 'general relativity & quantum cosmology' archives than for all the other archives. A second feature becomes evident in Table 1: a much greater proportion of citation references are added during the three months between April and

July 2003, for papers posted in 2002 than for papers originally posted over the other two years, and this applies to most of the archives.

The data in Table 1 can be explained by considering the time it takes to get a paper published, but normally a citation reference is available within one year and can in principle be uploaded at that point. The different rates between archives shown in Tables A and B are more complicated to explain, and the data invite further exploration of questions such as: (1) how to account for the enormous size difference between some of the archives; (2) how to explain the different upload patterns of papers and citation references; or (3) how to explain the large range of preprint/citation reference ratios. I will leave these questions for future research, and here examine what an automated daily upload of citation references may indicate.

The SLAC library reports daily to arXiv and provides data for an automated upload of new journal references. This SLAC service still primarily caters to HEP. There is much relevant material for some other disciplines, such as condensed matter physics and astrophysics, and SLAC attempts to cover papers on condensed matter that are cross listed into other archives of special interest to the HEP community. Furthermore, in connection with a recently established collaboration with Fermilab, SLAC decided to start uploading the full astrophysics archive.[41] When a similar status check was done in spring 2002 and shown to advocates of the arXiv system, the immediate response was that more is published in some journals. The rest are mostly dissertations, proceedings, and other materials that are not likely to be published in journals. The reason for this is that the automated processing is unlikely to find every reference that exists. And the much lower figures for the other archives should be explained by the lack of an automated citation harvesting and upgrading feature.[42] When a coordinator for mathematics papers was shown these data, he responded with a similar comment to the effect that the low citation reference figures for mathematics do not present everything that is published in conventional journals.[43] This is because no automated harvesting of citation references is implemented for mathematics. Although it

Two extracts, April 2003 & July 2003	Posted on arXiv 2000		Posted on arXiv 2001		Posted on arXiv 2002	
	Available citation references in each extract %	Difference	Available citation references in each extract %	Difference	Available citation references in each extract %	Difference
Astro-ph	20.81 -> 21.85	1.04	20.41 -> 21.01	0.6	28.99 -> 32.03	3.04
Cond-mat	36.69 -> 37.40	0.71	37.67 -> 39.10	1.43	26.83 -> 32.74	5.91
Cs	29.82 -> 30.04	0.22	29.84 -> 30.02	0.18	27.84 -> 28.52	0.68
Gr-qc	71.30 -> 71.86	0.56	66.72 -> 67.49	0.77	50.83 -> 59.29	8.46
Hep-ex	62.43 -> 62.57	0.14	54.10 -> 54.94	0.84	39.10 -> 48.77	9.67
Hep-lat	78.46 -> 78.46	0	86.71 -> 87.22	0.51	32.27 -> 37.14	4.87
Hep-ph	68.91 -> 69.00	0.09	69.14 -> 70.32	1.18	48.32 -> 60.04	11.72
Hep-th	75.04 -> 75.29	0.25	75.86 -> 76.58	0.72	56.50 -> 64.60	8.1
math	17.74 -> 18.44	0.7	14.12 -> 15.14	1.02	09.21 -> 10.43	1.22
Math-ph	38.65 -> 40.34	1.69	31.08 -> 33.67	2.59	26.01 -> 29.79	3.78
Nlin	36.02 -> 36.16	0.14	30.48 -> 30.91	0.43	19.01 -> 23.16	4.15
Nucl-ex	74.73 -> 74.73	0	71.21 -> 71.98	0.77	37.46 -> 59.28	21.82
Nucl-th	74.27 -> 74.37	0.1	73.79 -> 74.42	0.63	47.88 -> 61.10	13.22
Physics	43.17 -> 43.43	0.26	32.47 -> 33.28	0.81	18.74 -> 21.16	2.42
Quant-ph	38.50 -> 39.08	0.58	37.29 -> 38.48	1.19	25.46 -> 30.94	5.48

Table 1
Increase in available citation references in April-July 2003, for papers originally published in 2000, 2001 and 2002. Courtesy of Cornell University Library.

Notes: See Tables A and B in Appendix. There is a clear increase in references added to the 2002 papers over this period. The trend is most noticeable for papers submitted into the 'hep', 'nucl', 'gr-qc', 'quant-ph', 'cond-mat', and 'nlin' archives.

seems clear that almost every archived paper eventually goes into conventional publication, the problem remains to better automate the 'harvesting' of information on journal publication.[44]

If the explanations provided by arXiv advocates are taken seriously, two challenging questions arise. (1) If the centralized and automated author self-archival system can provide all the mediation needed for communicating and sufficiently certifying ongoing research, why do all the scientists still submit to journals?(2) If journal publications are indeed so important to scientists, why are submitters so lax about manually providing citation references to their already posted submissions in arXiv?

In the following section, I will argue that the utilization of new, effective and fascinating facilities such as arXiv has had very little impact on the broader social and cultural meaning of conventional journal publication. To simplify greatly, one can say arXiv represents an evaluation and local agreement of 'significance', while conventional journal publication represents an evaluation and broader social agreement about 'verity'.[45]

The Certified Journal is a Cultural Object

The early sociology of science identified the origin of the conventional refereeing system with the developing science societies of the seventeenth century. Institutionalisation of peer review came with the first journal publication, the 'Philosophical Transactions' of the Royal Society - more precisely, with the control that its council possessed over the content of the journal (Zuckerman & Merton, 1971). Ever since then, journal publications have played an important role in identifying a scientific elite. The early literature agreed that the process of peer evaluation, by virtue of the culturally granted authority of formal refereeing, provided the warrant for scientists to build on the cumulative work of their peers. It also offers a status hierarchy based on honour and esteem.[46]

Literary strategies play an essential role in the conversion of research into publication. Such strategies mediate scientific research and reasoning because much of the ongoing work during a research period is forgotten when the publication process, framing those events, takes over.[47] Shapin and Schaffer's (1985; Shapin, 1984) notion of 'virtual witnessing' describes a literary technology used as part of the early experimental life which contrasts from 'direct' witnessing of an experiment. Given the limited access provided by an account of an experiment, the importance of having credible witnesses, and the limitations to a replication of experiments, literary technology became a means to broaden the circle of witnesses. The style of writing and the introduction of pictorial illustrations helped to establish an authority of the written paper.

