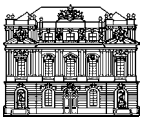
Michael Nentwich

cyberscience

Research in the Age of the Internet

Chapter 2

CYBERSCIENCE: THE NEW TOOLS – THE NEW WORKING ENVIRONMENT



Austrian Academy of Sciences Press
Vienna 2003

Submitted to the Austrian Academy of Sciences on 10 April 2003
by Gunther Tichy, member of the Academy

British Library Cataloguing in Publication data.
A Catalogue record of this book is available from the British Library.

All rights reserved
ISBN 3-7001-3188-7
Copyright © 2003 by
Austrian Academy of Sciences
Vienna

Austrian Academy of Sciences Press
Tel. +43-1-5129050-3405, Fax +43-1-51581-3400,
Postgasse 7, A-1010 Vienna
Email: verlag@oeaw.ac.at
http://hw.oeaw.ac.at/cyberscience

Layout, cover & type-setting: Manuela Kaitna, A-1080 Vienna
Printed and bound in Austria by Manz Crossmedia GmbH & Co KG, A-1051 Vienna

DETAILED LIST OF CONTENTS

2 CYBERSCIENCE: THE NEW TOOLS – THE NEW WORKING ENVIRONMENT

This chapter provides an overview of the technological elements of cyberscience. I shall present in a synoptic way and describe from a technical and functional point of view the technologies and applications involved. In this context, I shall also discuss some trends and influences on academia from other "continents of cyberworld" (such as business software or E-commerce). Passim, I shall also analyse the technical factors influencing academic communication patterns, as outlined in 1.2.3.1.

The purpose of this chapter is to serve as a reference point for all other chapters when it comes to the technical aspects (as they will be outlined only here).[143] For the purpose of the argument of this study, any intricate technical detail would be superfluous and misleading. Wherever possible, concrete examples from the CYBERLINKS (cf. 0.3.4.1) database are given.

This chapter has a first section presenting the basics, that is the computer and the network (2.1). The rest is structured according to the distinction I introduced in 1.2.1.[144] It distinguishes between machine-to-machine (2.2), people-to-machine (2.3) and people-to-people communication (2.4) in academia. Furthermore, there is a special section on the technical side of digital archiving (2.4.9), followed by a short outlook.

2.1 Basics

Before presenting the new tools for the three forms of scholarly communication, we need to discuss two basic elements of the new academic working environment, namely the computer and the Internet.

2.1.1 Computers: stand-alone and networked

The computer developed into a universal aid (Mittler 1996, 75): the workplace of the future may provide everything necessary for research and publishing. Mittler speaks of the "fight against the media discontinuity in academic work". Stichweh (1989) points at the concentration of resources at the workplace and at the fact that this is true for both the humanities and the sciences.

Speaking of the stand-alone device and depending on the research field, the computer serves as a powerful calculator, as a device for storage of data of whatever kind both numeric and other (such as text, graphics, pictures and sound) or as a text processor with layout capabilities enabling researchers to produce publications of very high layout qual-

[143] Note that this is not the place to give detailed technical descriptions, but rather short descriptions of the features and functionalities of the various elements presented.

[144] Note that this distinction serves mainly the purpose of structuring here but is not clear-cut as some technological elements may be partly listed under more than one heading (for an example, see in fn. 237).

ity ("camera-ready"). This is certainly not the place to describe in detail how a computer functions. It suffices to say that today, it is a universal machine: With the right software or a gifted programmer, a computer can do a wide range of things. There will only be very few researchers denying the power of this device.

Take as an example the writing of an academic publication. Up-to-date word-processors are able to present your text in a way in which you can view the whole structure of the text at once.[145] This allows you to browse through the text and re-arrange it in a matter of seconds. A contents or tables list is produced within seconds, too. Powerful add-ons help you to insert quotes from your literature database and will format the bibliography according to the style required by the publisher. You can give the digital version of the text to a colleague who can edit it and comment on the screen so that you see the amendments and corrections and can either accept or reject them with a click of the mouse. Data presentation tools enable the writer to insert dynamic charts and graphics still connected to the original data. This allows the researcher to correct data without having to remake the chart or graphic; and so on and so forth. The point is that it is extremely fast and comfortable to do this if compared to the time before the advent of the computer.

The true power of the computer is unfolded as soon as it is hooked up to a network, that is, if it is connected to other computers. With the networked computer, data-exchange of all sorts becomes easily feasible. Only if networked, is the computer able to concentrate most of the resources necessary for academic work, with the effect that it becomes a true "universal aid". The network connection allows accessing remote databases, downloading documents and publications, exchanging information and communicating with other researchers. It is this communicative potential which is the core of cyberscience.

2.1.2 Internet and WWW

Had this report been written several years ago, I would have had to include a distinction between the Internet and other academic networks (e.g. Werle/Lang 1997). In addition, I would have had to describe a number of different services running on these various networks, such as Gopher or WAIS (e.g. Meier/Wildberger 1993). Today, academia has almost entirely opted for the Internet which was originally one of a few co-existing networks.[146] Furthermore, for most scholars, it is above all E-mail and the World Wide Web (WWW) with its standard protocol[147] HTTP that runs as an application protocol above the network layer of the Internet.[148]

[145] In Microsoft WORD this function is called "Outline View".

[146] For an interesting overview of this early phase of the Net and the difficulties of even sending E-mails from one network to the other see Quarterman (1986).

[147] In relation to the Internet, a "protocol" is a set of rules that determines how two systems interact, in particular protocols set the terms how a client programme and a server programme interact in order to perform a particular task. It is the World Wide Web Consortium or W3C which organises and manages the further development of the WWW and its protocols, which secures interoperability between the various protocols and which sets standards (<Cyberlink=739>).

[148] One speaks of three main "layers" of the Internet: the physical layer (the cables), the network layer (everything to do with connecting), and the application layer (what the standard user actually is confronted with). In this overview, I do not go into the details of the physical nor the network layer, which is ruled by the protocol TCP/IP, as I am mainly interested in the application layer.

The WWW has, in contrast to its predecessors, a graphically oriented user interface, which allows exploring the information available in an intuitive way. All information is stored in files, the majority in the text-oriented HTML format which nevertheless allows for inclusion of graphics and other layout elements as well as video, audio and even little pieces of software (so-called "applets"). The pages are linked with so-called hyperlinks. Hyperlinks are specially marked words or graphical elements (such as "icons", that is small symbolic pictures). The computer mouse is the key device when the user "browses" from page to page: When "clicking" on a hyperlink, the underlying address will be reached and the object (document, picture etc.) will be transferred ("downloaded") – no matter on which server it resides. By this token, the whole WWW is presented to the user as one immense unit, interwoven with a myriad of hyperlinks. The basic elements of the WWW, the webpages (see below 2.4.3), reside on web servers in directories which are addressable via the HTTP protocol.

There are different types of Internet addresses. URLs (uniform resource locators)[149] are composed of different identifiers. First of all there is the IP-address:

- The *IP-address* is the physical Internet address; in the current version (IP version 4) it is a string of 32 ones or zeros (i.e. 32 bit), for better readability normally written as four octets separated by dots, like 111.222.33.444. Each PC connected to the Internet has either a permanent or a temporary number (the latter is only valid during the present session).
- The *domain name* is an "alias" for the 32-bit-IP-number, an alternative, very popular and user-friendly identifier which is widely used instead of memorising IP-numbers; each domain name is a unique element of the Internet's Domain Name System (DNS). The DNS is like an automatic information system. It is a hierarchical (tree structured), highly distributed and dynamic database, consisting of tens of thousands of name servers on the different levels of the system, with the Root Name Server and the Top Level[150] Domain Name Servers at the top being the most important computers of the Internet. The DNS has several functions, but is mainly used for resolving domain names into IP-addresses. Nevertheless, the DNS is not the WWW directory, and it cannot be "searched through" like a phone book.

WWW addresses are human-intelligible addresses, which start with the domain name, that is, with the server's name, followed by the directory hierarchy and the file name at the end.[151] The problem with these is that they may and do change frequently which leads to the well-known and annoying "404 errors" indicating that the address is no longer valid and the document is not found. In particular in an academic environment (but not only there), where exact and persistent referencing is important, initiatives are under way to arrive at a persistent addressing scheme. This comes under the label of "(persistent) identifiers" (for an overview see Lynch 1997):

[149] According to the W3C, URL is no longer used in technical specifications. It is (was) an informal term associated with popular identification schemes, like http, ftp, mailto, etc.

[150] "Top level" signifies the highest order, that is, in particular the country domains like ".de", ".at" as well as ".edu" (for the US higher education scene) or ".com" for (US and worldwide business); within each top level domain, there are second level domains, for instance in Austria, ".gv.at" for all government-related web sites, ".ac.at" for the academic network etc.; the third level is, within the country hierarchies, the level of individual institutions, like ".oeaw.ac.at" for the Austrian Academy of Sciences.

[151] For examples, see below in 2.4.3.

- *URNs (universal resource names)*[152]: This is an initiative taken by the Internet Engineering Task Force (IETF) to standardise the syntax of a name for persistent referencing of digital objects which will parallel the present system of URLs. Probably analogous to the domain name system, there will be many decentralised "resolution databases" (resolving persistent URNs to possibly changing URLs denoting physical locations of files) which will guide the browsers to the right place (urn:)[153]. URNs are identifiers with an institutional commitment of persistence and availability, like for instance the system of the so-called
- *PURL (persistent URL)* initiative[154]: The idea is that dedicated intermediate resolution services (so-called PURL servers) resolve requests for URLs with the up-to-date address of the resource (file). As long as a more sophisticated system is not yet put in place, the Online Computer Library Center (OCLC) initiated such a centralised database system to redirect browsers from a registered PURL to the actual URL.

It is expected that at some point, the new schemes will resolve the present problems of no longer valid URLs. Meanwhile, two initiatives from the publishing sector should serve the same purpose:

- *SICI (serial item and contribution identifier)*: Based on the international standard serial number (ISSN), SICI not only identifies serials, but also individual issues and even articles within a journal. So far, it is mainly used to streamline interlibrary loan and document delivery.
- *DOI (digital object identifier)*: The idea is to attribute to each digital publication unit an identifier that would not be altered wherever the digital object actually resides. A DOI server would resolve DOIs to up-to-date URLs if needed. DOIs seem to be compatible with the URN framework. Lynch (1997, 6) proposes to skip the misleading, too comprehensive label DOI for something like "Publishers Object Access Identifier" (for a description and discussion see Davidson/Douglas 1998).

Related to these ideas is *XLL ("Extensible Linking Language")*, a possible solution to the problem how information-rich inter-linked document structures can be constructed without the above mentioned problems. One option is to separate link storage from document storage: "XLL links are no longer bound to occur with the data or indeed in the same documents as the data that they refer to. Instead they can be produced and stored separately, and (potentially) managed independently." (Carr et al. 1998; see also Carr et al. 1995; Carr et al. 1996a, for a practical development of this ideas). In 2001, the W3C finally approved the XML Linking Language (XLink) Version 1.0 recommendation.[155]

Three further elements of the application layer of the Internet need to be addressed in this overview. The "telnet" protocol enables direct access to another computer over the net. The user needs an account (a "user-id" plus password) on the remote machine to open a "session". Depending on his/her privileges on this machine, s/he is able to communicate with the data and run programmes on it while sitting at another machine. Many programmes are only available on central and powerful servers accessible to decentral users. The next protocol is the file transfer protocol (FTP) which is used to transmit files from one computer to another (for instance from the desktop to the web server). Finally, the Simple Mail Transfer Protocol (SMTP) is the basis of E-mailing (see below 2.4.1).

[152] Both URLs and URNs are URIs (Uniform Resource Identifiers) which is the generic name of all names/addresses that are short strings that refer to resources.
[153] <Cyberlink=430>; <Cyberlink=397>.
[154] <Cyberlink=431>.
[155] <Cyberlink=843>.

