Electronic Publishing in Science

J.C. Sens reports on the joint Council of Scientific Unions (ICSU) – UNESCO conference of experts entitled *Electronic Publishing in Science* that took place in Paris on 19-23 February.

Electronic publishing in science started many years ago as an activity from the bottom-up. It has now reached the esoteric realm of high councils in international organizations, from where the field can be viewed in all its splendour, without hindrance from mere soft- and hard-ware problems. One such exercise took place recently at UNESCO in Paris where UNESCO and the press section of the Council of Scientific Unions (ICSU) jointly organized a conference of experts.

It became immediately apparent during introductory presentations on the state-of-the-art that electronic publishing is having a profound impact not only on communications among scientists, but also on education, on access by developing countries to information, and on the roles of learned societies, commercial publishers and indexing services, librarians, databank creators, archivists, copyright lawyers, and network operators.

The prolific growth of the information industry and its destabilizing side effects are characterized by many striking statistics that were quoted during the conference (see insert). With these statistics setting the scale of the problems, presentations focussed on several specific issues such as the digital library, pricing policies for journals, electronic archiving, copyright, standards, and network access in developing countries.

Traditionally, scholarly journals have been produced by learned societies (mainly the case in the US) and by publishing companies (mainly the case in Europe). In spite of much publisher-bashing by the scientific community, it seems unlikely that this pattern will change very soon.

Digital Library Initiative

Two years ago, the National Science Foundation, NASA and the Advanced Research Project Agency in the US launched the Digital Library Initiative involving a four-year, 24 M$US grant to six universities to develop prototypes of systems for the electronic distribution and retrieval of scientific journal literature. The University of Illinois' contribution was described in some detail at the conference. Its purpose is to build a large-scale, multi-journal/multi-publisher production testbed for scientific journal articles. This effort, along with the experience gathered at the five other universities, is likely to lead to a pilot of a digital library of the future two to three years from now.

In the Illinois project, collaboration agreements have been reached with several scientific societies and institutes, among them the American Physical Society (APS) and the American Institute of Physics (AIP). They provide the project with electronic copies of articles which will be published in their journals. Each article contains the complete text, graphics, images, tables, and equations in the Standardised Generalised Markup Language (SGML) format, with publisher-defined document-type-definitions (DTD) to delineate the document structure (title, authors, *etc.*). Commercial publishers are also supplying articles in this format. The project will grow to cover all articles in about 30 science and engineering journals.

The project intends to provide access to World-Wide Web (WWW)-based document repositories and links to full texts, digital images and video, and other information resources. At present, the focus is on the submission of publisher-provided articles, where the emphasis is on the conversion from printing-and-mailing to Internet-based retrieval and distribution. However - and this is a crucial point - the model will apply to any suitably organized repository of manuscripts, provided they are supplied in standard format.

Preliminary findings at Illinois indicate that SGML is sufficiently fine-grained to permit the sophisticated indexing and retrieval needed in a full-text environment, but has shortcomings when used as a display tool for mathematical formulae. The project uses SoftQuad's *Panorama* SGML viewer, implemented over Internet using Netscape's *Navigator* browser. This makes it possible to view documents incorporating WWW link protocols.

In the final version, documents will not be delivered by publishers, but will be searched for by users and downloaded. Open Text Corporation's *Open Text Index* search engine will be used for searching.

Problems under investigation include conversion from SGML to the Hyper-Text Markup Language (HTML) DTD used by WWW, difficulties in dynamic updating due to bandwidth limitations, and "tag normalisation", *i.e.*, the conversion of tags as defined in various DTDs to a normalised form. To overcome this last problem, the project recommends the use of the International Standards Organization's ISO 12083 physics and maths DTD as the standard for all publishers.

Non-Technical Issues Looming

Looming in the background of such digital libraries are copyright issues and pricing policies. No less than four plenary presentations were devoted to legal questions, *e.g.*, the intellectual property rights of electronic manuscripts residing in commercial databases; the futility of a reward structure for creators based on the right to

Some Statistics

Computer **performance** has increased a million-fold over the last twenty years. Within the European Union, over 25 % of offices and 15 % of homes are equipped with computers.

