How four design flaws weaken CoRR's potential value.

A Computing Research Repository: Why Not Solve the Problems First?

A.J. van Loon
R&D Text Consulting
P.O. Box 336
6860 AH Oosterbeek, The Netherlands
tom.van.loon@wxs.nl

Abstract

The Computing Research Repository (CoRR) described by Halpern is potentially a powerful tool for researchers in computing science. In its current form, however, shortcomings exist that restrict its value and that, in the long term, might strongly undermine its usefulness. Important aspects that have insufficiently been taken care of are (1) the quality—and consequently the reliability—of the material stored, (2) the still restricted submission of material, which implies that other sources have to be consulted by researchers as well, (3) the still unsound financial basis of the project, and (4) the confusion that may easily arise when a "preliminary" version is stored in the CoRR, while a different final version is published in a journal. It would certainly be worth while to address these problems—and to solve them—before the CoRR is promoted worldwide; a premature promotion might lead to disappointment of the users and thus to a premature death.

H.3.7 Digital Libraries*—online publishing, information retrieval, document management*

Keywords*—refereeing, versioning, superdatabase, publishing costs*

1. Introduction

The contribution by Halpern (2000) in the present issue, being a revised version of earlier publications (Halpern, 1998; Halpern and Lagoze, 1999), is in itself a definite proof that the computing research repository (CoRR) as described by him is not yet truly mature. However interesting and promising the project is, one must conclude that there are still so many pitfalls that it seems advisable to stop the current project temporarily, to rethink it over, and to have a restart after the main problems will have been solved satisfactorily, rather than going in with the current situation. With respect to Halpern's (2000) contribution, one must conclude that:

- It was apparently necessary to publish the (adapted) material three times for the (largely) identical readership, so that one must conclude that at least the two earlier versions were, according to the author, not entirely satisfactory, and thus of a suboptimum quality;
- It was apparenly not sufficient to submit the paper to the CoRR for dissemination among all potentially interested readers, so that one must conclude that the author's confidence in CoRR's present role is not great;
- It is possible to easily find out the consecutive order of the three successive versions on the basis of different years of publication, but this would have been much more difficult—if possible at all—if the three versions had been published in the same year. In such a case confusion could arise easily as to the question which is the "definite" version.

ACM Journal of Computer Documentation, 2000,24:64-71
 ISSN 1527-6805

These critical remarks are not made because I do not recognize the importence and the potential value of CoRR, but because CoRR must get a true chance to fulfill its envisaged role, not hampered by shortcomings that can be avoided relatively easily. My remarks are therefore just meant to show some of the shortcomings.

…CoRR includes some pitfalls that should be removed before operation becomes truly feasible

This makes it even more important to consider the possibilities that CoRR has to offer, and to analyse in what way CoRR is new and can add value to existing possibilities. Where CoRR appears to add value, it should be analysed whether the practical operational functions have been designed optimally, or whether improvements are possible. And—last but not least—it should be analysed whether CoRR includes some pitfalls that should be removed before operation becomes truly feasible.

Such analyses must be carried out keeping in mind what CoRR is meant for. According to Halpern, the main functions are:

- To disseminate research results with a minimum loss of time.
- To have a database where relevant material can be found, and from which it can retrieved.
- To have it function without cost for the user.

Such objectives are truly important and worth striving for, but one might wonder whether they are realistic, particularly on the long term when the original "push" by people directly involved will have come to an end.

2. Avoiding loss of time

Halpern (1998, 2000) states, as an opening statement, that "[c]omputing research relies heavily on the rapid dissemination of results." This statement is somewhat enigmatic, because research is not based on the dissemination of results, but rather on the availability of previous results, thus on their retrievability. It is beyond doubt, however, that quick dissemination of research results contributes significantly to the progress of science, and delay in publication time (commonly 7-11 months for printed material) (Zhou, 1995) increases research costs, among other reasons by duplication of research. More rapid publication is, obviously, favorable not only for research in computer science, but also for science in general (and particularly the "hard" sciences). It is therefore interesting to see what other sciences did do to promote quick dissemination.

Quick dissemination is not a goal in itself, nor is it a need for the progress of science, as shown clearly by history. The need for quick publication comes mainly from the researchers' side, who want to have their material published as quickly as possible, largely because of the economic aspects involved (patents, priorities, fund raising, etc.). The researchers' requests for quick publication are the main reason why journals started competing. This led, among other things, to electronic publishing. It is good to remind here that this form of publishing is, as a rule, not yet (?) economically profitable for the publisher, nor does it meet all the demands of the users. This aspect seems out of scope here, however.