2.1.2.1 The future of the Internet

Traffic on the Internet has grown considerably over the recent years due to both an increasing number of users and ever more non-text (such as audio, video and graphical files) being transmitted. The capacity is being expanded constantly, the current aim being powerful broadband networks, more or less[156] separated from the rest of the Internet. Earlier, this came under the labels of "Global Information Infrastructure" (GII) and of "Internet II"[157]. Today, there are a number of high-speed networks like the "very high performance Backbone Network Service" (vBNS+)[158] and Abilene[159] in the US and GÉANT[160] in Europe. These new networks should be able to transmit in real-time, that is, without (considerable) delays, live video and audio streams. Only then, will a number of the services discussed in the following sub-sections (for instance video-conferencing) have a chance to be practised on a broad level:

Table 2-1: Necessary bandwidth for selected Internet services

Service	Necessary bandwidth
Interactive educational software	Gbps
Scientific modelling	100s of Mbps – Gbps
Teleworking/distance education & training	6-10 Mbps
Desktop video-conferencing	6-10 Mbps
Publications	100s of Mbps

Gbps = Gigabytes per second (1,000,000,000 B/s)
Mbps = Megabytes per second (1,000,000 B/s)

Source: Axmann/Payr (1999, 41)

Table 2-1 shows the bandwidth needs for some of the cyberscience services. If we compare this with the actual bandwidths, that is the speed of the present network, a gap can be detected. On the website of the European high-speed network GÉANT[161], it is shown that most of Western Europe is connected with 3,5 Gbps, with a number of areas far below the one Gbit/s threshold. However, these figures are misleading, as it only relates to the so-called backbone network, that is the "expressway", but not to the finer grains of the academic networks. Many research institutions have connections with only 10 or 100 Mbps; tele-working researchers access their institutions mostly with slow mo-

[156] Actually, all scientific and academic networks worldwide are certainly connected with the rest of the Internet, that is the many, many commercial networks. Many "bridges" between these different networks exist. The contracts between the academic network providers and the other Internet service providers (ISP) stipulate the conditions of use of the academic networks for non-academic purposes. In general, through-traffic from commercial networks to other such networks through academic networks is to be avoided. The intra-academic traffic uses to a very high proportion only the academic networks because the academic routers (network nodes) normally use as "open and best routes" the intra-academic lines.

[157] <Cyberlink=856>.

[158] Started in 1995 as vBNS and is now provided by a private firm for the National Science Foundation (NSF); <Cyberlink=855>.

[159] <Cyberlink=857>.

[160] <Cyberlink=858>.

[161] <Cyberlink=858>.

dems (56 Kbps). Furthermore, this bandwidth of the access route has to be shared by all users at one institution. Hence, if you need a 2 Mbps connection for a simple desktop videoconference, but your colleagues next door are surfing the web, downloading data and running remote programmes on a server at the same time, the present bandwidth is not enough. The same account can be made with regard to the higher levels of the network. While it seems that the multi-Gbps networks in existence and under construction would satisfy most needs of academia, there are still bottlenecks. In addition, for many applications, such as high-quality videoconferencing or secure remote control of measurement devices, the establishment of so-called virtual private networks (VPN) seems necessary to improve the "quality of service". Contrary to the way the present Internet works,[162] VPNs rely on dedicated lines, i.e. exclusively reserved connections between two or more points within the network. Only with the worldwide introduction of the next-generation Internet Protocol IPv6, the so-called "resource reservation protocol" will provide the quality of service needed.

Not an extension of current Internet technology, but something qualitatively new could develop on the basis of the Grid projects. The DataGrid[163] technology is currently being developed for high performance distributed computing (see below 2.2.1), but may be opened, in the not so distant future, for both researchers from other disciplines, and perhaps the general public, too.

2.2 Machine-to-machine communication

While, in this study, I will be interested mainly in computer mediated communication involving researchers, there are, nevertheless, a number of interesting applications which involve mainly communication between machines in the academic network. Although initially started by people, the computers can fulfil a variety of tasks for the researchers without the latter being involved all the time: distributed computing, on the one hand, and what we may call "intelligent, semi-autonomous information retrieval", on the other hand, will be presented in the following.

2.2.1 Distributed computing

There are computing tasks which are so huge that it would take even the big so-called mainframe and super-computers very long to fulfil. Often these large computers are not available at all or not available long enough. Some of these tasks can be decomposed in smaller units, computed separately and then recomposed at the end to get the overall result. Distributed computing or data-processing is a way to carry out such large computing projects by, first, decomposing the overall task, second, distributing the units among a large number of decentral computers, and third, administering the re-composition of the partial results.

[162] No direct connection is established between two points (as with the telephone), but everything is split into many small packets which all use the same lines and which are queued according to the first-come-first-serve principle at each routing node in the network.

[163] <Cyberlink=248>.

Distributed computing can either be done on a relatively small scale with, for instance, the participants of a distributed research group and their computers. The other alternatives are mass projects. In the latter case, there is a central website distributing client programmes (small pieces of software which include the computational core and which handle the communication with the central server) and the task units. Often the client programmes run in the background of the client computer (without the user noticing because resource management is tuned in a way to give priority to the application in the foreground). Alternatively, they run as a screen-saver, i.e. only if the user does not use the input/output devices of the computer (screen/keyboard/mouse) at the moment. The client programme connects to the main server, gets a new working unit, disconnects, works on it, and when ready, connects again, sends back the results and retrieves a new working unit and so on and so forth.[164]

The same idea can also be used for massive distributed simulation: some simulations can only be implemented by distributing the task among a large number of processors. Furthermore, the co-ordinated steering of telescopes in the framework of targeted astronomical observation is a special case of distributed computing, too.

In the US, a national initiative is under way under the label HPCC ("High Performance Computing & Communication") aiming at developing the "next generation" soft and hardware for distributed computing and the like on a large scale. The next step in the evolution of the HPCC is called "Networked Scientific Computing" (NSC) and will allow the scientific community to use the high performance communication infrastructure (vBNS+, GÉANT etc., see above 2.1.2.1) with the aim to integrate the heterogeneous networked hardware and software resources as a single "meta-computer"[165]. Equally, at CERN, the future computing infrastructure will be built in an EU project with a view to providing intensive computation and analysis of shared large-scale databases, from hundreds of Terabytes to Petabytes, across widely distributed scientific communities.[166] Scientific disciplines interested in such computing power are to be found not only among high-energy physicists, but also in for Earth observation, genome exploration – that is where large-scale, data-intensive computing is essential.

2.2.2 Semi-autonomous information retrieval

Retrieval of information can either be an activity of researchers communicating with machines (databases, archives) or a semi-autonomous activity of software tools. While I will look into the former in 2.3.4 below, the latter will be presented here.

2.2.2.1 On the path to the Semantic Web: meta-data, XML et al.

The WWW is, for the most part, rather unstructured. The information units are, in general, not ordered. Searching the web is therefore mainly done on the basis of full-text search, that is words or combinations of words have to be found in the text of a web document. This is rather inefficient as a word can have different meanings in different con-

[164] Prominent examples of mass projects of distributed data-processing are to be found in astronomy (SETI: <Cyberlink=249>) and cancer research (United Devices: <Cyberlink=412>), see 2.2.1.

[165] Quoted from <Cyberlink=736>.

[166] DataGrid project (<Cyberlink=248>); on an international level, the various regional "grid" projects will be merged to the "World Wide Grid".

texts so that you may find totally unrelated documents. In other words, although everything on the Internet is machine-readable, the data is not machine-understandable. One way of overcoming this shortcoming, is to use a standardised (or controlled) meta-language to describe the content of a document. Meta-data is "data about data". Searching for a word in a particular section of the meta-description would "hit" only related documents. The W3C provides for the Resource Description Framework (RDF)[167], based on XML (see below), which lays the foundation for processing meta-data: it secures interoperability between applications that exchange machine-understandable information on the Web. Among others, the RDF solves the problem of relationships between different meta-data, e.g. between authors and addresses if there are more than one (Berkemeyer/Weiß 1999, 1096).

There are a number of initiatives, in particular within the academic and librarian communities, to use meta-languages for this purpose. One of them is Dublin Core (DC)[168], a standard for meta-data to describe bibliographic entries and publications. In the meta-section of such a DC-compatible web document, the author, the title, the publication date etc. are included in a uniform format. See also the meta-data initiative CARMEN[169], equally intended to make the academic knowledge retrievable in a more targeted form (Plümer 2000). Another example is the Open Archives Initiative (OAI)[170] which standardises the description of items included in electronic archives to enable searching in distributed archives. In the framework of OAI, the Academic Metadata Format (AMF)[171] is being developed.

The likely successor of the current language of the web, the *extensible mark-up language* XML[172], incorporates meta-language of all possible sorts in order to describe each document extensively. It is a very flexible instrument, which can be adapted to a variety of purposes and content. Meta-tags are pieces of text, invisible for the user[173], in a standardised form embedded in the document. It depends on the context what the content of the meta-tag includes. Consequently, there are many different variants of XML, such as MathML[174] with specific tags used by mathematicians. Hardie argues that XML will "move the Web from its present unstructured form to a semi-structured form (...) It is the machine to machine interactions that will take the Web to the next level of usefulness, primarily by providing the humans with better quality, relevant information." (1999, 11)[175]

The present overall aim of the W3C is it to create a so-called "*Semantic Web*" (cf. Berners-Lee 1998). It is an extension of the current Web in which information will be given "well-defined meaning, better enabling computers and people to work in cooperation. It is the idea of having data on the Web defined and linked in a way that it can be

[167] <Cyberlink=263>.
[168] <Cyberlink=252>.
[169] <Cyberlink=51>.
[170] <Cyberlink=60>.
[171] <Cyberlink=805>.
[172] <Cyberlink=372>.
[173] You may, however, have a look at them when reading the source text of a web page.
[174] <Cyberlink=543>.
[175] "XML has the potential to make nearly all Web stuff – publishing, translations, searching, identifying – nearly friction-free. (...) Once we get to nearly friction-free publishing, everything may change (again). When software agents roam the Internet initiating 'automatic auctions' for answers to questions, or when a printstore the size of a Kodak kiosk serves made-to-order books ordered from Amazon.com and BarnesandNoble.com in whatever size and format you specify, it's hard to know how small, medium, or large publishers (and authors) will fare." (Jensen 1998, 4)

used for more effective discovery, automation, integration, and reuse across various applications."[176] In this context, "semantic" means that the elements of the web are tagged in order to allow retrieving its meaning in an automatic way. A number of initiatives intend to provide universal ontologies for the web, that is joint terminologies between members of communities of interest. On a general level, the W3C defines WebONT[177]. On this basis, specific ontologies for all fields would be developed. For instance, WEBONTO[178] is intended to support the collaborative development of such ontologies.

2.2.2.2 Knowbots, infobots, software agents

"Knowbots" ("knowledge acquisition robots") or "agents" or "infobots" are pieces of software, which, on the basis of an individual (search) profile given to it by its "master" (user), search in the Internet for specific information. The better structured, the better organised the Internet becomes, the more efficient these knowbots will be. One way of achieving this is by meta-tagging the documents, i.e. by describing documents on a formal level (not directly visible to the human reader, but to the software; see above 2.2.2.1).