One-half of the world's population has **never placed** a telephone call. Some 80 % of the world's population, and 24 % of the world's scientists, live in developing countries. Developing countries account for 2 % of the production and use of the indexed output of scientific information; the quality of research and the economics of scientific publishing are the main factors responsible for this small amount.

The number of **Internet users** is expected to reach 200 million by the year 2000. At the end 1995 there were 70 000 WWW sites worldwide, with some 800 new ones being added on average every day; the principal scientific fields on WWW are engineering (880 sites), computer science (727), medicine (612), biology (509), earth sciences (473), and physics (469).

In a survey of 4000 scientists, less than 20 % make frequent use of **current awareness** services.

Scholarly **literature** doubles in volume every 10 – 15 years. Worldwide, there are some 90 000 periodicals and 1500 electronic journals; of the academic journals, some 50 are peer reviewed.

Between 1988 and 1992, five mid-west US research libraries cancelled **subscriptions** to 13 021 titles (some 5.7 % of the total). Yet these libraries ended up paying 30.5 % more for the remaining subscriptions. Average subscription prices have risen by about 10 % *p.a.* over the last four years, and continue to increase. Subscription prices have doubled since 1985.

The electronic publishing **market** in Western Europe lags some 3 – 5 years behind the market in the US.

J.C. Sens, who works at CERN, chairs the EPS Publications Committee's Task Force on Electronic Publishing and is a member of the IUPAP Working Group on Communications in Physics.

copy in a digital, zero-cost reproduction environment; the notion of transmission to and display on home computers, for instance, as a violation of copyright.

In spite of being the creators of intellectual property, scientists are largely unfamiliar with, and on the whole totally allergic to, the complex world of legal terminology, most of which is carried over from the music, motion picture and television industries. Indeed, it is with amazement that one hears of notions such as encryption (requiring a software key to read material), tattooing or watermarking (creating a permanent, indelible mark in a digital record) and fingerprinting (like watermarking, but adding a mark to identify the user at the time of use) as ways to combat piracy in electronic publishing.

The basic motivation for much of this wrestling with copyright stems clearly from the fact that commercial publishers have been well served by the now largely defunct paper-only journals. Manuscripts provided free-of-charge or paid for by scientists, free-of-charge value addition by unpaid editors and referees, copyright sign-over by authors who assign exclusive reproduction and distribution rights to publishers, sale-back to the providers, and subscription fee income from libraries in advance of incurring print and distribution costs cannot fail to leave a feeling of nostalgia among the beneficiaries of this 300-year-old system at the threshold of the electronic era.

Copyright legislation is national; electronic publishing is international. Constructing multimedia products - by adding graphics, sound and video to text, from sources which may themselves be compilations - makes adherence to copyright virtually impossible. Digital libraries are competing with database vendors in providing access, and are thus becoming primary suppliers of information ..... without the copyright, *etc.* The list of problems is long and few solutions are in sight.

In electronic publishing it usually is an institute or university which owns the copyright of papers produced by its scientists. So the transfer of copyright by authors to publishers is in fact illegal. Without a legal transfer, the publisher is neither an author nor a producer, and is thus unprotected. Yet publishers have made major investments, in both the paper and electronic sectors. This paradox has led to a draft *International Convention for the Protection of Publishers of Electronic Editions* by the International Publishers Copyright Council, and a draft *Directive for the Protection of Databases* by the European Commission. These initiatives aim to create a new right, and to allow database owners to exercise copyright, *i.e.*, prevent unauthorized, *i.e.*, free-of-charge, extraction and/or use, of all material in a database, including papers by authors who have not signed away copyright and favour free circulation.

The Scientists' Approach

Legal problems take on an entirely different dimension when seen from the "Ginsparg perspective" (P. Ginsparg, a high-energy physicist, pioneered the Los Alamos National Laboratory e-print server in the US which now covers nine domains of physics and has mirrors in Italy and Germany). This perspective finds conventions such as the ones mentioned as being merely pacts with the Devil, as devious as they are ineffective, since on the information highway of the future, scientists will take matters into their own hands and simply ban commercial science publishers.