Virtual witnessing relieves the tension of being excluded from the core group of direct witnesses, but contemporary esoteric science introduces the further problem of expert readership, which imposes a tension between core-groups and surrounding social domains. Even the *virtual* witnessing of research results in contemporary physical science and mathematics for example, usually limited to core-group members and their research associates; people outside the inner circle are unlikely to find such results intelligible or useful. The core group's literary technology produces more or less meaningless prose to all but those that have been extensively trained and socialised into a highly specialized practice.[48] Although colleagues in the close-knit networks do not even read conventional journal publications, because they have long since read the preprints, it remains the case that scientists are not promoted and do not receive grants and other professional rewards if they do not produce conventionally certified publications in respectable journals. The arXiv system has had very little if any influence upon such professional certification. Advocates of the system are aware of this fact: they recognize the added value of journal publication for (in their terms) 'extra-social' purposes. For example, a mathematician I corresponded with described two stages of promotion at the mathematics department of the University of California, Davis:

> Both merits and promotions are first voted on by the department. If the vote is (sufficiently) favourable, the chair writes a letter to the administration recommending the action. Then the dean approves the action on the advice of an 'ad hoc' committee. In the case of UC Davis mathematics, the Dean

> oversees *math*, *physics*, *chemistry*, *statistics*, and *geology*, and the 'ad hoc' committee is drawn only from these departments. Merits stop there. Promotions pass on to the next level, at which the committee on academic personnel' (CAP) advises a 'vice provost'. CAP is a campus-wide 9-member committee. It is as likely to have a *historian* or a *veterinarian* as a mathematician.[49] [emphasis added]

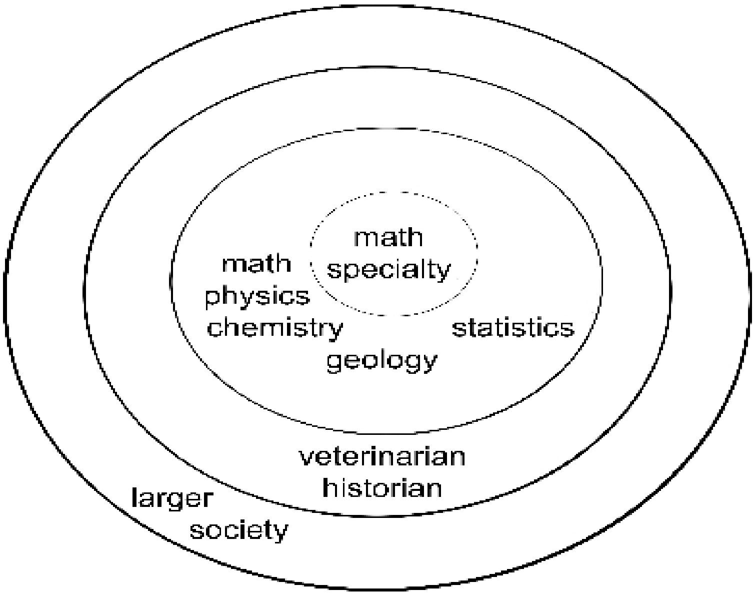

FIGURE 3
Core-group scientists in an academic institutional setting

When a bull's eye model for core-groups and their surrounding social domains (Figure 3) is used to characterize the promotion of the hypothetical mathematics researcher at UC Davis described in the above quote, the certification provided with formal peer review and conventional journal publication clearly serves assessments by groups *outside* the core of the candidate's discipline.[50] A distinction can be drawn in this case between (1) scientists as expert readers and (2) scientists as authors.[51] The reader uses the electronic medium interactively, and does not necessarily keep track of paper versions or citation references. An arXiv user reads preprints and consults directly with colleagues to resolve any uncertainties. However, when acting as authors, scientists have no choice but to care greatly whether or not their papers have been or will be formally published, because they must compile a record of research activities in a widely recognized and culturally entrenched form, recognizable to 'outsiders'.[52] This is an odd situation.

Core-groups in a given science do not wait for certified publication before consulting research results, and have more or less ceased reading journal papers in their final published form. Once certified, such papers have no direct 'use value' for hands-on practice. However, even though journal publication remains crucial for short-term professional evaluations, for longer-term historical judgements of a given contribution – determining whether or not it signifies a 'permanent' contribution and represents 'excellence' – such publication may provide a premature indication, and is thereby not very 'useful' either.

It is entirely possible to run into a dilemma when some of the consequences of the current situation are examined further. Assume that arXiv's proponents are right when they claim that the only purpose remaining for a costly formal peer review process is to certify members and maintain order in a discipline. The question then is: Is the validation process that is now in place through the electronic and networked system not adequate for that purpose as well? The necessary facility may be already be present, in a crude way, so that the only thing needed to certify individual papers would be something like an automated online expert voting system with regular deadlines. There would not be any need to publish papers separately, in print or electronically, through conventional peer review processes, because the papers are already in circulation and the real problem is simply to establish a way to formally encode certification that particular contributions have been graded and passed inspection. Consequently, a researcher would be able to produce the certificate when applying for grants or promotion. For arXiv advocates, this purpose could be accomplished through a simple addition to the arXiv preprint identifier, which would look something like a grade letter 'V' for verified, 'yy-mm' for year and month of certification, and a two step evaluation, 'nn-nn' checked off for general confirmation of important results and un-checked general confirmation of expected results.[53] A before and after example would look like this:

> **Journal-ref:** Nucl.Phys. B648 (2003) 131-160
> **Certification-ref:** hep-th/0209045: V 03-05, 3-1

Formally, both resemble certificates, and there is nothing to prevent the process from being implemented, since all the relevant technologies are available and the argument for economic and practical feasibility is widely understood. A problem does not arise until the current state of the world gently whispers that the '**Journal-ref**' is actual code but the '**Certification-ref**' is not. '**Certification-ref:** hep-th/0209045: V 03-05, 3-1' is currently only an imagined code and it remains to be determined who are the expert voters.[54]

The divorce of the two values associated with publication - immediate efficiency in practice, that is to say direct *use value* on one hand, and social-cultural *token value*, on the other hand - might explain why arXiv patrons are not under much pressure to manually upload their citation references to arXiv. The citation references are authoritative cultural objects, not useful objects like preprints in the 'lab'.[55] This also leaves unanswered a question about the extent to which traditional journal publications lend authority to the preprint exchange. Could the preprint exchange possibly exist without them? Historically, the two systems have operated in connection with one another. DESY collected both published and unpublished works and so did CERN. PPF and PPA were two components of the same publication, and when examined closely it appears that the authority of the arXiv repository depends on citation references.[56]

Traditions in journal publication only reinforce the social mechanisms already at work in all knowledge cultures: a pecking order in the professional life based on the evaluation of knowledge-production.[57] Scientists are sensitive to notions of prestige and status: they want the best references and citations from authoritative sources. Conventional journals confer status for papers, and colleagues certainly notice where their peers are published. It may be a laudable goal to defy the practices of profiteering trade publishers, particularly in response to an unpleasant monopoly-like situation. But it would pose a serious problem to be in direct opposition to the publishing practices of learned science societies, such as the British Royal Society or the American Physical Society, with their firmly rooted and broadly recognized authority and long tradition.