New generation agents may solve the perceived problem of information overload by filtering the information stream for the individual researcher (LaPorte et al. 1995; see also Harnad 1990, 3). Other types of agents will play a role in negotiating information access rights, that is they will take over the task of identifying oneself for all databases and resources visited. Others are being developed with a view to enable collaborative, so-called "multidisciplinary problem solving environments" for distributed computing in the framework of the next-generation Internet.[179]

2.2.2.3 Knowledge discovery: data mining, web mining, bibliomining

Data mining is the process of finding new and potentially useful "knowledge" ("patterns") from existing data that are recombined in so-called data warehouses. Web mining focuses on Web documents and Web databases. There is also a special variant called "bibliomining" which is the combination of data-mining, bibliometrics, statistics, and reporting tools used to extract patterns of behaviour-based artefacts from library systems.[180] Data mining, web mining and bibliomining are covered by the term knowledge discovery.[181]

2.3 People-to-machine communication

Academics do not only communicate with other academics (see below 2.4), but also with computers containing data and information needed for their daily work. Here, future screen technology plays an important role. Furthermore, I shall present the idea of interactive electronic reading devices, of ubiquitous computing and telework, of databases and digital libraries, of remote control, and print-on-demand.

[176] From the Semantic Web Activity Statement on <Cyberlink=443>.
[177] <Cyberlink=740>.
[178] <Cyberlink=738>.
[179] E.g. SciAgents <Cyberlink=736>.
[180] <Cyberlink=772>.
[181] See the Knowledge Discovery resources at KDNuggets <Cyberlink=773>.

2.3.1 Screen technology

The most important hardware interface to cyberspace are, so far, computer screens which present all types of data to us, the majority of which comes in visual formats, that is texts and graphics. On the path to cyberscience, academics sit in front of screens ever-longer periods of time. Screen technology is developing and improving fast. However, at the time of writing, most people would print interesting articles from E-journals and only very few would read whole papers on screen. While the latest generations of screens do not suffer any more from the old disadvantage of flickering (due to low framerates), the available and common screen technology is still far from satisfying the basic needs of constant screen reading:

- *Portability*: Screens need to be portable to everywhere just as sheets of papers are: they have to be light and should be independent from power supply for long periods (low electricity consumption);
- *Portrait format*: Some (for instance Fidler 1998b) argue that the traditional landscape format of computer screens is a hindrance on the path to on-screen reading. This might vanish in favour of the portrait format or future screens may give publishers and users the option to choose either orientation. "This means that electronic editions of publications can be built more successfully on the traditional portrait-oriented, page-based format of printed publications and typographic documents." On the question landscape vs. portrait see also Wearden (1998b);
- *Independence from surrounding light conditions*: It should be possible to look at screens from at any angle without deterioration of the visibility of the screen content, disturbing shadows or reflections.

The newer screens used in notebooks (LCD technology) come ever closer to these aims but have not yet reached them. So far, the various E-book devices[182] are still bulky and rather unconvincing alternatives to the printed book. There is, however, a technology on the horizon, which might be able to fully meet these criteria. It became known under the label of "E-paper" – thin sheets of plastic that can display digital content like screens. The prototypes of E-paper have characteristics very similar to ordinary paper when it comes to surrounding light conditions and consume much less electricity than today's standard screens.[183] Examples are SMARTPAPER[184], E·INK[185], CHOLESTERIC LIQUID CRYSTALS[186] (West 1998), and EPYRUS[187]. For a comparison of available display technologies see Doane (1998). It may well be that very shortly, there will be portable reading devices with a large memory chip. They might be no thicker than an issue of an academic journal, but could carry and display whole personal libraries. This would revolutionise the way we think about and use digital information.

[182] For the current state of affairs in this market, see <Cybercategory=36>; in particular EBOOKS.ORG gives an up-to-date overview of all devices (dedicated readers, handheld devices, tablets, webpads) with pictures and comparative data (<Cyberlink=25>).

[183] They only consume electricity when changing the content of the page, not – as today's screens – even when sustaining the image. As texts are the predominant form of presentation of academic content – at least so far – changes of screen content while reading should only happen rarely.

[184] <Cyberlink=446>.

[185] <Cyberlink=29>.

[186] <Cyberlink=326>.

[187] <Cyberlink=913>.

2.3.2 Interactive electronic reading devices

Just as academic or professional reading is more than just consuming characters printed on sheets of paper, but also involves annotating, highlighting etc., the future might bring about reading devices which support interactive reading in the electronic world. See e.g. Xerox' XLIBRIS[188] prototypes (Schilit et al. 1998b). One of the possible features of the new devices may be what Schilit et al. call the support for "fluid movement among different styles of reading and different document activities" (ibid., 3). The knowledge worker of the future might be able to use his/her *online* annotations to put together new queries, to form annotated clippings of the documents retrieved, to skim over the text etc. Furthermore, the Xerox prototype also automatically generates 'further reading' lists.

It is conceivable that the "reading device" of the future (perhaps on the basis of E-paper, see above 2.3.1) will be connected to the Internet. This may allow retrieving files from a remote document host to which the reader has access, annotating them and storing them again at the same location. As a consequence, the academic reader may have access not only to the documents stored in the personal (next generation E-paper-) notebook and to remote E-publications and databases, but as well to his/her own annotations – a situation very similar to the present state of affairs, although in digital format.[189] Such annotations could also be shared among a research group. Indeed, one of the areas of activity within the W3C initiative of a Semantic Web (see above 2.2.2.1) is the development of remote shared annotations to Web documents in the Annotea[190] project.

2.3.3 Telework and ubiquitous computing

The network enables scholars to work not only at their office desks, but also from home with their connected home PCs (telework). Ever more researchers do not store their data (documents, mails, notes) on the local drive of their PC or notebook, that is on a computer which, as a rule, is not always turned on and hooked up to the network, but on network drives. By this token, the data is available around the clock and, depending on the security policy of the respective computer department, from (almost) everywhere.[191] There are even data services offering virtual network drives over the Internet for everyone, even if not connected to any institutional network.[192] Depending on the sort of network, access to the data may be more or less convenient. In some cases, the user will not even notice the difference between a local and a network drive (for instance in a WINDOWSNT network)[193]; in other cases special programmes (FTP or Telnet clients) will have to be used. The Internet Messages Access Protocol (IMAP) enables ubiquitous access to one's mail archive (see below 2.4.1). Furthermore, mobile communication is increasing in bandwidth and hence mobile audio and video conferencing may be available soon at sufficient quality.

[188] <Cyberlink=28>.

[189] However with the important difference that access to the personal files (documents, annotations) is not restricted to one's office or library, but would be ubiquitous.

[190] <Cyberlink=741>.

[191] An additional advantage of network drive storing of data is the automated backup service.

[192] See for instance the list of Protector (<Cyberlink=751>).

[193] Such a network can be local (local area network – LAN) or more extended (wide area network – WAN). If access is provided over the Internet – as in most cases in academia – the institutional networks often have a public and a private part, that is an area to which only the members of the respective institution have access. The latter is called Intranet.

As soon as connections to the worldwide network become ubiquitous[194], mobile computing (notebooks, E-paper) will make research work in the digital world, in principle, independent from space and location. Xerox' experiments with TELEWEB (Schilit et al. 1998b) enable offline working with online documents, PARCTAB (Want et al. 1996) integrates palmtop devices into office networks, DYNOMITE (Wilcox et al. 1997) improves the ordinary notebook with organising and audio functions.

In addition, there is the vision of "ubiquitous computing" (UC) of technology development in the field of ICT. Computer based devices will become so small, will be produced so cheaply, will co-operate seamlessly and will be so easy to use that they will assist a multitude of everyday activities in an almost imperceptible way. This goes far beyond a notebook and digital personal assistant (palmtop) connected via a mobile phone to the Internet. The idea is that many artefacts of everyday life become "smart" or "intelligent" in the sense that they include a computer chip capable of carrying out specific tasks. While most concrete examples of UC reported so far belong to home technologies (such as the refrigerator ordering new milk), there are also possible applications in academia. For instance, the pencil with which a researcher makes a note might be able to store the note and transfer it to one's office PC.

2.3.4 Databases: digital libraries, archives et al.

While until not so long ago data were collected in book-long lists or card files, today, digital databases of all sorts are becoming ever more widespread. Note that Internet users in general, and researchers in particular increasingly interact with databases without even noticing.[195] While this is not the place to explain the details of current database architectures,[196] a number of useful distinctions are made in the following with a view to understanding what Finholt/Olsen (1997, 30ff.) call "people-to-information links". I deal with libraries and databases under the same heading because in the age of cyberscience they go together intimately, as will become obvious immediately.

2.3.4.1 Typology

Nearly all academic libraries have gone online. We may distinguish the following types:[197]

- *Online Public Access Catalogues* (OPAC) are bibliographic databases which, in general, include only meta-information (accession number, author, title, source, keywords, status etc.) about the physical holdings of a particular library, but not the full text. In

[194] For instance, there are already Internet cafés providing for wireless Internet access of the notebooks of their clients.

[195] As we shall see in 2.4.3 below, WWW sites are increasingly dynamic, that is the individual pages seen by the users are not stored as static, individually addressable files, but are the result and representation of a database query.

[196] It suffices to say that there are various types with the so-called relational databases being increasingly widespread. They use relations or two-dimensional tables to store information and the information in a series of tables can be linked through common columns. This is a very powerful – although not universal (cf. Hockey 1997b on the limits) – way of structuring information with a view to allow sophisticated search-queries.

[197] Harter (1996b) gives a short overview of the history of the term "digital library" and related terms and then concentrates on the properties of a digital library, distinguishing between a narrow (i.e. based on the traditional library), a broader and a broadest view (i.e. loosely based on the current Internet).

other words, this is the digital version of the old file cards catalogue.[198] While earlier OPACs were only accessible via the telnet protocol, today most libraries have a WWW interface (so-called WebPACs).199 Many OPACs are part of a library network (or "clump") and can be searched simultaneously.[200]

- *Digital libraries* (DL) are structured collections of digital information which are provided via digital (global open) networks with mainly the Internet or the WWW as the user interface (Grötschel/Lügger 1996, 7). The main point is that the full text of each item is available for downloading by users.[201] A DL may be run as a database, but often a simple (structured) list of available texts suffices.[202] On different approaches to construct and manage DLs, namely as a closed system (on the basis of database systems like HYPER-G)[203] or as an open system (on the basis of the WWW with so-called distributed linking services[204]), see Carr et al. (1996b).[205]
- *Virtual libraries* are not primarily document collections held at a particular server, but rather access points or link collections (or portals[206]) pointing at full text material available online at servers world-wide. In this sense, a "virtual library" is a library with no physical place to go to and no collection to look at in analogue format. Such a library exists only inside servers scattered around the globe and is often managed by people equally distributed all over the world. There is a central WWW virtual library site[207], which is a collaborative endeavour of many virtual library projects around the world.[208]
- *Hybrid libraries*[209] are libraries which offer both physical access to print holdings and online access to digital resources, that is they are half traditional libraries and half digital/virtual libraries.
- *E-library* is used in our context here as a generic notion comprising all variants of libraries using some form of ICT, be it a simple OPAC or a sophisticated virtual library databases.

198 "ICT has expanded delivery methods for bibliographic databases and has created better options for storage, search and retrieval." (OECD 1998, 200)

199 The largest of all libraries with a public access interface is the US Library of Congress (<Cyberlink=743>).

200 In Austria, there is a statewide academic OPAC system called "Österreichischer Bibliothekenverbund" (<Cyberlink=742>); see also the Scottish clump project CAIRNS (<Cyberlink=744>).

201 Perhaps the most prominent example of a DL is Project Gutenberg (<Cyberlink=26>); another general DL is the Internet Public Library (<Cyberlink=398>). In addition, there are many field-specific DLs. For instance, the ACM (Association for Computing Machinery) has a large DL (<Cyberlink=113>).

202 In contrast to the US Gutenberg project, the German variant Gutenberg-DE is not a database (<Cyberlink=313>).

203 Now: HYPERWAVE (<Cyberlink=43>).

204 DLS (<Cyberlink=418>).

205 There are many national and international digital library initiatives (for an overview see e.g. OECD 1998, 203).

206 For these notions, see also below sub-section 2.4.3.

207 <Cyberlink=603>.