Some two and a half hours were set aside at the conference for scientists to present their views. This was ample time, however, to make clear to an audience consisting largely of representatives of commercial interests of one sort or another, that electronic publishing replaces paper journals not by electronic journals but by publications residing on a scientific organization's, rather than a publisher's database. Issues of cost, peer review, archiving, and copyright must be re-examined from this vantage point.

It is tempting to quote from Ginsparg's (written) contribution to the conference: "The essential point is now self-evident: if we, the researchers, are not writing with the expectation of making money directly from our efforts, then there is no earthly reason why anyone else should make money in the process (except for a fair return on any non-trivial "value-added" ..... or except if, as in the paper-only era, the true cost ..... is sufficiently high)". Another nail is then driven into the coffin of commercial interests by the observation that journals in several fields of physics fail to fulfil their dual role of reporting research results (owing to today's preprint activity) and of serving as a vehicle for career advancement and grant allocation (since a name on a paper, as opposed to say "recent activities", plays a minor role, in decisions on jobs and grants). "Certain physics journals", it is then concluded, "play no role whatsoever for physicists. Their primary role seems to be to provide a revenue stream to publishers, a revenue stream invisibly siphoned from overhead on research contracts through library systems." Translate "overhead on research contracts" by "government funding of universities and institutes", and we have the situation in Europe. A revenue stream, one might add, that some publishers can probably do without, since physics represents only a minor fraction of their income.

This may seem just rebellious rhetoric, were it not for the fact that in subsequent presentations it was shown how two, albeit relatively small, learned societies, the International Union for Crystallography (IUCr) and the American Astronomical Society (AAS), have been able to set up fully electronic journals, with all manuscripts accepted, edited, refereed, archived, and distributed electronically (insert). These two example, represent admirable illustrations of what a small team of experts, librarians and the staff of a learned society can accomplish. They contain some valuable lessons (insert) for societies, which have yet to go electronic.

Regional Publishing Offices

The APS, in a short contribution, made the all-important announcement that it is "re-engineering" its entire publishing process, by constructing a "seamless electronic web from preparation by the author of a manuscript, through sub-

Electronic Publishing in Science

Some 156 people (25 % UK, 14 % US, 40 % continental Europe, 3 % South America, 18 % others) attended the well-organized ICSU conference *Electronic Publishing in Physics*. Summaries by invited speakers with expertise in specified domains were followed by working group sessions. This format proved to be effective and well-suited to the main goal of the conference, namely the formulation of recommendations to UNESCO's Director-General for further action. All presentations were well-prepared and highly informative. Of the 21 speakers, 12 were from the UK and the US. This is not surprising, since these two countries are the most active in electronic publishing. Among the scientific disciplines, physics was relatively well-represented, and life sciences poorly represented. Representatives of commercial interests outnumbered those from learned societies. Bandwidth, economics and the extension of networking were not on the agenda.

mission, refereeing, editing, production, and distribution". Furthermore, the society will install an e-print server that "will make available unrefereed and unedited articles in all fields of physics. This server will be linked to potential submissions to and refereeing for the APS journals, although authors will be free to submit their work to any other journal", or leave it unpublished. With these initiatives, the world's largest physical society is now addressing the heart of the problem. It is moving one step beyond copying paper versions to screens, leaving cost-increasing current-awareness services (*e.g.*, indexes) to others, and halting increases in the income needed to maintain both paper and electronic versions.

It is only through such measures that dialogue, on the time scale of an office conversation, between scientists and between authors, editors and referees can become a reality, that paper versions become manageable in size and cost, that publication delays can be reduced, and that libraries will be able to upgrade their services, reduce cataloguing work and devote more of their budgets to books.

In another short presentation, the International Union of Pure and Applied Physics (IUPAP) Working Group on Communications in Physics proposed the creation of "Regional Publishing Offices (RPOs)" to be managed by regional physical societies. RPOs would act as clearing houses for manuscripts submitted by authors usually, but not exclusively, located in a given region. All manuscripts would be collected on regional databases (logically one, physically one or more per region, with all regions interconnected) and manuscripts to be published would be translated into SGML, and made available to the publisher of the author's choice. There are six regional societies around the world which could qualify for participation, so the project would provide a single standard for the world's physics literature.