2.1 Limited time saving

Electronic publishing saves the time of "classical" reproduction by printing. These savings are currently minimum, since texts (and figures in increasing amounts, too) are submitted nowadays commonly in electronic form. "Typesetting," preparing a layout, proof correction and printing are thus quick processes if a good infrastructure exists; it takes commonly less than a week. Another delay is the distribution of the journals by post. This takes in most countries much less than a week; international delivery can be speeded up through airmail or priority post and then takes less than a week, too. This implies that the time saving by electronic dissemination instead of by distributing printed journals is at most in the order of two weeks. This is, as

a rule, negligible in comparison to the time involved in the research and also in comparison to the time needed by the researcher to write his article.

Why then publish electronically? The answer is: cost saving on the long term (see also early views: Borgman, 1986; Mitterer et al., 1992; Hammond and Shipley, 1994). The commonly heard argument that so much more time is saved because the refereeing procedure can be avoided, is obviously not valid. Refereeing is certainly an important cause of publication delay, but has a most important underlying reason: quality control. This is exactly why the great majority of electronic journals still have a considerable time lag between the submission of a manuscript and its publication (commonly in revised form, after adaptation on the basis of the referees' comments). Considering the fact that media with a high-quality reputation will not be inclined to let their high-standard material "drown" in a pool of unrefereed (not necessarily low-quality) manuscripts, it is hardly plausible that the "open protocol will encourage other scholarly archives to join in this framework," as Halpern anticipates.

In fact, it seems that Halpern encourages the submission of immature material ("Authors have 24 hours to withdraw or revise a paper"). But why should a researcher submit an immature report? Or does Halpern expect that authors will, within one day, come to the conclusion that their results are incorrect, because they will be contacted by users? This could only be true if each new submission would immediately be traced, downloaded, read and commented upon by users of the repository. This does not sound realistic. If such immediate comments are necessary for control of quality, it would be much more effective to adhere to the classical refereeing procedure. But this is exactly what is not intended (The submitted materials "are checked (by moderators…) only for relatedness to the topic area, but not quality or novelty").

2.2 Related loss of quality/reliability

It is apparently Halpern's intention to submit material to CoRR and to have it available through CoRR without any delay on the basis of refereeing ("Submissions are not refereed"; "By eliminating the time consuming and expensive process of peer review..."). This option is acceptable for researchers—who are always short of time and funds—only if CoRR is not meant to be reliable, but only informative about what other people are doing. No true scientists would ever rely for future research on material that has not been critically reviewed by experts. Consequently, one must deduce that CoRR's main function, in Halpern's vision, is alerting computer scientists that specific people are working on specific topics. But this raises the question why reports should be stored in full at CoRR, and why it would not be sufficient to use CoRR as a center where one can submit his name (with data like fax and e-mail addresses) in combination with the fields of interest that one is working in. This would be cheaper, better surveyable, and easier to maintain (persons could, for instance, be removed if they have not updated their data in a year's time). The CoRR data could also include a reference to the researcher's institutional or private web site, where the contributions that a researcher thinks worthwhile can be found (and possibly downloaded).

2.3 A possible solution: immediate refereed publication

There is a solution for the above apparent controversial choice between "rapid publication" and quality control. It is the development of a new infrastructure for refereeing. It is very well possible to publish an article, including refereeing and editing (thus meeting the same high-standards that are adhered to by present-day scholarly top journals), within 24 hours after submission (obviously only if the quality is good: revisions that the author should make will cost extra time, but it would be unrealistic to blame a journal for insufficient quality of a submitted article). One should also realise that editing becomes less time consuming if the manuscript is submitted exactly according to specific technical requirements (cf. Barclay and Murray, 1995).

The possible new infrastructure that could

achieve this—but that would require considerable financial investments before it would save much more money—has been described in the past few years (Dong and Van Loon, 1997a,b, 1998b). It seems therefore out of scope here to detail the possible setup of such an infrastructure. It might be useful, however, if CoRR is intended to play the role that Halpern envisages, to discuss within ACM, preferably in combination with organizations such as IEEE-PCS, whether CoRR could be designed and developed in such a way that its present disadvantage of insufficient quality guarantee could be replaced by the huge adventage of almost immediate publication of guaranteed high-quality material.