208 The use of the two terms "digital library" and "virtual library" is often not consistent in the literature and sometimes differs from the above definition. In this quote by Owen, I would have rather used "virtual" instead of "digital": "The digital library is not a library in the traditional sense, but an often global organisation of scientists or scholars who use advanced technology to create and share information over the network." (2000, 4)

209 This term is used in the U.K. Electronic Libraries (eLib) programme (<Cyberlink=154>).

Most E-libraries not only offer digital access to their own holdings, but also to other databases (which are, in principle, independent from the library setting and can also be accessed individually on the basis of individual or institutional subscriptions):

- *Abstracting databases* developed from the earlier paper-based indices or abstract services which collected the meta-data of academic (mainly journal) literature, keyworded each article and took over an existing or edited a new abstract. Today commercial publishers publish them online. Some of them cover wide areas of science and research[210], others are specialised in a discipline or sub-discipline[211]. Recently, these databases have begun to be integrated with the full text journal databases (see below) so that the user can directly access the full text (when her/his home institution has paid the respective fees to the full text database provider).
- *Reviews databases* are similar to abstracting databases, but offer a genuine review of it, that is a short note or longer article written by a non-anonymous author giving an overview and a critical appraisal of the reviewed article.[212]
- *Full text journal databases* are at the core of hybrid libraries, but may also be accessed without any institutional connection to such a library. There are a number of providers of such databases, none of them being comprehensive. They are either large commercial publishers[213] or university or library based consortia[214]. Recently, some of these databases have gone well beyond simple lists of journals and journal articles and include sophisticated, automatically generated hyperlinks. These links allow users to "jump" directly from one article to the next either because it is quoted in the first article or because it includes the same set of keywords qualifying it as a "related article" (see below 2.3.4.3).

Furthermore, there are various databases made by the research community itself, by international organisations, governmental bodies or commercial providers:

- *E-(pre-)print archives* are collections of digital research papers. There are central archives where all full texts are submitted to and stored on one server – these are called "E-print archives".[215] And there are decentral archives where the central archive website holds only the common search-engine but not the original working papers – called "meta-archives". The papers are stored decentrally, i.e. on the servers of the publishing institutions; only meta-data[216] and perhaps indices[217] of the papers are to be found locally. In general, the researchers run these archives themselves. The Open Archives initiative aims at making these archives interoperable and therefore searchable across sites.[218]

[210] Among the larger ones are Journals@Ovid (<Cyberlink=704>) and ISI Web of Science (<Cyberlink=488>)

[211] E.g. the Zentralblatt for mathematics (<Cyberlink=749>).

[212] An innovative review database is the Faculty of Thousand (<Cyberlink=646>); also the Zentralblatt mentioned in the last fn. contains reviews.

[213] Such as Elsevier's ScienceDirect (<Cyberlink=746>) or Springer's LINK (<Cyberlink=747>) services.

[214] Such as the Journal STORage project (JSTOR; <Cyberlink=322>).

[215] Certainly the most prominent example is arXiv, the world-wide physics pre-print archive (<Cyberlink=216>).

[216] A typical example of a meta-data-only E-print archive is the huge Research Papers in Economics archive (RePEc; <Cyberlink=214>).

[217] In this context, "indices" means condensed, machine-readable versions of texts; the purpose of this is to accelerate the search-engine as, by this token, it is not necessary to connect to the remote/decentral sites each time a full text search has to be carried out. An example of this type is the European Research Papers Archive (ERPA; <Cyberlink=215>).

[218] <Cyberlink=60>.

- *Other academic databases* may have diverse content and diverse providers. They may collect statistical or other numeric data, texts for linguistic purposes, images of archaeological artefacts, legal texts etc. (see also 7.2.4.3). We may distinguish between two different architectures. Most databases are typically *central* that is all the data is located at one place (server), but accessible from everywhere. Others have a *decentral* architecture which means that the data is dispersed over more than one server; the software integrates the data into a virtual unit.[219] The "decentral type" may be increasing as it corresponds to the networked nature of the WWW.
  An interesting platform for academic databases in the social sciences is SYNAPSEN[220]. This is a hypertextual card file (Krajewski 1997) which enables users to store both bibliographic information and other ideas, interconnected with hyperlinks. Krajewski is planning to make SYNAPSEN apt for collaborative and online access.
- *Knowledge bases*, in particular on the Web, are special databases aiming at providing neither data nor information, but knowledge – a very ambitious aim. The idea is to represent knowledge (declarative information) in some formalised way (classification, ontologies) and inter-link it in a database. Artificial intelligence research focuses on inferences from large bodies of declarative information. Such knowledge bases – or to use a related label: Semantic Webs – will be constructed collaboratively. There are already a number of tools available and many projects under way (cf. e.g. Euzenat 1998).[221] See also 6.2.2.1 on hyperbases.
- *Digital academic archives* are collections of digital documents of all sorts (text, video, audio) relevant for a particular research field. In many cases, the items of such archives are scanned and digitised. Mueller (2000a, 8) speaks of the "digital surrogate archives" which contain also retrospectively converted matter.[222] Digital archives may be holographic (images, scans) or allographic (searchable full-text) or hybrid (ibid., 9, quoting Nelson Goodman). The latter form combines page images with indices[217] based on optical character recognition (the errors in recognition are low enough not to spill the overall indexing result). Mueller (2000a, 9) distinguishes further between
  - *macro* archives which are "massive collection[s] of documents submitted to the minimal processing that makes them searchable by standard search tools"; and
  - *micro* archives where "bod[ies] of materials [are] coded specially to be processed by more granular search tools" – a much more labour-intensive type.

2.3.4.2 Access to databases

From the point of view of how the database can be accessed, one can distinguish between
- *offline databases,* which are distributed in CD-ROM format;
- *online databases,* which can be accessed through the Internet, mainly using the telnet-protocol; and
- *web databases* that have a WWW interface.

[219] An interesting example in this respect is the Distributed Annotation System (DAS) in biology: the annotations to gene sequences are not stored centrally, but decentrally and in the same format; the central server integrates the annotations to a particular sequence and presents them to the user in a unified form (Biodas <Cyberlink=765>).

[220] <Cyberlink=69>.

[221] E.g. the High Performance Knowledge Bases (HPKB) programme and its followers (<Cyberlink=423>).

[222] E.g. the Thesaurus Linguae Graecae (<Cyberlink=343>), Women Writers Project (Orlando Project; <Cyberlink=371>), Perseus (<Cyberlink=373>), JSTOR (journals; <Cyberlink=322>).

Today most online databases are accessible through web search forms. Therefore, online and web database are used synonymously in most instances. From a practical point of view, databases and archives can be accessed either by

- *browsing*, that is by stepping through (hierarchical) categories or chapters and subchapters (this was the only way to retrieve information from lists, catalogues, indices, encyclopaedias, dictionaries etc. before the advent of cyberscience); or by
- *searching*, that is by using a search-engine; this allows not only to directly access items in the database, but also to easily find related items in different categories. Often, databases have two query forms: one simple for quick access, which gives you only access to a limited number of fields (e.g. only author and title), and a more sophisticated one, which allows the user to formulate very targeted search queries by combining search words in different fields of the database ("this author in this period of time with a title similar to this, but excluding that"). Sometimes, there is also an "expert modus" which gives the user the opportunity to formulate queries in a special query language. The latter is the most sophisticated form of communicating with a database, but the user needs to master this special language to obtain convincing results.

In most cases, both access routes are available. Increasingly, retrieval software is labelled "intelligent". This denotes the development of less crude and more targeted retrieval software. While early search-engines may only be able to find words, combinations of words or parts of words, eventually in a particular context, newer retrieval algorithms may behave more "intelligently". This means that they can "learn" users' interests or use meta-data embedded in the database (e.g. XML – see already above 2.2.2.1).

A special form of access to a data base is offered by the statistical office of Canada, for instance, called remote data access (RDA)[223]: researchers write and test their own computer programmes using a file with artificial data on the webpage. After the test, they can send these programmes via the Internet to Statistics Canada, where they are run on the original (official) data file. The results are then sent back to the researcher.

Many databases are not accessible for free. There are various forms of access control:

- *User-ID and password*: the user enters a restricted area of the database homepage after having successfully identified him/herself;[224]
- *IP-address check*: the database checks the IP-address of the computer from which the access request has been sent. In case the IP-address is valid, the user is granted access. The list of valid IP-addresses can either be on a case-by-case basis and/or on a domain basis, that is all IP-addresses of a domain, e.g. an institution, are allowed to access.

With regard to restricting the use of retrieved items from databases and archives (such as documents, articles etc.), the commercial publishing industry is presently developing the "*Digital Rights Management*" (DRM) system.[225] The idea is that in all copyright-protected digital material, all allowed uses of the particular purchaser are encoded (in a non-erasable way). Software reading the document (or playing the music or presenting the movie) checks whether the requested use is permitted and acts accordingly by either denying or granting it. Given the various problems involved (privacy issues, undue

[223] <Cyberlink=731>.

[224] Bishop (1998) discusses ways of measuring access, use and success of digital libraries and presents various user studies performed at the University of Illinois to test their digital library project (DeLIver). One of the main results is that authentication and registration procedures presented an enormous barrier to use (many things can go wrong).

[225] See e.g. <Cyberlink=759> and for the development of an open standard <Cyberlink=758>.

restriction of fair use etc.), it is not yet clear whether DRM will be implemented and in what form (see chapter 9).

With a view to guarantee the authenticity of digital documents (and also be able to trace copyright infringements), a number of technologies have been developed. Among them is *digital water marking* which is often part of DRM. Special software[226] embeds information about the author into digital files (in particular audio and video, but also other documents). This information, when decoded with the appropriate software, can reveal things such as the author's address, terms of use, copyright date etc. Watermarks are not removable and not alterable. The information does not degrade with file duplication and does not perceptively disrupt the original data file.

2.3.4.3 Cross-linking

Increasingly, digital publications (cf. 2.4.4) stored in databases refer to each other. Hitchcock et al. discuss at length various models aiming at linking from E-journal papers to other papers and resources (Hitchcock et al. 1997c). They argue that content integration will drive web publishing, i.e. making content easily accessible across the boundaries of databases and bundles offered by commercial publishers. They describe how it is technically feasible to create links automatically whereby the links are stored in so-called link databases ('linkbases'). A document will be represented by a core information unit in which meta-data about the various representations of the document is stored: where the full text in which formats is stored, where the bibliographic entry, where an abstract from which services etc. In an open environment, links may be enhanced and bi-directional. In the future, links

> "could be differentiated by colour (or simply pruned according to particular thresholds) according to whether they are hand-authored specific links, machine-generated general links, recently created links or links belonging to (for example) a course tutor. This adds to the user's control over the view of the document's connectivity, which currently exists only at the macro-level by including or excluding whole linkbases." (ibid., 14)

CrossRef[227] is an initiative of commercial publishers to allow for cross-links between journal articles of different publishers. The system is based on the DOI scheme (cf. 2.1.2). Also the Open Journal project up to 1998 and now the Open Citation project[228] aim at creating an environment in which links back and forth between papers is possible, starting with the E-print archive in physics (e.g. Hitchcock et al. 1997b). See also the other advanced schemes in the physics HyperCite[229] and the APS Link Manager[230] (Hitchcock et al. 1997b, 5) Among others, also the Web of Science by ISI[231] is an implementation of sophisticated linking technologies "putting users in control" (Hitchcock et al. 1998).