The APS and IUPAP initiatives, although developed more-or-less independently, are in fact in complete accord with one another. Perhaps the greatest beneficiaries of such proposals will be developing countries. Instead of having to buy information from commercial publishers, and then having to appeal to charitable organizations to pay for it, there will be, after initial investments, equality between all regions in the handling of and access to physics publications worldwide. How well this works will mainly depend on the cost of access to wide-area computer networks.

The IUCr and AAS Approaches to Electronic Publishing

The IUCr project began with the definition of a standard ASCII file, the *Crystallographic Information File* (CIF), having simple syntax rules. At present, only manuscripts in CIF format are accepted; they are submitted to the journal's office, usually via e-mail, and processed automatically. The manuscripts are validated with dedicated software, turned into PostScript files and returned to the author, along with an error report, for checking. Editors then receive, from the journal office, the text in proof format, the report of the checking and Internet FTP (File Transfer Protocol) access to the CIF file, the last to enable the editor and referees to view the manuscript on their screens. Editors and referees communicate via e-mail. Once editing has been completed, the CIF file is updated and archived in the journal's database. As to delivery, the list of contents (title, authors and synopsis) of each journal issue is placed on WWW. URLs point to the CIF files and chemical diagrams, which may be downloaded and displayed by the reader free-of-charge. The full text is as yet not available on WWW, but will be so in the future, probably for a fee. Annual sets of issues will be available on CD-ROM. Absence of adequate WWW and FTP services in much of the world and the high cost and unpredictable performance of networks have compelled the journal to maintain hardcopy distribution for the time being.

The AAS publishes *Astrophysical Journal* and the *Astronomical Journal*, which together comprise one-half of the world's peer-reviewed literature in astronomy. An electronic version of the Letters section of *Astrophys. J.* has been available on WWW since July 1995. Authors submit manuscripts in a AAS version of $LaTeX$; editors receive the manuscripts electronically or by telefax, and forward the corrected versions to the publisher. At this point, the $LaTeX$ version is converted automatically into SGML with tags defined by *ISO 12083*, and then entered into a database. Translation is required since $LaTeX$, although containing logical markup, is not equipped to provide hypertext links and links to databases containing the material referenced in the manuscript. The next step is copy editing on the screen of the SGML version and the automatic insertion of pointers to the abstracts database. This requires the two databases to have compatible standards and software. After assembly of text and figures into the definitive SGML-formatted manuscript ready for archival, two versions are derived: an on-screen version, via automatic SGML-HTML conversion, and a version in Adobe's *Portable Document Format* (PDF) using PostScript to PDF conversion to enable subscribers to locally print the article using Adobe's free-of-charge *Acrobat* reader. The SGML version is archived and made available for deriving new tools and functions as and when they are developed. The electronic version is presently released one month before the paper version and since its launch, the AAS system has served an average of 40 000 files each week.

Some Lessons

• The IUCr project addresses a highly specialized, virtually closed community, so it contains a somewhat parochial undertone: authors must submit their articles in a format (CIF) which is special to crystallographers, a feature that precludes simple access to the databases of other fields of physics. Access to WWW and on-screen display are via *ad hoc* software, and will remain so for future developments as long as the input is in CIF. This may be adequate for a closed community, but cannot serve as a general model. The AAS project has a wider base: it requires a given version of $LaTeX$ as input, and converts to SGML, the universal standard for database content and hyperlinks.

• It is important to realise that a database must satisfy stringent requirements (searching; access to other databases; coding so that different versions of the same manuscript can be extracted).

• Mathematical formulae are represented as images in a standard file format (GIF), inserted one-by-one into the text. This is time consuming and unsatisfactory. The *ISO 12083* DTD offers an alternative, but at present it also has problems [see *EPSIG News* 1 (2) 1995, published by the Electronic Special Interest Group which maintains and promotes *ISO 12083*].

• HTML DTDs are generally unable to cope with the large number of special characters needed for scientific documents and "it seems that HTML 3.0 (the latest version of the HTML DTD presently being reviewed by a Internet Engineering Task Force will) not emerge as a standard" [P. Boyce, *Computers in Physics*, January 1996].