3. The need for an extensive database

Why are researchers interested in a central place where information about research results is available? The answer is clear: research results are published in a wide variety of media: in printed form alone there are hundreds of thousands of journals nowadays (Page et al., 1987)—probably some 800,000 (Dong and Van Loon, 1997a); many of them are devoted, in full or in part, to technology, including computer research. Because of this massive character of the flow of information, it is no longer possible to have direct access (through libraries or otherwise) to the large majority of them. This is the main reason why secondary (and tertiary) media developed. Electronic databases, with their much larger capacity, have taken over the role of printed secondary journals, but they still have several disadvantages:

- Databases are, as a rule, restricted to one discipline or a few disciplines, which becomes increasingly problematic considering the ongoing integration of research disciplines, particularly in innovative fields.
- No database in a specific field covers all published articles in the field (let alone unpublished internal reports), so that always several databases must be consulted (with a lot of duplicate results), without the guarantee that all publications that should be covered are covered (indeed, it seems fairly sure that important publications will be lacking).
- Databases provide increasingly more references to publications, but there is commonly no or hardly any indication about the quality of the works referred to, so that the researcher has to browse to increasing numbers of articles.
- Databases can be run economically only if they serve a large public, which implies, however, that they are rarely providing the data that specific researchers need for specific purposes.

3.1 Rejected options and consequences

These shortcomings have apparently not all been recognized explicitly by Halpern, but it is obvious that he was not satisfied with the existing ones: there would have been no need to establish a new one (CoRR), if other satisfactory options were available. It is interesting in this context that Halpern states that the option of joining LANL (the Los Alamos National Laboratory) was rejected because "it did not provide an interface to which other repositories could link". One must assume that Halpern has not only established this fact, but has also tried to convince the Los Alamos people to provide a suitable interface. This attempt was apparently in vain. It seems therefore highly improbable that all LANL material will be submitted to CoRR, which implies that at least one most important source will not be accessible to computer researchers through CoRR. This makes, in fact, CoRR one of the numerous databases to be consulted by researchers, which takes away one of the most important advantages that it might have offered.

A similar problem seems to apply to the option of becoming a node in NCSTRL (the Networked Computer Science Technical Reference Library), which—as Halpern states—"did not have all the software necessary for running a repository." Wouldn't it have been easier, more cost-effective and more practical for the computer-science community if such software

had been provided to NCSTRL rather than founding its own repository with the required software (but also with personnel, housing and other costs)?

3.2 A possible solution: a superdatabase

The problem of relevant data being dispersed over numerous databases is well recognized, but not solved. This is due to the fact that most databases have required huge investments, and that no owner/exploiter would like to join other databases, unless this would guarantee larger profits than going on itself. Since such guarantees are, as a rule, impossible, the current situation seems to be frozen.

One can only change the situation by adopting a new philosophy: the old data should, at least for the time being, be left in the databases/repositories where they are now, but new material (with their relevant data) should be stored in one single repository. What could be a more suitable development than storing in such a new database/repository all new manuscripts that are delivered in a prescribed digital form? The CoRR only partly fulfills the requirements, as it is restricted to computer science. This is the more unfortunate, because interdisciplinary and multidisciplinary work becomes increasingly essential for scientific progress; in addition, researchers become more and more aware that they often need research results from fields outside their own, narrow discipline.

For this reason it was proposed before (Dong and Van Loon, 1997a, 1998a) to start some kind of superdatabase, where all relevant data regarding new articles published in electronic journals (but also manuscripts submitted in electronic form to printed journals) are stored. When these data are retrieved, the researcher also finds indicated where, how and at what cost the original material can be retrieved. This superdatabase thus could meet at least one of the objectives that CoRR will not be able to reach in its present form or in a form that could logically develop from its present status.

It is clear from the data provided by Halpern that CoRR is insufficiently attractive now, especially for this reason. How else could one explain that "interest [in co-sponsoring] comes from both the United States and Europe" (a very vague statement) and not from countries such as Japan and India, where computer science is at a high level and where the researchers are very keen to cooperate with western colleagues? The relatively small interest is understandable if one keeps in mind the data regarding the LANL use: even this well established and fairly well known repository has a relatively low number of submissions ("It has now over 75,000 eprints, is growing at a rate of about 25,000/year..."), which indicates that it is more interesting as one of the possible sources of information ("...handles over 70,000 transactions/day, and has over 35,000 users") than as a medium through which authors/researchers want (or expect?) to disseminate their results.

4. The cost of CoRR

4.1 The Problem

A free service is difficult to realise. "Currently, CoRR is riding on the coattails of funding provided to LANL and NCSTRL, and this should suffice for the foreseeable future. The long-run funding situation is not yet clear." Many non-profit projects start this way, and experience shows that a good performance is commonly achieved; only, however, for the time that the initiators are still actively involved. They do a lot of work, often without being paid, while out-of-pocket costs are paid by an organisation. Such payments are, however, doomed to stop when a new management sets different priorities or when a budget cut is required.