A very sophisticated cross-linking experiment is PeP (Perspectives in Electronic Publishing)[232]. This is a bibliographic database which provides access to freely available full text in a special field (E-publishing) – the full text of (most) articles included does, however, not reside on that server, but on the original sites. If the user grants permission,

[226] E.g. Digimarc (<Cyberlink=780>).
[227] <Cyberlink=376>.
[228] <Cyberlink=59>.
[229] <Cyberlink=778>.
[230] <Cyberlink=777>.
[231] <Cyberlink=488>.
[232] <Cyberlink=494> (Hitchcock 2002).

an "applet" (tiny computer programme) transforms the full text while downloading it from the original location and before presenting it to the reader. The text you see is amended with special PeP-links that may lead you either to the full text or bibliographic entry (from quoted literature) or to a list of related bibliographic entries (from specially marked keywords throughout the texts). These keywords belong to a growing thesaurus of E-publishing managed by the bibliographic database. By this token, the reader browsing through this wealth of published information can move quickly to related and further articles.

2.3.5 Web search-engines and directories

Given the lack of a coherent structure of the WWW when it comes to content[233], one of the most frequently used services in the Internet are the various search-engines. While search-engines are often found also at individual homepages giving you direct access to the content of that homepage, web search-engines give you access to the whole WWW or parts thereof. Although most search-engines are hybrid, we can distinguish three basic forms:

- *Web directories* provide the user with a hierarchically structured access to addresses in the Web. The user browses top-down through the different categories, which are built-up either by user-suggestions or by an editorial board.[234]
- Genuine *search-engines* let you search in huge index databases which contain the full text and the meta-data (if available) of all harvested webpages. Special programmes (called "harvesters" or "spiders") regularly visit all sites registered in a core list and upload condensed versions of the webpages to a central database. There are various algorithms (programming procedures) to present the user with a most appropriate search result.[235]
- *Meta-search-engines* let you search via more than one search-engine (or database) at the time: you use one form to enter your query and the meta-search-engine sends it to a number of search-engines and collects and presents the results in some coherent form.[236]

2.3.6 Web forms

On WWW pages you often find forms to fill in. The data entered in these forms is sent to a server when you click on the "send" button. The data can be of various types: from simple comments (that is as a web-surrogate of E-mailing) to placing orders for books or reports, to questionnaires. From a technical point of view, the coding language of the WWW (HTML) provides for these forms as data input. What is done with the data depends on the receiving end, that is the programme on the server. In simple cases the data is just

[233] In contrast to the hierarchical technical structure of the Internet with IP-addresses and the domain name system.

[234] A typical web directory is ALTAVISTA (<Cyberlink=761>).

[235] A typical, and at the time of writing the most successful web search-engine is GOOGLE (<Cyberlink=760>).

[236] A typical general example is METACRAWLER (<Cyberlink=762>); in economics, for instance, there is a specialised meta-search-facility (<Cyberlink=614>); another example is the virtual catalogue for German libraries KVK (<Cyberlink=663>).

mailed to a particular E-mail address[237]. In more sophisticated cases the data serves as input for search-engines (see above) or for amending databases.[238] Even entire project proposals of far reaching financial and other implications can nowadays be submitted via web forms.[239]

2.3.7 Remote control et al.

What Finholt/Olsen (1997, 30ff.) call "people-to-facilities links" plays an increasing role in some science fields. Remote control is understood here in a broad sense including all sort of remote handling of instruments and computers. For instance, researchers may access special-purpose-computing resources, such as supercomputers and sensor-based instrumentation. They have data viewers at their disposal, which display the current modes and status of remote instruments and experiments (e.g. the temperature or the co-ordinates of a robot). These remote control devices and processes are at the core of the so-called virtual laboratories.[240] Under the heading "Scientific instruments", the OECD report even speaks of virtual instruments because "the user, rather than the instrument maker, determines precisely what the equipment does by matching the software to the sensor needed for each measurement" (OECD 1998, 210). In other words, as a scientific instrument has both a hardware and a software component and the latter has become increasingly important, the software necessary to run the hardware can even be located elsewhere. However, only the two together (hardware and software) form the instrument (and are linked to each other and to the user via ICT). In this sense, the instrument is "virtual" as there can be many of them (using the same hardware) at the same time.

2.3.8 Speech recognition

New human-machine-interfaces have been developed which enable speech-to-text as well as text-to-speech conversion. The idea is not to use the keyboard, the mouse and the screen to communicate with computers (and the Internet), but to use the microphone and natural language. Latest advances are surprising and may well revolutionise the way we think about people-to-machine communication.[241]

[237] And can then be said to be not people-to-machine, but people-to-people communication, see below 2.4.

[238] The CYBERLINKS database gives examples for all three types: the feedback form sends a mail to this author, the search form triggers a database query, and the "add link" form manipulates the database entries.

[239] PROTOOL is the respective tool of the European Union (<Cyberlink=96>).

[240] For an example see the High Flux Isotope Reactor virtual laboratory in Oak Ridge (<Cyberlink=753>).

[241] See the Natural Language Software Registry (<Cyberlink=768>) and as an example WaveToText (<Cyberlink=769>).

2.4 People-to-people communication

What Finholt/Olsen (1997, 30ff.) call "people-to-people links" (2001, 22f.) is at the heart of this study: innovative tools to enable communication between researchers. Based on my distinction in 1.2.1.1 (means, properties), the following table lists the new tools:

Table 2-2: Typology of electronic people-to-people communication tools in academia

	Co-operation (conversation and correspondence)	**Publishing**
One-to-one	E-mail E-conferencing tools (bilateral type) chat tools	–
One-to-many/few	Mass E-mail E-conferencing tools (lecture type)	E-newsletters Homepages Self-publishing
Few-to-few	Discussion lists E-conferencing tools (seminar type) E-teaching tools Groupware, CMS Chat tools	(Discussion lists: skywriting) Project-specific working paper series
Many-to-many	Distribution lists (newsgroups) Link collections Shared databases Software sharing	E-(pre-)print series E-journals Frequently asked questions (FAQs)

The following sections will present these cyber-tools in turn.

2.4.1 E-mail

Besides the increasingly ubiquitous WWW, E-mail is certainly the most prominent and widespread application of cyberscience. E-mail was one of the very first things researchers did with the new network. In general, users in academia have an E-mail account with their local university computer department, which runs a "mail server". This server provides for digital mailboxes for each user in which incoming messages are stored until the user connects to the server in order to retrieve them. This can either be done by a local E-mail programme residing on the computer of the user (such as EUDORA, PEGASUS and Microsoft OUTLOOK) or by a server-based mail programme (e.g. the telnet-based programme PINE, or the web-mailers like IMP or MAIL2WEB). Mails can either be retrieved and stored in local mail directories (and deleted from the server) or they can remain on the server. IMAP is a relatively new protocol that is not yet used widely in academia. It provides for a standardised format of mail directories and messages that are stored centrally. This enables users to access their mailboxes and stored messages in mail directories wherever their present location and regardless of the hardware and software at hand.

By this token, researchers access their mails and mail archives both in the office and from home, and from abroad (for instances at conference venues).

E-mail can not only be sent to one, but easily to many addresses (one-to-many; in the case of very many addressees, we speak of mass mail; see also below). Furthermore, the option to attach documents to E-mails is increasingly important in academia. The documents reach the addressee together with the mail message. They can be stored locally and opened with the appropriate programme. This can be done with all types of digital files, such as texts, images, audio files etc. While E-mails were initially rather simple and plain text, formatted text in rich text format (RTF) or in HTML have become ever more widespread recently. Today, text emphasis and colour, different font sizes etc. are often found in E-mails. There are still a number of older mail programmes that have problems with „rendering", i.e. representing more sophisticatedly formatted messages and there are incompatibilities between mail programmes and between servers (in particular when it comes to attachments). In general, however, E-mail works very well.

2.4.2 E-lists

Electronic mail is not only used for bilateral (or one-to-many) but also for multilateral communication. We can distinguish two ideal types according to the *principal* communicative activity:

- *distribution lists* where information of all kind (e.g. calls for papers, conference announcements, advertisement for publications, academic gossip etc.) is shared among those subscribed;
- *discussion lists* where subscribers engage in an exchange of mails related to continuously changing topics; one may further distinguish between
  - *forums for questions and answers* where the main purpose of "postings", i.e. mails to the list, is to ask for specific bits of information, and
  - *forums of debate* where the focus is on the exchange of opinions and academic discourse.

One speaks of "threads" to denote related postings, i.e. postings with the subject line.[242] Most lists are a combination of both main types, i.e. with varying shares of discussion and distribution. To my knowledge, there is no empirical study giving any indication of the shares in reality. However, based on my expert interviews, it seems that the majority of lists are distribution lists with elements of the "questions and answers" type, rather than genuine forums of debate. Even those of the second type have a considerable share of information distribution, but are stricter when a member of the list abuses the list by advertising or sending off-topic stuff. As it becomes easily confusing and frustrating to join a debate by many (also in a real world setting), it seems likely that debate rather takes place in small lists or among very few active participants in a larger list. The larger the list the more it will be of the distribution or "question and answer" types.

From the perspective of the technical implementation, there are various forms of E-lists (namely mass mailers, list server lists, newsgroups and web boards):

[242] Note that it requires disciplined posters in order to change the subject line if a new topic is started (often people only press the Reply-button and automatically use the old subject line) Otherwise, you often find unrelated contributions in the same thread (see 4.2.2.2).

- *Mass mails*[243] are the simplest type of distribution lists. They are not handled by a mail server programme but by the client-side mail software.[244] In this case, the user sends the mailing simultaneously to any number of addresses. However, this is only convenient if the list is not too large as the administration of such lists may be cumbersome. Everything related to the "subscription" has to be done manually. Furthermore, the addressees have no control over their being on the list. In a strict sense, this is not multilateral (many-to-many), but one-to-many communication – although, if the addressees' list is not suppressed, every addressee may use it for his/her reply.
- *List server lists* are based on a so-called "list server" programme (like "listserv", "listproc", "mailman" or "majordomo") which administer lists. People can subscribe and unsubscribe by sending a short note to the list server. The subscribers use their own E-mail client to communicate with the list server. The server distributes E-mails sent to the list's address (the so-called "postings") to all E-mail addresses on the subscribers' list. Subscribers get mails to the list either immediately and separately, or in "digested" form, that is all postings of a particular period in a single mail with a table of contents. The so-called "owner" or "moderator" of the list is able to perform a number of administrative tasks via the list server programme, such as adding or removing subscribers, filtering out "spam" (i.e. unrelated E-mails) etc.
- *USENET newsgroups* differ from list server lists in so far as the user does not subscribe to a list (or group) by registering his/her E-mail address with a central server, but by advising a local programme (the newsgroup reader) to look up whether new postings in a particular group have arrived. While list server users wait for new mails from the list to arrive in their standard mailbox, newsgroup users load a special programme. Postings to the group remain on the central server. In some clients, the postings are presented to the user in a subdivided window showing both the list of old and new postings as well as the text of posting which is marked in the list. There is no moderator in such newsgroups. While there it is no central directory of list server lists,[245] it is rather easy to find appropriate newsgroups as each "news server", i.e. a server hosting own newsgroups, also "mirrors" hundreds of other newsgroups around the world in well structured directories. However, there are not too many academic newsgroups available.[246]
- *Web-boards* are different once again. They are hosted on (dynamic) WWW pages, which show in a structured manner all previous postings. If someone wants to reply to a message, s/he fills out a web-form. The entries in the form are received by a programme, which then adds the messages to the webpage at the appropriate place. There are web-boards which require subscription so that postings are only accepted if you enter a user-ID and password in the form, but most of them have no such restrictions.

[243] This function is called differently in the various E-mail clients, for instance "nickname", "contact" or "distribution list".

[244] From this simple form we would have to distinguish the professional mass mailing programmes which make life easy for advertising companies, but also increasingly annoying for users. These programmes have various programme modules that gather E-mail addresses from the web, administer them in a local database and use them for mass mailings with thousands of mailings sent out in minutes.

[245] See, however, the unofficial and not constantly updated DSEJ (<Cyberlink=180>).

[246] Lewenstein acknowledges that the USENET system "is still not designed for professional scientists", and predicts that "the distinction between 'the' net (the whole, publicly accessible cyberspace) and smaller electronic spaces with more limited access may be seen more often" (1995, 143).