Concluding Remarks

Revolutions do not happen overnight. Subscriptions to print journals decrease steadily by approximately 3-5% per year, and online subscriptions are slowly taking over.[58] More peer-reviewed papers are made freely available on the web, new online-only journals crop up and arXiv grows steadily, disseminating submissions from all around the world.[59] A survey in 2002 showed that respondents across all disciplines overwhelmingly agree (either moderately or strongly) that they are confident that they can find more relevant information on the Internet than was available only 2 years earlier, and they overwhelmingly disagree (either moderately or strongly) that the Internet has not changed the way they use libraries.[60] Publishers of the learned science society journals whose members use arXiv have accepted the preprint culture. For example, the American Physical Society journals have allowed direct electronic submissions from arXiv since November 1996, the *Journal of High Energy Physics* (JHEP) since 1997, and the UK Institute of Physics' journals have done so since March 2001.[61]

I have attempted to show how arXiv differentiates between solutions to what I call epistemological and logistical objections. It may be unnecessary to answer the epistemological objection – that arXiv will disseminate poor science and misinformation – when the system is limited to the exchange of research literature among members of core-groups. One reason why this is possible is because formal peer review only checks papers for obvious errors and potential significance for ongoing research. What is correct and useful science on the other hand, must be determined by members of core-sets and core-groups, who usually are the only readers with sufficient interest and technical know-how to assess the usefulness, value and validity of the contents. If they can get their hands on fresh research results, they prefer not to wait for that research to appear in a journal.

Logistical controls – designed to forestall what I call the logistical objection that arXiv will produce a flood of irrelevant papers – are managed separately but their operation still remains problematic. While arXiv attempts to automate the vetting of membership, relevancy, and even subject matter, human input is needed to complete these tasks adequately. Furthermore, broad social agreement is lacking on how logistical controls should be managed for certifying or validating the literature in a wider social–cultural context. These observations call for a comparison of communication networks in academia. How are networks constituted and maintained in disciplines other than those discussed in this paper? How esoteric and close-knit do the disciplines grow and consequently, to what extent are quick 'on-topic' or 'off-topic' selections used in their publishing practices? In other words, how are controls implemented in the everyday scholarly practices and the editorial settings?

While these questions remain unanswered, the relationship between unrefereed arXiv preprints and their formally refereed counterparts also remains unclear. The arXiv case demonstrates how validity of research is granted on two levels, first in the open exchange of unrefereed papers and later by way of scientific publications that lend the vital token value to the work. The existence of the latter has granted the former the authority to prevail and become an established structure for communicating and validating research, and it still does, while at the same time the demand for rapid exchange and implementation of ever more efficient logistical controls becomes the norm of the scientific practice, weakening the status of traditional publication. The effects will be felt in wider social–cultural contexts and journals are prone to change.

There are further uncertainties about this two-headed process of research validation that arXiv has introduced. I will give three examples. (1) Citations play an important role inside and outside of core-groups to determine, at least to some extent, the influence of a given paper in a research community. ArXiv preprints by established authors may be heavily cited before they are formally refereed.[62] (2) In 2002, the journal *Classical and Quantum Gravity* published an paper by a group of gravitational wave physicists that caused much stir with some other members of the core-set. The group deliberately did not post the preprint in arXiv because they knew it would attract criticism from other core-set members and probably never be accepted. Members with opposing views felt forced to publish a formal response, whereas if the paper had only been posted with arXiv they could have directed their criticisms through informal channels.[63] (3) The 'reverse Alan Sokal hoax'[64] is a case that left open whether language can be employed to fake physics and get hoax papers through peer reviews. This case invites the speculation of what would have happened, had the works in question been submitted first to arXiv and that way exposed to the research community prior to formal review.

These three examples are all somewhere on the margins, where it is not clear what the authority is of either the journal publication or the unrefereed preprint exchange, but case studies of such incidents could perhaps bring to fore the essence of the problem of understanding the relationship between them. I like to argue that the arXiv case is enlightening in the sense that its operations have introduced a two-headed beast, while it remains largely hidden to what degree members of other academic sectors communicate their works-in-progress and rely on peers for validation while composing papers and writing reports. The arXiv case is also helpful in the sense that it demonstrates how aggregating and disseminating research literature is largely a matter of logistics. It is an attempt to organize and communicate mass amounts of mainstream works. ArXiv is very successful at doing so, and its advocates remain proud that almost all the papers eventually are published with reputable journals: 'The standard of work archived at Los Alamos is very high. 70% of papers are eventually published in journals and another 20% are in conference proceedings' (O'Connell, 2002).

Notes

I wish to thank all those who contributed to the writing of this paper by providing me with corrections and useful tips, by communicating individual issues over time and by reading over this text at various stages and pointing out to me ways to better capture the basic themes. In alphabetical order, Harry Collins, Robert Evans, Paul Ginsparg, Greg Kuperberg, Michael Lynch, Heath O'Connell, Andrew Odlyzko, Trevor Pinch, Thorsten Schwander, Simeon Warner and the anonymous referees. I also wish to thank the School of Social Sciences at Cardiff University for providing me with an office in the beautiful Glamorgan Building while I was working on the first draft of this paper under Harry Collins' supervision, my employers and co-workers at Cornell University Library, in particular Thomas Hickerson and Marcy Rosenkrantz, for all their patience, and my arXiv co-workers for putting up with a rather odd arXiv admin-persona for so long.