It may be true that CoRR can keep on riding the coattails of LANL for some time, and its position might certainly become stronger during this time, but it is unlikely that such sponsoring will continue forever. LANL might be forced, as soon as a dip in the economy takes place, to get rid of such projects. It thus seems that the present-day financial basis for CoRR is insufficiently sound. That is no good start for such a project, because one failure for financial reasons might imply a regrettable drawback that could easily reduce the chances for a financially more sound project for quite a long time.

4.2 A possible solution: downloading at cost price

It is not entirely clear what Halpern means with the "70,000 transactions/day, and ...over 35,000 users" of LANL. What percentage of these numbers might characterise CoRR? If it would, in the longer term, be in the same order, a fairly sound financial basis might be obtained by making customers pay a small fee; not for browsing through titles or abstracts (I think that such free browsing is essential for an effective use) but for downloading. If, for instance, 70,000 articles would be downloaded per day, while the cost of maintaining the repository would be in the order of $7,000, one could charge the customers for $0.10 for each document retrieved. Even if the cost per day would be in the order of $70,000 (but this would imply yearly costs of over $25 million!), a charge of only $1.00 would be necessary to cover the cost.

Such charges are peanuts if compared to the prices paid nowadays for subscriptions to databases or scientific journals. One must therefore conclude that such a fee could not be an obstacle for a researcher, particularly if compared with his research costs. A free delivery would therefore not be truly helpful in the quick dissemination of knowledge. On the contrary, it would pose a serious threat considering the probability of financial failure at some time.

5. What is the latest version?

As stated in the introduction, Halpern (2000) is a revised version of Halpern (1998) and of a largely similar publication by Halpern and Lagoze (1999). This is no problem, because it is clearly indicated. But what if such a reference to earlier work is not made by the author? One should realize that many journals (in fact, almost all top journals) refuse manuscripts that have been published in the same or a largely comparable form before. Authors who want to have their material published in print, will thus be seduced not to refer to the earlier version deposited in CoRR.

CoRR may not think republication a problem ("The committee decided not to require any transfer of copyright or publication rights"), but it is not very likely that most authors will update their version in CoRR if they have published their work elsewhere (in print or in electronic form) and have signed a statement that the material was not published before. If they pretend to have forgotten the submission to CoRR, they may escape penalties. But if a publisher finds out that an earlier version was published (in spite of a signed statement that there is no earlier publication) and that it was updated afterwards, there is a great chance that the author will be prosecuted and, if so, he will certainly not be able to escape. This makes it unlikely that all CoRR material (or even a majority) will represent a latest version. In addition, authors may really forget to update their CoRR material, they may die before being able to adapt or remove the CoRR version, etc.

The consequence is that CoRR will include a possibly large percentage of material that is not up-to-date, i.e. newer versions may have been published elsewhere, and the user who wants to retrieve material from CoRR is not informed about this. This may result in references to versions that are outdated, but also in research projects carried out on the basis of previous work that has been found incorrect, etc.

Even the possibility of having various different versions of the same material is scientifically alarming. This aspect should not be underestimated, and CoRR should find a solution that is much better than a general disclaimer (which is not of any good for a user, and which may even help undermine the status of CoRR).

6. More is to be said

The above remarks concern only some general aspects, without going into much detail. Obviously, the provoking article by Halpern deserves comments in much more detail, particularly because the CoRR initiative might change procedures in literature search

considerably. It seems out of the scope of the present contribution, however, to deal here with all minor aspects (other comments might deal with a great deal of them). One should therefore not come to the conclusion that my criticism is restricted to the aspects of reliability/quality, overview, cost, and up-to-date character. These four main topics deserve much more consideration, however, before a large-scale implementation of CoRR should start.

My remarks are for a large part based on experience with technical and scientific journals, and with the articles published in them. These journals are all published in printed form, some of them with an electronic version. This makes it possible to find out which characteristics printed and electronic material have (or should have) in common, and which are (or should be) different. One of the most important characteristics of both is the absolute need of reliability. This should also hold for CoRR. In the presentation by Halpern, however, CoRR is meant apparently for an unstructured mixture of both high-quality manuscripts that may also be published somewhere else, and unrefereed reports of unknown quality and reliability. It is this aspect that worries me most of all: who would really be interested in a huge reservoir of data if one cannot rely on them? It is well known that the growth of the number of informative technical documents on the web poses more and more problems and requires measures (Rowe, 1997), including the way of presentation (Hysell, 1998). This will also be true for CoRR. Modern society—and the scientific community in particular—requires increasing quality (cf. Alexander, 1999). Presenting material does no longer suffice: it is just the electronic evolution that has changed technical communication into the direction of knowledge management (Leonard, 1998). It seems to me that CoRR is not based on this requirement, and one wonders why the various problems connected with this aspect are not dealt with first.