While most academic lists are list server lists, there are not too many academic newsgroups (but thousands of leisure and special interest groups) and relatively few webboards.

In administrative terms, there are

- *unmoderated lists* where there are no restrictions to postings to the list; either everyone subscribed is allowed to send messages or even not subscribed persons can; and
- *moderated lists* with a so-called moderator which filters all mail (or particular categories of mails) to the list with the power to deny forwarding of unrelated or otherwise unsuitable mailings; often, the moderator is the driving force behind such a list and frequently posts her/himself with a view to generate debate; furthermore, the moderator can deny individuals access to the subscribers list or even unsubscribe users.

Moderation of a list is a time-consuming and often tricky business. The moderator needs to establish a balance between, on the one hand, regulating what is necessary to secure a genuine debate or that the list is not abused and, on the other, leaving the necessary freedom to the subscribers so that they do not feel censored. Rost (1998c) discusses the problem of the undemocratic threat of moderation by a single person and proposes, as a remedy, the so-called "scoring server". Subscribers to the list would send simple scores about each contribution to the E-mail address of the scoring server which analyses them together with data about how many replies a contribution has triggered (in other words: how long the thread became). Regular reports inform the subscribers about thread-quality and content-quality with the aim of discouraging low quality contributions.

Among the distribution lists, we can further distinguish between

- *one-way lists* where only a tiny group of people or even only one member of the list is allowed to post messages, and
- *both-ways lists* where each subscriber is also allowed to send messages.

Scholarly associations often use the formers in order to distribute newsletters, publication announcements and other associational mailings. Also E-journals use them to inform their subscribers of newly published papers or issues. They come close to mass-mailings but differ in that the subscribers may – in most cases – subscribe or unsubscribe at their own discretion. This first type is logically not apt for a discussion list.

Most listserv and newsgroup servers archive all postings to a list or group. Hence, postings do not disappear, but are still accessible later (as long as they are kept in the archive, cf. 7.4.1).

2.4.3 Homepages et al.

The WWW allows both individual researchers and research institutes or universities to have a cheap "window to the world" which is furthermore easy to maintain (see 6.4.4.2). The words used to denote the WWW pages are confusing. I propose here the following usage:

- a *webpage* is the individual file, mostly written in HTML format;
  (A typical address (URL) of a webpage would be:
  http://www.oeaw.ac.at/ita/cyberscience.htm – which is the main page of this project; it contains a description of this project and links to further related webpages.)
- a *homepage* is the starting page of a coherent set of webpages of one institution or person; sometimes the notion of "homepage" is also used to denote the whole set of webpages;

(A typical URL of a homepage would be: http://www.oeaw.ac.at/ita/ – which is the page of the home institute of this author; it gives access to many further webpages containing information on a variety of aspects of the institution.)

- a *website* is the top unit and may consist of one or several homepages which are all reachable under the same main name and which all reside on the same server. *Webspace* is mainly used synonymously with website.
  (A typical address (URL) of a website would be: http://www.oeaw.ac.at/ – which is the site of the parent institution of this author's home institute; it contains many more homepages than the one of the above institute.)
- a *domain* is the technical term for all elements of URLs between the initial two slashes ("//") and the third slash ("/").[247] For practical purposes, "domain" is often used synonymously with "website".
  (Top-level domains are those defined by the two, three or four letters between the last dot (".") and the third slash. In my example above, `at` is the top-level domain. `ac.at` is the academic network domain in Austria; `oeaw.ac.at` is the domain of the Austrian Academy of Sciences; www.oeaw.ac.at is the WWW domain of this institution, but there is also a mail.oeaw.ac.at domain for the E-mail server etc.)

As regards content, I speak of a *web portal* if the website is designed to give comprehensive access to all resources related to a particular topic. While link collections only point to resources residing on other servers, portals offer in addition information residing on their own web server, such as introductory comments, newsletters, overviews etc.

Furthermore, webpages can be either

- *static*, that is they are fixed files whose content can be changed by editing the file itself; or
- *dynamic*, that is they are generated each time a user requires seeing them (via his/her web browser) on the basis of the current entries in a database at that moment.

So far, most webpages in academia are of the static type, but the trend to dynamic pages is growing. In general, websites are a combination of both static and dynamic elements (like the ones mentioned above).[248] Webpages can be very simple (as most academic homepages still are), that is mainly text-oriented. Some, however, feature sophisticated (and often expensive!) designs with graphic elements and little scripts embedded which animate it. For instance, the links to further pages can be so-called "buttons" which change colour or form when the mouse-tip "touches" them. Often the content of a page (whether retrieved from a dynamic database or read from static files) is separated from the presentation. For instance, so-called cascading style sheets (CSS) are like "templates" which advise the browser how to present each element of the content.

Producing ("editing") webpages can be easy (if one does not intend to produce dynamic multimedia pages), as there are a number of tools available to help. In particular, most of the widespread word-processing software enables users to store their text files in HTML format. So far, however, the results are not convincing in visual terms, as HTML is less layout-oriented. Furthermore, there are many different web-browsers in use, of-

[247] See already above, 2.1.2.

[248] For instance, the CYBERLINKS collection is a database. Except for a few static pages (e.g. the "disclaimer" page), everything the user gets to see is a representation of a search result in the link database. Every link is stored in a "table" with a number of columns for identification number, short title, URL, date of entry, descriptive text etc. The script to access the database produces an HTML file on the basis of the search result in combination with a number of text elements, such as the footer and header of each page.

ten in older versions, too. All of them "translate" (or "render") HTML code slightly differently. What a user sees on his/her computer screen when producing the page is therefore often not the same as everyone else is seeing. This makes webpage editing tricky and visiting "home-made" homepages often frustrating. In general, however, the quality of webpages is constantly rising, also in academia in both the content and the layout dimensions.[249]

2.4.4 Academic E-publishing

Under this heading, I shall present both the various new formats of publication and the tools for the organisational side, i.e. applications for electronic submission, reviewing, editing etc.

2.4.4.1 Formats

Meanwhile, all forms of academic publishing have their counter-part in the electronic world. There are electronic working papers, pre-prints, off-prints, journal articles, single-authored books, text (student) books, conference papers, newsletters etc. In addition, a number of new forms have seen the light of the "cyber-day" (see below 7.2.4). E-publishing is evolving (see Table 2-3 below). In the early days, E-books[250] were distributed on diskettes (e.g. Kolb 1994); today, they mostly come on CD-ROM[251] or are available online[252]. The early E-journals were distributed by E-mail (via distribution lists)[253] or via Gopher or anonymous FTP; today, they are almost exclusively online. However, some of them also offer a yearly CD-ROM for archiving purposes.

Working papers and pre-prints have been (and sometimes still are) distributed by E-mail, but there is now a growing trend towards "up-loading", that is sending them in electronic version, to servers which make them accessible for everyone over the Internet. There are working paper archives[254] and E-pre-print servers[255] (see above 2.3.4). The same is true for conference papers: while before the conference, they are still distributed among the panellists by E-mail, they are increasingly made available on conference servers.[256] The typical post-conference proceedings volumes are nowadays often replaced by either these online servers or by CD-ROMs sent to all participants. The latter has the advantage over the standard book format that all contributions can be included in full length, also with additional figures or databases.

[249] See (Nolte 1998) for a comparative description of the state of art in Germany in 1998.

[250] For an overview on E-books, see e.g. (Rink 1999; Vinzant 1998); a user's wish list is presented by Bryant (1997).

[251] See for example the English Bible project of University of Michigan Press (Miller-Adams/Trager 1998) which includes digital pictures of papyrus of various sources, accompanied by a text book, including English and Greek transcripts.

[252] See, for instance, the German legal textbook on Internet law (<Cyberlink=671>); further E-books are to be found as dissertations on university servers, e.g. the Networked Digital Library of Theses and Dissertations (NDLTD <Cyberlink=286>).

[253] E.g. "EJournal" (<Cyberlink=729>).

[254] For instance ERPA (<Cyberlink=215>).

[255] The first and most prominent being arXiv (<Cyberlink=216>).

[256] An example is the "PROceedings" website of the American Political Studies Association (APSA) which makes the conference papers available for one year until the next annual meeting (<Cyberlink=728>).

Table 2-3: Academic format and current standard in technical implementation of E-publications

		Technical implementation			
		Offline		Online	
		Diskette	CD-ROM	E-mail	WWW
Academic format	Working paper			x	x
	Conference paper		x	x	x
	Journal article		(x)	(x)	x
	Book	x	x		x

As indicated in the above table, we can distinguish between offline (CD-ROM and diskette), and online or Internet publishing (E-mail and WWW). In this study, I focus on the latter. Note, however, that often CD-ROM publications also come in an online format. The advantage of CD-ROM is that the present[257] slowness of the network is no factor and that sophisticated proprietary software may be included (which is much more restricted in online publishing).[258] W. S. Strong points at the differences between CD-ROM and related technologies, on the one hand, and online publishing, on the other: While with the former, a tangible product is physically transferred, this is not the case with the latter. Therefore, the former is much more analogous to classic publishing (1995, 1).

Academic E-publishing comes in various formats (see Hitchcock et al. 1996, 7f. for a comparison in 1995 STM online journals; also Hitchcock et al. 1997a, 11ff.), listed in Overview 2-1:

TYPE	DESCRIPTION
Page image	Graphical, not textual representation of each page
HTML	The standard format of the WWW which comes in both "raw" and "enriched" flavours
ASCII	Plain text
TeX	Widely used in physics and mathematics
SGML	Standardised General Markup Language*
XML	Extended Markup Language, there are various "dialects" specific to purposes and fields
PDF	"Portable Document Format" from ADOBE
PS	PostScript, originally also from ADOBE
RTF or DOC	Rich Text Format or Word format from MICROSOFT

Overview 2-1: Formats (technical) of E-publications (* see fn. 259)

[257] See, however, the initiative to create Internet II (2.1.2 above).

[258] Armstrong (1998, 28ff.) also discusses the pros and cons of the two media and different approaches of U.K. publishers.

[259] SGML is an international standard for the definition of device-independent, system-independent methods of representing texts in electronic form. Both HTML and the various XMLs conform to this standard (for an introduction see <Cyberlink=340>). "SGML is all about structure and mean-

The latter three are so-called "proprietary" formats meaning that firms have developed them and can and do develop them further at their own discretion. The former seven are general standards with an international procedure to amend them. The formats are also distinctive with respect to how much information can be included, in particular as regards layout. While for instance, a heading of level no. 1 can in HTML be displayed differently in different web-browsers, a PDF or TeX document looks the same everywhere. In addition, many E-publications come in compressed ("zipped") form in order to minimise file size and hence storing needs and transmission times. Users need a de-zipping programme to be able to read such files.

As to electronic journals[260], we have to distinguish between

- *E-only journals* which are online-only and have no printed counter-part (in this category, I also include those E-journals which print a yearly volume for archival purposes while delivery of the newest issue is only online); among these, there are two forms:
  - E-journals with *virtual issues*, that is submitted, refereed and accepted articles are collected and published ("put online") together at pre-set dates (similar to P-journals);
  - E-journals with *continued publication*, that is each submitted and refereed article is published as soon as it is accepted and formatted;
- *virtual journals* are E-journals which present online collections of relevant papers from a broad range of "source" journals in a particular field; the editor selects otherwise published articles, but does not accept new manuscript submissions;[261]
- *parallel P+E-journals*[262] which come in both formats; in general, the electronic version precedes the printed one.

An increasing number of traditional P-journals have been converted to P+E-journals over the recent years. Today, a newly founded academic journals typically starts as parallel journal (if the publisher is a commercial enterprise) or as an E-only journal (mainly if it is run by an academic association). There are only a few cases, so far, where a P-journal was converted directly into an E-only journal or where a commercial publisher started an E-only journal.