[1] Method of research: I was employed as a programmer/analyst specialist at Cornell University Library from January 2001 until August 2003. For almost two years during this time, I was assigned technical administration for supporting daily submissions to arXiv. I acquired first hand knowledge of the inner workings of the system as both a technical and a social object, and was involved in some of the institutional migration process from Los Alamos National Laboratories to Cornell University Library, including broad examination of system needs, planning new developments and addressing policy issues. The contents of this paper are based on participant observation (and comprehension; Collins, 1985:171-172). The text draws from my knowledge and experience, and from my understanding of the burning issues debated by advocates of the arXiv model and related initiatives. I draw from personal/professional communication that includes extensive unrecorded conversation in the workplace and off the record; in person, at meetings and over the phone. This communication is supplemented with notes and email exchanges to follow up on key topics. All statistical data that I have collected and processed are publicly accessible using Internet harvesting protocols and I contacted my correspondents regarding the use of spoken and written remarks and gave them copies of a draft to examine.

[2] Arpanet was conceived and planned by Joseph Carl Robnett, 'Lick' Licklider, and others, for the Information Processing Techniques Office (IPTO) of US defence research, of which Licklider was the first director. An interactive overview of the email and Internet and related histories is available at http://livinginternet.com.

[3] A trend in the development of the service was to add specialities that cater to interests of high-energy physics, although the growth of the system has stretched to other communities. Archives included as of October 2004 were:

High energy physics - theory, since 08/91

High energy physics - lattice, since 02/92

Mathematics (includes subject classes), since 02/92

High energy physics - phenomenology, since 03/92

Astrophysics, since 04/92

Condensed matter (includes subject classes), since 04/92

General relativity and quantum cosmology, since 07/92

Nuclear physics - theory, since 10/92

Nonlinear sciences (includes subject classes), since 01/93

Computer science (includes subject classes), since 01/93

High energy physics - experimental, since 04/94

Nuclear physics - experimental, since 12/94

Quantum physics, since 12/94

Mathematical physics, since 09/96

Physics (includes subject classes), since 10/96

Quantitative biology (includes subject classes), since 09/03

[4] See for example Bijker (1987), on the theory of invention and Hughes (1985, 1986) on ignoring disciplinary boundaries and on technology as problem-solving systems and means for reordering the world. Also see Woolgar (1991) on configuring users and machines.

[5] The notion of 'core-group' (a group of researchers and research units who actively and interactively engage in a scientific research area) was coined by Collins (1985; see discussion in Collins, 1999). The larger community that I refer to as 'HEP' comprises more than one core-group, and much of the literature that is of interest to that community is read across the boundaries of core-groups. This can surely raise a problem when preprints are inspected, but that problem is the same as with formally published works. In both cases uncertainty can arise, but I argue that the delay of formal publications gives arXiv salience for immediate research uses, and more immediate communication channels (consultancy for example) are used to settle uncertainties introduced with preprints.

[6] For an insiders' account of the early history of technical implementations for arXiv and its patrons, see Ginsparg at http://arxiv.org/blurb/blurb.ps.gz.

[7] The page number refers to the available online version at http://arxiv.org/blurb/blurb.ps.gz.

[8] This section largely follows a historical overview by Heath O'Connell (2002), published with High Energy Physics Libraries Webzine and available at http://webzine.web.cern.ch/webzine/6/papers/3/. It also draws from discussions with correspondents that helped clarify my own understanding of this and other historical records.

[9] Ibid., at http://webzine.web.cern.ch/webzine/6/papers/3/index.html#SPIRES%20and%20Internet.

[10] See Donald E. Knuth's web page at http://www-cs-faculty.stanford.edu/~knuth/, also about TeX use at http://www.tug.org/.

[11] The arXiv operation is all Linux based using mainly the Perl programming language, and it has until very recently been exclusively Unix/Linux oriented and expected its patrons to feed only source data from papers. Patrons have always been expected to be competent in scientific computing, which includes an understanding and being up-to-date about the operational details behind digital imaging and image conversion, writing and compiling TeX source, postscript interpretation of documents, the use of font files, document style files and other factors that are at play in the arXiv system.

[12] See https://en.wikipedia.org/wiki/Citebase.

[13] See their site at http://www.openarchives.org. This technology is referred to as aggregation services of digital libraries.

[14] On a technical note, it is interesting that, although the development and maintenance of the auto-TeX compiler has contributed to the 'open-source' GhostScript developments in accordance with 'open-source' ideology, and that the design of arXiv has been all free-ware computing, the arXiv system source has never been made available to the public under any scheme such as the Open Source Initiative, see http://www.opensource.org. One reason for this is that it has never been sufficiently complete as a system to be one single software package, which is transferable (and manageable) as such, but is a compilation of many such packages. See also http://sourceforge.net as one example of large scale, Open Source development organization.

[15] See Bijker 1987 and Pinch and Bijker 1984, 1987. But referring simply to a 'technological system' already suggests a system of social arrangements (Woolgar, 1991; MacKenzie & Wajcman, 1985).

[16] Quantitative biology started posting on arXiv in September 2003. See also the 'Math Front' at UC Davis for an entry to all the math specialties that use arXiv, at http://front.math.ucdavis.edu/.

[17] See Ginsparg (2001). He offers a detailed discussion of these figures. They are used in a talk given in 2001 and are clearly 'ball park' figures. However, they measure against APS revenues as an example of a professional society disseminating output, against Elsevier as an example of a for-profit trade publisher, and so forth. An average figure was estimated of about US$50,000 per produced scientific paper in HEP. This cost would typically cover the expenses of research: namely, salaries, overhead and experimental equipment, but not the editorial, printing and distribution costs of the final paper. In other words, the output would still need to be published. Then a comparison was made between a pricey trade publisher in the field with US$10,000-20,000 per paper in revenues, a professional society in physics estimated to generate about US$2,000 per paper in revenues, an operation that takes data feed from an existing print publisher and converts it to Web readable formats, operating at approximately US$100 per paper and more. It was also pointed out that managing the conventional editorial process could never cost less than around US$1000 for each peer reviewed and published physics paper, given the way this process is generally orchestrated. However a point was made in personal communication about this particular claim, to the effect that the figure reflects the cost of peer review in physics by the American Physical Society journals, and that the cost of the peer review process can vary greatly, depending on discipline and the type of process.

[18] For example Greg Kuperberg, professor of mathematics at UC Davis, argues that open reviews, such as Math Reviews and Zentralblatt, provide a much more helpful screening and evaluation of the mathematical literature than the anonymous refereeing system is capable of (Kuperberg, 2002).

[19] See also Kuperberg (2002). Both Ginsparg and Kuperberg argue that there exists a persistent and misleading idealisation of the role of formal peer review.

[20] Ginsparg, (2003) asks if the peer review could be better 'focused', as he puts it.

[21] The author probed specially for this view among arXiv programmers in a project meeting with them 12th of February 2002.