7. Conclusions

The CoRR project can be a valuable tool in computer-science research, but its potential applicability should not be limited beforehand, because pitfalls have not been noticed or because possible problems are neglected. CoRR deserves an optimum start.

It seems worthwhile to discuss the many aspects of CoRR that seem not yet optimum during a series of conferences or workshops, where experts should be invited to unravel all the potential shortcomings, and to discuss possible solutions, such as I have tried to do shortly in the above remarks about a few aspects. Only on the basis of such discussions a form might be found that avoids a premature, unnecessary death of a promising development.

References

Alexander, S.M. (1999). Communicator skills in a changing world. In T.J. Malkinson (Ed.), Communicating jazz: improvising the new international communication culture, *Proceedings of the 1999 IEEE Professional Communication Conference* (pp. 79-83). New Orleans: IEEE.

Barclay, R.O. and Murray, P.C. (1995). Virtual blood, real sweat, no tears: lessons learned from making a publication about electronic publications. In *Proceedings of the International Professional Communication Conference IPCC'95* (pp. 106-109). Savannah, GA: IEEE.

Borgman, C.L. (1986). The user's mental model of an information retrieval system: an experiment on an online catalog. *International Journal of Man-Machine Studies, 24, 46-64.*

*Dong, G. and Van Loon, A.*J. (1997a). Immediate publication requires a new infrastructure. *Science International,* 64, 15.

Dong, G. and Van Loon, A.J. (1997b). Publishing refereed and edited scientific and technical articles within 24 hours. *European Science Editing,* 23, 73-77.

Dong, G. and Van Loon, A.J. (1998a). A technological and infrastructural challenge: a superdatabase for accessing all electronically published science and technology articles at once. In T.J. Malkinson (Ed.): A contemporary reconnaissance: changing the way we communicate.

Proceedings of the 1998 International Professional Communication Conference (pp. 131-137). Quebec, Canada: IEEE

Dong, G. and Van Loon, A.J. (1998b). Publishing high-quality articles within one day: a new challenge in communicating science and technology. In: T.J. Malkinson (Ed.): A contemporary reconnaissance: changing the way we communicate. *Proceedings 1998 International Professional Communication Conference* (pp. 329-338). Quebec, Canada: IEEE.

Halpern, J.Y. (1998). A computing research repository. *D-Lib Magazine*, November 1998 (http: //www.dlib.org/dlib/november98/11halpern.html).

Halpern, J.Y. and Lagoze, C. (1999). The computing research repository: promoting the rapid dissemination of computer science research. In *Proceedings of ACM Digital Libraries '99* (pp. 3-11). New York: Association for Computing Machinery.

Halpern, J.Y. (2000). CoRR: A computing research repository. *ACM Journal of Computer Documentation,* 24, 41-47.

Hammond, J. and Shipley, R. (1994). Considerations for the use of multimedia with scientific and technical information: a case study. In *Proceedings of the International Professional Communication Conference IPCC'94* (pp. 304-310). Banff, Canada: IEEE.

Hysell, D. (1998). Meeting the needs (and preferences) of a diverse World Wide Web audience. *Proceedings of the 16th Annual International Conference on Computer Documentation SIGDOC 98* (pp. 164-172). Quebec, Canada: Association for Computing Machinery.

Leonard, D.C. (1998). Electronic evolution: from technical communication to knowledge management. In: T.J. Malkinson (Ed.) A contemporary renaissance: changing the way we communicate. *Proceedings of the 1998 International Professional Communication Conference* (pp. 9-19). Quebec, Canada: IEEE.

Mitterer, J., Lungu, D., Carey, T. and Nonnecke, B. (1992). A "reader-centered" approach to online information. In P. Trummel (Ed.): *Proceedings of the International Professional Communication Conference IPCC'92* (pp. 312-316). Santa Fe, N.M.: IEEE.

Page, C., Campbell, R. and Meadows, J. (1987). *Journal Publishing*. London: Butterworth.

Rowe, J. (1997). Hypertext to hypermedia and beyond—the evolution continues. In *Crossroads in Communication—Conference Proceedings 15th Annual International Conference on Computer Documentation SIGDOC 97* (pp. 237-241). Salt Lake City, UT: Association for Computing Machinery.

Zhou, F.C. (1995). The publishing period of science periodicals should be shortened as soon as possible. *Press and Publishing Journal*, 20 March 1995, page 5.