2.4.4.2 Tools for editors and publishers

Academic journals in paper traditionally used the letter mail and more recently the fax to administer all communication between the submitters, editors, referees, authors, readers and publishers. The Internet now provides for a number of electronic tools to take over.

ing, and has little or nothing to do with appearance; PDF is all about appearance, and has little or nothing to do with structure and meaning." (Kasdorf 1998, 3) More about SGML in Kasdorf (1998) who praises SGML for its potential for the publishing business in general, not only for E-publishing. He describes XML as the SGML for the web.

260 For a discussion of definitions and categories of E-journals see the overview of the (older) literature in McEldowney (1995, chapter II; Kling/McKim 1999).

261 E.g. the Virtual Journals Series in Science and Technology (<Cyberlink=748>).

262 Kling/McKim (1999) distinguish, in this category (which they call "hybrid journals"), further between "p-e journals" (paper journals with an online companion) and "e-p journals" (E-journals with a limited distribution in paper form). A few examples for the latter can be found in artificial intelligence research, such as JAIR (<Cyberlink=457>), which as a yearly bound hard copy. The first category is, however, by far more important as practically all previous P-journals have been uploaded to the Web by now.

The simplest way is to use E-mail for most or all mailings, including submission and subsequent handling of the article manuscripts themselves which can be attached to the E-mails. Indeed, many journal editors accept E-mail submissions and correspond with referees by E-mail. There are, however, much more sophisticated systems that also handle the storage of file and keep track of referees' reports, the different versions of manuscripts etc.

For instance, Appel describes an E-mail-based system applied for a computer science journal (1996). Wheary et al. present the sophisticated software which has been developed for the "Living Reviews in Relativity" (1998). As much of the article processing as possible is automated with a set of PERL scripts to be addressed through a JAVA graphical user interface. Given the innovative features of the journal, such as a reference database which not only stores the bibliographic reference but also the exact location of quotes in all review articles, this would be a tremendous task to do by hand. Pope/Miller (1998) describe the very useful, web-based software used by the editorial team of "Conservation Ecology". Everything possible is being automated, from uploading articles to communication with referees and formatting of articles.

There are also a number of initiatives with a view to making these insular solutions available for editors of other E-journals, too. For instance, EPRESS[263] is a system originally developed for a sociological journal. Unfortunately, lack of funding prevented the developers from making an alpha version publicly available (for a technical description, see Zhang 1997, §11ff.). Another such tool, ESPERE[264], is being developed in the framework of the UK Electronic Libraries Programme (eLib) and so far only available for the members of the developing consortium. While EPRESS and ESPERE support traditional peer review models with anonymous referees sending in their opinions on which the editors base their decision, the Digital Document Discourse Environment (D^3E)[265] software supports open peer commentary: For each article, a discussion list and a special homepage is set up. All contributions (comments, criticisms, amendments etc.) are linked to the specific paragraph in the original text, submitted for discussion. The software package includes sophisticated tools for the editors to manage this open discussion and consolidate and thread the comments (for details, see Sumner/Shum 1997).

For the publishers, there are also a number of tools available, in particular content management systems (CMS) (Kartchner 1998). Such systems administer the data repository, provide for workflow schemes, editorial tools and output utilities to help the publishers in a digital environment (see also below 2.4.6).

2.4.4.3 Tools for authors

Nearly all academic authors today use word processing software to write their manuscripts. While this is not the place to report on them in detail, I shall, nevertheless, make some general comments. First, there is commercial software (like for instance Microsoft's WORD) and there is non-commercial software, for instance based on the LINUX system. Commercial software has, so far, lots of functions for almost every purpose. Free software is, in general, less sophisticated, but this is changing fast. Another advantage of commercial software is its widespread distribution. Being "compatible" with fellow re-

[263] <Cyberlink=168>.
[264] <Cyberlink=164>.
[265] <Cyberlink=57>.

searchers, that is being able to exchange files without the trouble of converting them to different formats, often requires using the same software as everyone else in the closer research community. Free software requires more technical knowledge but is known to reward the user with a more stable performance. In addition, compatibility with proprietary file formats is high on the present agenda of open-source developers.

Authoring tools for web publishing (HTML pages) are equally available. However, if my scenario that hypertexts may play a more important role in academic writing (see 6.2.2.1) becomes reality, easy-to-use hypertext editors will be needed. DISKURS[266], the software used to write an experimental conference paper-hypertext in the framework of this study (Nentwich, 2000; see also above in 0.3.4.2), is the attempt to enable the academic author to write hypertexts without having any particular knowledge of hypertext marking. It is just one step with a view to explore the potential. A good editor is essential since all the technical issues involved in making a hypertext distract from the main task, i.e. writing and thinking. To my knowledge such editors specifically targeted at academic writing do not exist yet (despite the fact that the WWW has been around for a couple of years already)[267], but will certainly be developed. They will help organise a hypertext by assisting in making the right type of links and modules. Furthermore, these tools will also do the meta-tagging and storing in the right place of the net environment. These tools will help the authors to create "meaningful links" (see 6.2.3.2) semi-automatically.[268]

2.4.4.4 Print-on-demand and electronic document delivery

Circulation of publications in academia is often rather limited. Therefore, print-on-demand (or publishing-on-demand, both abbreviated as PoD) may be a solution. Note that this is not about E-publishing, whereby documents are stored on web-accessible servers and printed ("on demand") on the personal printers of the customer, but about print (P-) publishing.[269] The difference to traditional P-publishing is that documents, in particular books or reports, are printed by the publisher but only when they are requested, and only the quantity requested. The printing is done by professional printing presses that are specially constructed for fast runs with low copy numbers.[270] PoD customers order their personal copy directly with the publisher and receive it by mail. Bennett calls this "just-in-time scholarly monographs" (1998).

Furthermore, this allows for user customisation and for personalisation of documents. Not each copy of a book or report has to have the same number of chapters, the same layout (e.g. font size), or the same page size. For instance, there is the idea to allow students to select which chapters of a textbook they want to buy. Using again a metaphor from industry, this would then be "mass customisation" in scholarly publishing. From the point of view of the publishers, this allows for one-to-one marketing and for a move from print-and-distribute to distribute-and-print modes of doing business (Interquest 1997, 7).

[266] <Cyberlink=32>.

[267] See however HYPERCARD, STORYSPACE, TOOLBOOK and other software that has been used to write fiction hypertexts (e.g. Becker 1995).

[268] DISKURS only creates vice-versa-links between modules. It is not yet able to create typified links other than path links.

[269] It is conceivable to combine E-publishing with PoD: customers could either download the digital version and print it on their own, or they order a nicely bound printed version.

[270] For instance, DOCUTECH is the Xerox trademark of a printing appliance particularly designed for print-on-demand (<Cyberlink=754>).

There are already a number of PoD publishers targeting academics. Often they offer not only PoD, but at the same time E-book publishing.[271]

A similar service is electronic document delivery, organised both among libraries[272] as well as on a commercial basis. Similar to the traditional system of interlending (interlibrary loan) of printed matter, the idea is that libraries having no access to a particular journal (electronic or print) may use this service to receive individual articles as E-mail attachments. Printed matter is scanned and sent, for instance, as a PDF file.

2.4.5 E-conferencing

E-conferencing software enables discourse among non-present researchers. With a view to its function, we can distinguish between three types:

- *bilateral* type: enables one-to-one conversation;
- *lecture* type: enables one-to-many broadcasting of speeches ("webcasting"); and
- *seminar or workshop* type: enables group discussion (few-to-few).

Note that a seminar E-conferencing tool can certainly also be used for bilateral discussions or lectures, but that some tools are not sophisticated enough to enable effective group discussions. Furthermore, the array of media included in the communication may differ. E-conferencing may include an audio channel and/or a video channel, or may be text-based. In addition, the sharing of documents (e.g. presentation slides – see below) may be possible. There are different technological possibilities for E-conferencing:

- *Telephone conferencing* which requires access to a bundle of parallel phone lines; it only provides for audio transmission and no special equipment is necessary.[273]
- *Videoconferencing*, that is with video cameras, special screens and a bundle of dedicated phone lines (or other data connections). There are dedicated video conferencing rooms in office centres, but sometimes also in universities. Both the equipment and the actual holding of a conference are rather costly and require that all participants have access to this equipment).
- *Internet* or *web conferencing*: here, each participant needs an Internet connection, conferencing software installed (see below) and a headset (i.e. a microphone and headphones). If, in addition, video should also be transmitted, a web-cam and a fast Internet connection are required. The live video pictures of each participant show on the computer screen. All this equipment is rather cheap. Internet connections are, in general, increasingly fast in academic institutions so that the potential net of communication partners is becoming larger.
- *Conference E-lists*: participants of the E-conference register with an E-list (of whatever type, see above 2.4.2) which is open for submissions of contributions for a limited time (from a day to weeks).[274] Normally, a moderator – like a panel chair – is engaged in initiating and steering the discussion, for instance by organising stimulating input

[271] For instance the German publisher BoD (<Cyberlink=492>).

[272] A German example is SUBITO (<Cyberlink=764>).

[273] This might also be done via Internet telephone providers, that is, one that does not use traditional phone lines, but rather the Internet, which is much cheaper.

[274] This is what the comprehensive Kovacs Directory (<Cyberlink=181>) means by "scholarly E-conferences". Typical examples were the so-called "think-tanks" of TechNet (<Cyberlink=122>) whose archives are still available online.

contributions (cf. Mills 1998). In this case no audio or video, but only text contributions are possible. While participants may react to each other's contributions immediately, this type of E-conferencing does not require synchronous communication.[275]

- *E-chatting* or *instant messaging* may also be used for academic conferencing;[276] it is again text-based[277], but in contrast to E-lists a synchronous medium, that is, it requires that all communication partners are online at the same time. Each participant has two windows on his/her screen: in the first, you can see what all other participants have just written; in the other, you write your own messages which appear in the first window shortly after you have hit the "submit" button on the last line, probably immediately followed by a contribution of another participant. In general, no record of such a "written session" is kept, but it is possible to do so. Special chat channels with academic topics are rare, but may have a potential (Orthmann/Näcke 1999, 4).

The following table summarises the key features of different types of E-conferencing as discussed above (including also bilateral distance conferencing/communication):

Table 2-4: Features of E-conferencing tools

	Media channels				Participants			Time
	Text	Audio	Video	Sharing	Bilateral	Seminar	Lecture	Synchronous
E-lists	x					x		
Chatting	x				x	x		x
Internet conference	x	x	x	x	x	x	x	x
Video conference		x	x		x	x	x	x
Telephone conference		x			(x)	x		x

As can be seen from the above table, "Internet conferencing" is both multi-channel and most flexible as to the number of participants. Compared to (not Internet-based) videoconferencing, the quality of Internet conferencing is, on the one hand, (still) lower. Video transmission is of lesser quality as the network is not able to transmit in "real time" that is without delay and seamlessly. On the other hand, it is much easier to set up a conference (you can do it from your desktop PC) and considerably less expensive. Start-

[275] For an early example see Freeman (1984, 202) who describes the "Electronic Information Exchange System (EIES)" which helped the emerging community of social network scholars in 1977 to meet and exchange in cyberspace with a message system. Another early system allowing both to set up related E-mail discussion lists and up- and downloading files was "PARTICIPATE" as described by Harasim/Winkelmans (1990). Woolley (1998) discusses the major design and organisational challenges of "web conferencing" by which he means E-mail conferencing in its various forms (E-lists).

[276] One example for a chatting facility in the Internet is ICQ (<Cyberlink=6>).