[22] See Traweek (1988), chapter 3, on this matter in the HEP community. She speaks at length about the sorting mechanisms and, interestingly, in her observation the young physicists are never made aware of this process. It is often only the 'bright and blunt bastards' that catch on to what is expected of them.

[23] For an interesting input, see Bachrach et al. (1998).

[24] To take some examples, the home of *Information Research: an International Electronic Journal*, offers a title page of well-known electronic journals that are peer reviewed but free of charge, at http://informationr.net/ir/titlepages.html. The Public Library of Science (PloS) is another initiative that calls on scientists in biology and medical research to resist the privatisation of scientific knowledge, see http://www.plos.org. Trade publishers have been able to establish a monopoly-like situation because university libraries and research institutions have no choice but to subscribe to academic journal publications in order to uphold proper status of their institutions. See also http://pubmedcentral.org an archive of peer-reviewed life science journals that are free of charge. Stevan Harnad is one pioneer who should also be mentioned: he has written numerous papers on the subject; see for example, Harnad (1998, 2001). His contributions include among other things the http://www.eprints.org/uk/ project. This project is dedicated to opening access, through so-called distributed author-institution self-archiving and is meant to operate alongside the formal peer review process and to take advantage of reference linking and aggregation services to maximize the distribution of works. He is also the founder of CogPrints.org.

[25] Just to give one example, the engineering library at Cornell University cancelled 35 subscriptions for an annual charge of US$59,000 in autumn 2003. According to a statement from the director of collection management, electronic-only format at that point was priced at 115% of the cost of a paper subscription with the same trade publisher.

[26] See Kling, Fortuna & King (2001) for an interesting treatment of the E-Biomed experiment for medical research (a mix of refereed and unrefereed works), which later became a non-profit scientific publishing of peer-reviewed works as the Public Library of Science (PloS) at http://www.plos.org/, and also PubMed Central, whose service offers free access to already existing life science journals at http://www.pubmedcentral.org.

[27] See for example Knorr-Cetina (1999) for comparisons of the cultures of high energy physics and molecular biology; also Collins (1999), on expert readership; Hilgartner (1997), on the Sokal affair; and Kling, Fortuna & King (2001), on the failure of the unrefereed exchange service of the E-Biomed proposal.

[28] See for example the first three pages of http://arxiv.org/pdf/hep-ex/0309017 (listing the BABAR Collaboration).

[29] Knorr-Cetina (1999). Chapters 7 & 8 discuss the notion of erasing and restoring the individual as an epistemic subject.

[30] Somewhat related to this is an interesting discussion about the debate between Thomas Hobbes and the experimentalists over the notion of 'public' in pubic witnessing in the early experimental life (Shapin and Schaffer, 1985). It is clear that high-level competence is mandatory in order to evaluate works, whether by direct inspection or aided by consultation from colleagues.

[31] See Hughes (1985, 1987) for an interesting discussion about system builders, ignoring disciplinary boundaries. He also suggests that analytical categories (political, science, technology, economics) should be used sparingly.

[32] See for example Bijker (1987), for a theory of invention; also Pinch and Bijker (1984, 1987) on the social construction of facts and artefacts; and MacKenzie and Wajcman (1985) for introductory essays.

[33] MacKenzie and Wajcman (1985) talk about building on already existing solutions, and about gradual developments, new combinations and new configurations, for which existing technology is a precondition. Hughes (1987) suggests that systems with high momentum tend to exert soft determinism on other systems, groups and individuals.

[34] Star and Griesemer (1989) discuss boundary objects, such as repositories and descriptive objects, and argue that they are plastic enough to adapt to various local needs, but robust enough to maintain common identities across sites. Seen in relationship to methods standardization, they point out that consensus is not a necessary condition for entities to cooperate, or for successful conduct.

[35] See for example Woolgar (1991), MacKay and Gillespie (1992) and Pinch and Trocco (2002). Also see Bohlin (2004) for a discussion of stabilization processes in the competition between new science communication regimes, as he describes them. Bohlin also points out how his use of the SCOT model is untypical, precisely because so much is unsettled; but he importantly points out that it will not be some sort of a system superiority that determines the outcome.

[36] A sizeable group of authors have problems using TeX when it comes to uploading their source files into an automated TeX compiler at arXiv. Astrophysicists have openly complained about having to compress their figure plots and colour photos 'more efficiently' according to the arXiv technical administration. See also an interesting and related discussion in Pinch and Trocco (2002) about how users of music synthesizers mostly use pre-set sounds and do not take advantage of the possibility to develop new sounds with the instrument.

[37] Greg Kuperberg, professor of mathematics at UC Davis leads a group of moderators for mathematics specialities who have shaped this process to their own special needs. Kuperberg also was commissioned to assist with an ongoing research project on the development of the auto-TeX system and the TeX services.

[38] It is also very interesting to look more closely at the involvement, over time, of the areas of physics that are remote from HEP; which already indicates a divorce of designers from patrons. Designers shift to operating as a group of socially and culturally distant software engineers in the frustrating situation of attempting to cater to needs of a largely unknown body of customers. It is a well-known situation that leads to all sorts of unpredictable and problematic results. See an interesting sociological perspective in Woolgar (1991).

[39] Refer to the following translations of the abbreviations used for the archives in the system:

astro-ph,	- - - -	Astrophysics
cond-mat,	- - - -	Condensed Matter
cs,	- - - -	Computer Science
gr-qc,	- - - -	General Relativity and Quantum Cosmology
hep-ex,	- - - -	High Energy Physics – Experiment
hep-lat,	- - - -	High Energy Physics – Lattice
hep-ph,	- - - -	High Energy Physics – Phenomenology
hep-th,	- - - -	High Energy Physics – Theory
math,	- - - -	Mathematics
math-ph,	- - - -	Mathematical Physics
nlin,	- - - -	Nonlinear Sciences
nucl-ex,	- - - -	Nuclear Experiment
nucl-th,	- - - -	Nuclear Theory
physics,	- - - -	Physics
quant-ph,	- - - -	Quantum Physics

[40] The so-called 'Journal-Ref' in arXiv is strictly used for existing journal citations when they become available and those references are either provided by services that automatically harvest them (SLAC) or they are provided manually by the authors themselves. Consequently, the column marked '%in print' in the Appendix (Tables A and B) refers to the percentage of papers in the database that are already published.

[41] From phone conversation and email exchange with library personnel at SLAC

[42] From phone conversation, office talk and email exchange with arXiv advocates and personnel.

[43] Greg Kuperberg at UC Davis manages this work and he also maintains the UC Davis Math Front to math papers in arXiv, see their site at http://front.math.ucdavis.edu/.

[44] From phone conversation, office talk and email exchange with arXiv advocates and personnel.

[45] Andrew Odlyzko framed a similar distinction in his paper 'Peer review and non-peer review', between significance and correctness, as he puts it, Odlyzko (2001b).

[46] See also Hagstrom 1965, chapters I & II.

[47] See Knorr-Cetina (1981), Latour and Woolgar (1979) for discussions of the scientific paper as an end product of research and a retrospective construction of that research.

[48] An interesting borderline case here, as regards physics, would be the story of the Bogdanov brothers. Apparently they employed the correct language and earned a degree in physics, but left the physics community uncertain as to whether their work was a hoax. One measure is to say that they never worked with anyone, and were never in collaboration with known bodies, which is usually questionable in the larger research community. A good entry to the Bogdanov case is a letter from John Baez to the Google newsgroup 'sci.physics.research' 2002-10-23, pointing out the possible hoax. It has links to their thesis, CVs and four paper references, see http://groups.google.com/groups?q=g:thl3105378894d&dq=&hl=en&lr=&ie=UTF-8&safe=off&selm=ap7tq6%24eme%241%40glue.ucr.edu

[49] From email exchange with an arXiv moderator, June 2003.

[50] This example is only one of many promotion mechanisms in academia, which apparently are quite similar from one institution to another, and on both sides of the Atlantic. Such similarity supports the main argument here, that academic status and prestige are obtained in an interdisciplinary social setting. However, there are outer social circles that differ importantly from those identified here. Two of my correspondents pointed out that the peer review process has direct use value, for example in mathematics and some areas of physics. Large numbers of scientists live on the outskirts of the practice, and therefore they rely on bibliographies, citation figures, and citation references in their work. For example, teachers in universities and institutes of higher education who teach the science to a certain degree, but are themselves not contributors to the construction of scientific knowledge.

[51] See for example Kuperberg (2002) and Guédon (2001).

[52] See also Ginsparg (2003) where he talks about a 'desired signal' that individual researchers or groups wish to or are required to 'transmit'.

[53] This is of course a playful example, but the general idea is in accordance with the type of peer review that physicists and mathematicians, using arXiv, can and do expect from conventional publications. The content is usually only checked for relevance to ongoing research; that is to say, it is accepted if it is interesting and not obviously wrong.

[54] Ginsparg (2003) actually discusses the possibility of gaining better focus to peer review process, and he asks if it is really the case that the 'desired signal' can only come from the traditional publication practices that cost $1000 for each paper.

[55] There is likely to be more to this situation. While the lack of motivation can be directly associated with career care-taking, two of my correspondents said they would like to see further studies on differences in rank or in age group, for example, given that younger persons might be more eager to supply an electronic system with a journal reference, in hope of attracting more readers through the online system, while established scientists can rely on their reputation and other channels. Also, there may be some relationship between the fact that the older generations used to have secretaries to typeset their papers and generally take care of things like getting papers to publishers and polishing curricula vita, while the younger and computer savvy generations have taken on these tasks themselves.

[56] This valuable point was made to me by Neil Stephens during a session in which I introduced a draft of this paper to colleagues in the Knowledge, Expertise and Science (KES) group at Cardiff University.

[57] See Guédon, 2001, chapter III.

[58] See Odlyzko (2001a) for statistics on the use of scientific literature in different formats and discussion about slow and fast changes.

[59] See http://arxiv.org/Stats/au_all.html.

[60] The full set of all 659 tables (35 questions) is available at http://www.diglib.org/pubs/scholinfo, Friedlander 2002.

[61] See also Bohlin (2004) for discussion of this trend in journal policy.

[62] A correspondent pointed out a single case in arXiv from November 1997 http://arXiv.org/hep-th/9711200 that wasn't formally published for roughly a year, but was the most cited paper in 1998. The author had 26 papers in arXiv dated before that one, and of these 25 published with elite publications, most of them already published at that point. Currently SLAC shows over 3250 citations of this one paper.

[63] From personal communication with an associate at Cardiff University. This paper is by P. Astone, D. Babusci, M. Bassan, P. Bonifazi, P. Carelli, G. Cavallari, E. Coccia, C. Cosmelli, S.D'Antonio, V. Fafone, G.Federici, S.Frasca, G. Giordano, A. Marini, Y. Minenkov, I. Modena, G. Modestino, A. Moleti, G. V. Pallottino, G. Pizzella, L.Quintieri, A.Rocchi, F. Ronga, R. Terenzi, G.Torrioli, M. Visco, co-authored 2002, available at http://arxiv.org/gr-qc/0210053.

[64] This example refers to the case of the Bogdanov brothers, see footnote 48

References

Bachrach, Steven, R. Stephen Berry, Martin Blume, Thomas von Foerster, Alexander Fowler, Paul Ginsparg, Stephen Heller, Neil Kestner, Andrew Odlyzko, Ann Okerson, Ron Wigington, Anne Moffat (1998) 'Who Should "Own" Scientific Papers?' *Science* 281(5382): 1459-60, available at www.sciencemag.org/cgi/content/full/281/5382/1459.

Bijker, Wiebe E. (1987) 'The Social Construction of Bakelite: Toward a Theory of Invention', in Wiebe E. Bijker, Thomas P. Hughes & Trevor J. Pinch (eds), *The Social Construction of Technological Systems* (Cambridge, MA: MIT Press):159-87.

Bohlin, Ingemar (2004) 'Communication Regimes in Competition: The Current Transition in Scholarly Communication Seen Trough the Lens of the Sociology of Technology', *Social Studies of Science* 34(3): 364-91.

Collins, Harry. M. (1985) *Changing Order: Replication and Induction in Scientific Practice* (Chicago, IL: University of Chicago Press).

Collins, Harry. M. (1999) 'Tantalus and the Aliens: Publications, Audiences and the Search for Gravitational Waves', *Social Studies of Science* 29(2): 163-97.

Friedlander, Amy (2002) *Dimensions and Use of the Scholarly Information Environment: Introduction to a Data Set Assembled by the Digital Library Federation and Outsell, Inc.*, available at www.clir.org/pubs/abstract/pub110abst.html.

Ginsparg, Paul (1994) 'First Steps Towards Electronic Research Communication', *Computers in Physics* 8(4): 390-96, (originally adapted from a letter to *Physics Today,* June 1992), available at http://arxiv.org/blurb/blurb.ps.gz.

Ginsparg, Paul (2001) 'Creating a Global Knowledge Network', Invited contribution for Conference held at UNESCO HQ, Paris, 19-23 Feb 2001, *Second Joint ICSU Press - UNESCO Expert Conference on Electronic Publishing in Science, during session Responses from the scientific community, Tue 20 Feb 2001*, available at http://arxiv.org/blurb/pg01unesco.html.

Ginsparg, Paul (2003) 'Can Peer Review be better Focused?', available at http://arxiv.org/blurb/pg02pr.html.

Guédon, Jean-Claude (2001) *In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing*, published by the Association of Research Libraries (ARL), 2001. Available at the ARL website: http://www.arl.org/storage/documents/publications/in-oldenburgs-long-shadow.pdf.

Hagstrom, Warren O. (1965) *The Scientific Community*, (NY: Basic Books).

Harnad, Stevan (1998, 2001) 'For Whom the Gate Tolls? Free the Online-Only Refereed Literature' *American Scientist Forum*, available at http://users.ecs.soton.ac.uk/harnad/Tp/resolution.htm.

Hilgartner, Stephen (1997) 'The Sokal Affair in Context', *Science, Technology & Human Values* 22(4): 506-22.

Hughes, Thomas P. (1985) 'Edison the Electric Light', in Donald MacKenzie & Judy Wajcman (eds), *The Social Shaping of Technology* (Milton Keynes, Bucks.: Open University Press): 39-53.

Hughes, Thomas P. (1987) 'The Evolution of Large Technological Systems', in Wiebe E. Bijker, Thomas P. Hughes, Trevor J. Pinch (eds), *The Social Construction of Technological Systems* (Cambridge, MA: MIT Press).

Kling, Rob, Joanna Fortuna and Adam King (2001) 'The Remarkable Transformation of E-Biomed into PubMed Central', *CSI Working Paper, Center for Social Informatics, Indiana University*, available at https://scholarworks.iu.edu/dspace/handle/2022/170.

Knorr-Cetina, Karin (1981) *The Manufacturer of Knowledge: An Essay on the Constructivist and Contextual Nature of Science* (Oxford: Pergamon Press).

Knorr-Cetina, Karin (1999) *Epistemic Cultures: How the Sciences Make Knowledge* (Cambridge, MA: Harvard University Press).

Kuperberg, Greg (2002) 'Scholarly Mathematical Communication at a Crossroads', available at http://arxiv.org/math.HO/0210144.

Latour, Bruno and Steve Woolgar (1986 [1979]) *Laboratory Life: The Construction of Scientific Facts*, 2nd edn. (Princeton, NJ: Princeton University Press).

MacKay, Hughie and Gareth Gillespie (1992) 'Extending the Social Shaping of Technology Approach: Ideology and Appropriation', *Social Studies of Science* 22(4): 685-716.

MacKenzie, Donald & Judy Wajcman (1985) 'Introductory Essay: The Social Shaping of Technology', in Donald Mackenzie & Judy Wajcman (eds), *The Social Shaping of Technology* (Milton Keynes, Bucks.: Open University Press).

O'Connell, Heath B. (2002) 'Physicists Thriving with Paperless Publishing', *High Energy Physics Webzine* 6, available at http://webzine.web.cern.ch/webzine/6/papers/3/.

Odlyzko, Andrew M. (2001a) 'The rapid evolution of scholarly communication', *Learned Publishing,* 15(1): 7-19. Available at www.dtc.umn.edu/~odlyzko/doc/recent.html.

Odlyzko, Andrew M. (2001b) 'Peer and Non-Peer Review', in F. Godlee and T. Jefferson (eds), *Peer Review in Health Sciences, 2nd ed.*, (BMJ Books): 309-11. Available at www.dtc.umn.edu/~odlyzko/doc/peer.review.txt.

Pinch, Trevor J. and Wiebe E. Bijker (1987) 'The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other' in Wiebe E. Bijker, Thomas P. Hughes & Trevor J. Pinch (eds), *The Social Construction of Technological Systems* (Cambridge MA: MIT Press).

Pinch, T. and Frank Trocco (2002) *Analog Days: The Invention and Impact of the Moog Synthesizer* (Cambridge, MA & London: Harvard University Press).

Shapin, Steven (1984) 'Pump and Circumstance: Robert Boyle's Literary Technology', *Social Studies of Science* 14: 481-520.

Shapin, Steven and Simon Schaffer (1985) *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life* (Princeton, NJ: Princeton University Press).

Star, Susan Leigh and James R. Griesemer (1989), 'Institutional Ecology, "Translations" and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39', *Social Studies of Science* 19(3): 387-420.

Traweek, Sharon (1988) *Beamtimes and Lifetimes: The world of High Energy Physics* (Cambridge, MA: Harvard University Press).

Woolgar, Steve (1991) 'Configuring the User: The Case of Usability Trials', in John Law (ed.), *A Sociology of Monsters: Essays on Power, Technology and Domination* (London and New York: Routledge): 58-97.

Zuckerman, Harriet A. and Robert Merton (1971) 'Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System', *Minerva* 9(1): 66-100.

Appendix

Table A

arch-ive/year	Online (N)	In print (N)	In print (%)	arch-ive/year	Online (N)	In print (N)	In print (%)	arch-ive/year	Online (N)	In print (N)	In print (%)
astro-ph/2000				astro-ph/2001				astro-ph/2002			
Total per year	6358	1323		Total/year	6819	1392		Total/year	7040	2041	
Average per month	530	110	20.81	Average per month	568	116	20.41	Average per month	587	170	28.99
cord-mat/2000				cond-mat/2001				cond-mat/2002			
Total per year	6037	2215		Total per year	7038	2651		Total per year	7780	2087	
Average per month	503	185	36.69	Average per month	587	221	37.67	Average per month	648	174	26.83
cs/2000				cs/2001				cs/2002			
Total per year	456	136		Total per year	553	165		Total per year	589	164	
Average per month	38	11	29.82	Average per month	46	14	29.84	Average per month	49	14	27.84
gr-qc/2000				gr-qc/2001				gr/qc/2002			
Total per year	1265	902		Total per year	1298	866		Total per year	1324	673	
Average per month	105	75	71.30	Average per month	108	72	66.72	Average per month	110	56	50.83
hep-ex/2000				hep-ex/2001				hep-ex/2002			
Total per year	748	467		Total per year	830	449		Total per year	810	316	
Average per month	62	39	62.43	Average per month	69	37	54.10	Average per month	68	26	39.01
hep-lat/2000				hep-lat/2001				hep-lat/2002			
Total per year	520	408		Total per year	587	509		Total per year	595	192	
Average per month	43	34	78.46	Average per month	49	42	86.71	Average per month	50	16	32.27
hep-ph/2000				hep-ph/2001				hep-ph/2002			
Total per year	4136	2850		Total per year	4238	2930		Total per year	4257	2057	
Average per month	345	238	68.91	Average per month	353	244	69.14	Average per month	355	171	48.32
hep-th/2000				hep-th/2001				hep-th/2002			
Total per year	3157	2369		Total per year	3194	2423		Total per year	3345	1890	
Average per month	263	197	75.04	Average per month	266	202	75.86	Average per month	279	158	56.50
math/2000				math/2001				math/2002			
Total per year	2999	532		Total per year	3237	457		Total per year	4246	391	
Average per month	250	44	17.74	Average per month	270	38	14.12	Average per month	354	33	9.21
math-ph/2000				math-ph/2001				math-ph/2002			
Total per year	533	206		Total per year	502	156		Total per year	715	186	
Average per month	44	17	38.65	Average per month	42	13	31.08	Average per month	60	16	26.01
nlin/2000				nlin/2001				nlin/2002			
Total per year	683	246		Total per year	689	210		Total per year	747	142	
Average per month	57	21	36.02	Average per month	57	18	30.48	Average per month	62	12	19.01
nucl-ex/2000				nucl-ex/2001				nucl-ex/2002			
Total per year	186	139		Total per year	257	183		Total per year	307	115	
Average per month	16	12	74.73	Average per month	21	15	71.21	Average per month	26	10	37.46
nucl-th/2000				nucl-th/2001				nucl-th/2002			
Total per year	987	733		Total per year	954	704		Total per year	1059	507	
Average per month	82	61	74.27	Average per month	80	59	73.79	Average per month	88	42	47.88
physics/2000				physics/2001				physics/2002			
Total per year	1156	499		Total per year	1238	402		Total per year	1238	232	
Average per month	96	42	43.17	Average per month	103	34	32.47	Average per month	103	19	18.74
quant-ph/2000				quant-ph/2001				quant-ph/2002			
Total per year	1530	589		Total per year	1923	717		Total per year	2188	557	
Average per month	128	49	38.50	Average per month	160	60	37.29	Average per month	182	46	25.46

Notes: The 'average per month' figures for 'in print' are rounded.

Table B

arch-ive/year	Online (*N*)	In print (*N*)	In print (%)	arch-ive/year	Online (*N*)	In print (*N*)	In print (%)	arch-ive/year	Online (*N*)	In print (*N*)	In print (%)
astro-ph/2000				astro-ph/2001				astro-ph/2002			
Total per year	6358	1389		Total/year	6819	1433		Total/year	7040	2255	
Average per month	530	116	21.85	Average per month	568	119	21.01	Average per month	587	188	32.03
cord-mat/2000				cond-mat/2001				cond-mat/2002			
Total per year	6037	2258		Total per year	7038	2752		Total per year	7780	2547	
Average per month	503	188	37.40	Average per month	587	229	39.10	Average per month	648	212	32.74
cs/2000				cs/2001				cs/2002			
Total per year	456	137		Total per year	553	166		Total per year	589	168	
Average per month	38	11	30.04	Average per month	46	14	30.02	Average per month	49	14	28.52
gr-qc/2000				gr-qc/2001				gr/qc/2002			
Total per year	1265	909		Total per year	1298	876		Total per year	1324	785	
Average per month	105	76	71.86	Average per month	108	73	67.49	Average per month	110	65	59.29
hep-ex/2000				hep-ex/2001				hep-ex/2002			
Total per year	748	468		Total per year	830	456		Total per year	810	395	
Average per month	62	39	62.57	Average per month	69	38	54.94	Average per month	68	33	48.77
hep-lat/2000				hep-lat/2001				hep-lat/2002			
Total per year	520	408		Total per year	587	512		Total per year	595	221	
Average per month	43	34	78.46	Average per month	49	43	87.22	Average per month	50	18	37.14
hep-ph/2000				hep-ph/2001				hep-ph/2002			
Total per year	4136	2854		Total per year	4238	2980		Total per year	4257	2556	
Average per month	345	238	69.00	Average per month	353	248	70.32	Average per month	355	213	60.04
hep-th/2000				hep-th/2001				hep-th/2002			
Total per year	3157	2377		Total per year	3194	2446		Total per year	3345	2161	
Average per month	263	198	75.29	Average per month	266	204	76.58	Average per month	279	180	64.60
math/2000				math/2001				math/2002			
Total per year	2999	553		Total per year	3237	490		Total per year	4246	443	
Average per month	250	46	18.44	Average per month	270	41	15.14	Average per month	354	37	10.43
math-ph/2000				math-ph/2001				math-ph/2002			
Total per year	533	215		Total per year	502	169		Total per year	715	213	
Average per month	44	18	40.34	Average per month	42	14	33.67	Average per month	60	18	29.79
nlin/2000				nlin/2001				nlin/2002			
Total per year	683	247		Total per year	689	213		Total per year	747	173	
Average per month	57	21	36.16	Average per month	57	18	30.91	Average per month	62	14	23.16
nucl-ex/2000				nucl-ex/2001				nucl-ex/2002			
Total per year	186	139		Total per year	257	185		Total per year	307	182	
Average per month	16	12	74.73	Average per month	21	15	71.98	Average per month	26	15	59.28
nucl-th/2000				nucl-th/2001				nucl-th/2002			
Total per year	987	734		Total per year	954	710		Total per year	1059	647	
Average per month	82	61	74.37	Average per month	80	59	74.42	Average per month	88	54	61.10
physics/2000				physics/2001				physics/2002			
Total per year	1156	502		Total per year	1238	412		Total per year	1238	262	
Average per month	96	42	43.43	Average per month	103	34	33.28	Average per month	103	22	21.16
quant-ph/2000				quant-ph/2001				quant-ph/2002			
Total per year	1530	598		Total per year	1923	740		Total per year	2188	677	
Average per month	128	50	39.08	Average per month	160	62	38.48	Average per month	182	56	30.94

Notes: The 'average per month' figures for 'in print' are rounded.