[277] Note that there is now a software available which adds voice to ICQ: QTALKA (<Cyberlink=7>); furthermore, most Internet conferencing tools for audio and/or E-conferences also provide for parallel chatting.

ing with the MBONE[278] initiative in the mid-90s (Grötschel/Lügger 1996, 14) originally based on video cameras, a variety of tools have been developed and implemented for Internet conferencing based on the new webcam technology. For instance, PLACEWARE[279] is a web-based tool also using telephone lines to connect the participants and "simulates a virtual lecture hall" (Finholt 2001, 23). Another widely used software is Microsoft's NETMEETING[280] (whose strengths and weaknesses are discussed by Finholt 2001, 22).[281] NETMEETING allows both for bilateral as well as group conferences and provides, in addition to audio and video transmission, for programme sharing.[282] Within the community of high-energy physicists, Virtual Room Videoconferencing System (VRVS)[283] has been developed with a view to integrating all different platforms and systems for E-conferencing.

Programme or desktop sharing is a unique feature of Internet conferencing software. Participants in the cnference can allow a live picture of either their whole desktop or of a particular window to be transmitted to the other participants. By this token, the participants are in a similar position as the auditorium in a face-to-face situation. They can see, for instance, presentation slides or the text of a document that is being edited during the debate. In a bilateral situation, it may even be useful to let the remote communication partner take over control of a window (and programme) on one's own computer with the idea of collaborating on a particular file simultaneously. Some programmes also allow for a combination of Internet conferencing (audio and/or video) and a parallel chatting session.

Seminar-type E-conferences with only a handful of participants are easily set up by individually opening up connections to all people. Larger conferences need a central server to administer the conference. CENTRANOW[284], for instance, offers an online environment for E-conferences without video and for up to five attendees, it is a free service. VRVS (see above) is a similar, but non-commercial platform that provides for so-called "reflectors" (i.e. specialised communication servers) which connect each user to a "virtual room".

2.4.6 Content management systems

A content management system (CMS) is a tool that enables centralised technical and decentralised non technical staff to create, edit, manage and publish a variety of content (such as text, graphics, video etc). There is a set of rules, processes and workflows that ensure a coherent, validated appearance of the output. The latter is often a website but may also be other kinds of publications. In a nutshell, CMS are collaborative tools for sharing information with a view to facilitating the co-operative production of knowledge representations.

[278] MBONE stands for "multicast backbone" (<Cyberlink=288>); in Germany, MBONE is now called DFN MULTICAST (cf. <Cyberlink=732>).
[279] <Cyberlink=733>.
[280] <Cyberlink=725>.
[281] Finholt's list of weaknesses of the use of a tool like NETMEETING: "over dependence on a small number of technically savvy users" and "awkward organization of conversational turn taking" as well as "high overhead associated with initiating data conferences" and "the importance of highly motivated NetMeeting 'champions' in getting groups over initial learning curves".
[282] Further examples are to be found in <Cybercategory=9>
[283] <Cyberlink=681>.
[284] <Cyberlink=1>.

Various CMS systems mushroom in the business world, but often originate from academic projects, such as the HYPERWAVE eKnowledge Portal or Information Server[285]. In academia, publishers are increasingly using such systems with a view to manage their large E-publishing projects. Another potential field of application in the academic realm is the virtual university, which needs to organise digital material for its students. Furthermore, it is conceivable that in large collaborative research projects or research institutes, CMS helps to optimise the research processes. Here, the dividing line between CMS and groupware is fading (see next sub-section).

2.4.7 Groupware

Groupware is software for computer supported collaborative work (CSCW). We may distinguish the following forms of groupware:

- *E-mail-based collaboration*:[286] This may be either on a (multi-)*bi*lateral basis or multilateral via E-lists. In its most basic form, scholars may engage in collaborative sequential text production with the help of present-day word processing software allowing for versioning, commenting etc. while sending drafts back-and-forth via E-mail.
  Valkenburg notes that "it has become increasingly clear that e-mail and e-mail lists lack functionality in a number of areas, even when supplemented with Web pages" (1998). For instance, versioning control and file sharing (that is managed access to the latest version of a document) cannot be done in a simple environment. Rost addressed this problem with his more sophisticated ObM – ("Organisation by Mail") server (1998c). Here, files are sent to and retrieved from the server via E-mail and administered there with versioning control (revision control system – RCS).
- *Asynchronous shared workspaces*, i.e. virtual spaces in which participants may up- and download documents, communicate and schedule meetings etc. These can be configured in a local network (i.e. not web-based; examples are Lotus NOTES or Netscape's COLLABRA[287]) or web-based like the free services BSCW[288] and WEB4GROUPS[289] (compared by Valkenburg 1998).
- *Asynchronous shared text production:* a few CSCW tools are specialised in online add-on-writing of texts via a web browser. In TWIKI[290], for instance, users edit text by typing in new text in a web form while using a very simple, but powerful markup language for layout called TWIKISHORTHAND; a special applet even allows drawing and editing graphics online. In OPENTHEORY[291], users add comments to a text, again via a web form (but with only limited formatting options), which then show as indented text below the commented paragraph. ENOTE[292], the electronic notebook, offers the members of a collaborative research group a common notebook on the WWW. With a view to writing web documents collaboratively, "Web Distributed Authoring and Version-

[285] <Cyberlink=44>.
[286] Note, however, that I stretch the notion of groupware when including E-mail-based collaboration.
[287] <Cyberlink=334>.
[288] <Cyberlink=734>.
[289] <Cyberlink=82>.
[290] <Cyberlink=184>.
[291] <Cyberlink=50>.
[292] <Cyberlink=752>.

ing (WebDAV)"[293] is being developed by working groups of the Internet Engineering Task Force (Burg et al. 2000, 4). Also MOO/MUD can be used for shared text production (Meyer et al. 1994), i.e. not web- but telnet-based collaborative spaces originating from the world of games (see below in 2.4.8).[294]

- *Synchronous conferencing tools* with shared-desktop facilities, where participants may all see and manipulate an area of their screen, e.g. for drawing or writing text (see already above 2.4.4.4). A highly sophisticated example is MINDMANAGER[295], graphical brainstorming software allowing for mind maps to be elaborated in a distributed setting.

The whole area of CSCW or groupware is highly dynamic, less because of important demand from academia (see 3.3.9) but because there is a growing market in the economy. Therefore, some of these systems also include so-called group decision support systems whereby voting and rating is supported (e.g. Alton-Scheidl et al. 1997). Valkenburg (1998) lists the following important yardsticks for assessing group collaboration tools: provision of private and public message boards; document sharing and versioning; integration of other existing means of (a)synchronous communication; integration of the interface with the desktop; security and access control to materials; voting and rating functions; scheduling features. Finholt (1997, 30) speaks of "awareness tools"[296] which are being developed to "reproduce in a distributed network environment the social cues and information that are normally available only in a shared physical setting" (e.g. "closed doors").

2.4.8 E-teaching tools

Although outside the main focus of this research-oriented study,[297] the new teaching tools are nevertheless discussed here since the majority of researchers also teach. A number of tools already presented in the previous chapters play a role in teaching, above all:[298]

- *E-mail* can and is increasingly used for the correspondence between teachers and students.
- *E-lists* serve well for questions and answers among seminar groups, including the teacher, and for group discussion.
- *E-textbooks* come in two flavours: either as a printed book with CD-ROM or WWW enhancement (for tests, multimedia such as animated graphics etc.) or online-only.[299]

[293] <Cyberlink=470>.

[294] An example of this kind is BioMOO, "a virtual meeting place for biologists (...) a place to come meet colleagues in Biology studies and related fields and brainstorm, to hold colloquia and conferences, to explore the serious side of this new medium" <Cyberlink=347>.

[295] <Cyberlink=727>.

[296] Not to be confused with those awareness tools that help web users keep track of changes on a list of websites or individual webpages.

[297] See, however, the overview in OECD (1998, 215ff.).

[298] "The tools for alternatives could be video servers with stored lectures by outstanding scholars, electronic access to interactive reading materials and study exercises, electronic interactivity with faculty and teaching assistants, hypertextbooks and new forms of experiencing knowledge, video and computer conferencing, and language translation programs." (Noam 1995, 248)

[299] Christie (1998) believes that, in the long run, it will be more cost-effective to lease or buy portable display devices for all students and to provide them with curriculum materials on CD-ROMs or through the Internet than to continue buying printed textbooks. In 1998 already, the Texas Board of Education launched a programme with a view to replace traditional books with E-books.

Highly sophisticated "web-based training" sites go well beyond the traditional concept of a textbook. All this comes under the label of "courseware", that is the combination of course (teaching) material and software to present it to the students in an innovative manner.

- *Homepages* of the instructor or university department are well suited to host all kind of course material for downloading such as syllabi, articles in the students' reserve and further documents.
- *E-conferencing* tools are, finally, the most advanced cyber-application for E-teaching. Two of the variants (lecture and seminar types) discussed above play a role here: *tele teaching* is the synchronous transmission of a lecture to remote locations; and *virtual classrooms* provide for face-to-face like situations in smaller groups (up to perhaps 15 students in a class). In the latter case, the whole array of features discussed above may be displayed, such as application sharing, text-based chatting and simultaneous audio/video debate.

Under the label of "distributed teaching" a combination of some of the above tools (except E-conferencing) serves to organise asynchronous courses where each student may follow his/her own path and pace, at whatever location. The instructor is available via bilateral E-mail and in E-lists and feeds course material into his/her webpage. Exams are taken via a web interface. Eventually, students and teacher also meet in person, combining the strengths of both distributed and traditional teaching. Commercial software packages like WEBCT[300] or the open source alternatives like ILIAS[301] are increasingly widespread, in particular in the US, and are intended to organise such courses. They include both the administration of students, the setting up of E-lists, E-mail accounts, dedicated course webpages for downloading the material, scheduling devices etc.

Many universities today have their "tele-teaching" programmes and often dedicated support offices for university teachers willing to develop online courseware.[302] In the most basic form, there are probably only very few universities left with not a single course on the Web, at least in the form of online downloadable course material. A few universities are more advanced, such as the Open University in the UK[303], well known for its tradition in distance learning. There are already genuine virtual universities, which rely fully on ICT-based teaching and administration, such as Jones International University[304].

From a technical point of view, a variety of technologies are being used and tried out for tele-teaching and virtual classrooms. Kirkup/Jones (2000) compare the strengths and weaknesses of different teaching technologies: print, radio, audiocassette, educational broadcast TV, pre-recorded TV, videocassettes, computer-based learning, multimedia, audio-conferencing, live interactive TV, video-conferencing, computer-mediated conferencing.

A new alternative to audio-video-conferencing tools are MOO/MUD-based virtual environments called "Multi Academic User Domains" (MAUDs). Purely text-based MUDs and their object-oriented variant MOOs are virtual reality spaces, accessible usually via telnet which incorporate communications and other functions. "Although MUDs were

[300] <Cyberlink=447>.
[301] <Cyberlink=551>.
[302] To name but a few European examples: Teleteaching Mannheim (<Cyberlink=70>), VIRTUS Cologne (<Cyberlink=86>), Virtual University WU Vienna Living Lectures (<Cyberlink=331>), Projektzentrum Lehrentwicklung University of Vienna (<Cyberlink=649>).
[303] <Cyberlink=477>.
[304] <Cyberlink=507>.

originally developed in order to facilitate the playing of (...) games, these systems are now being developed for academic uses such as distance education and virtual conferencing. MAUDs are systems which are dedicated to these uses."[305] There are examples of MAUDs from philosophy, like "freiraum"[306] and Painted Porch[307]. The combination of E-mail, the WWW, eventually audio transmission and MAUDs (or similar environments) is also known under the name of Virtual Educational Environment (VEE).

In general, courseware is a booming market, just as textbooks. But not all of it is commercialised. For instance, MIT's OpenCourseWare[308] project has announced making available selected course websites in a variety of fields. A project with a community approach, at least in the beginning, but which is heading for commercialisation is PolitikON[309], an initiative in the German political science community. Here course modules and other teaching-related resources are currently being developed by the participating scholars and will be made available for all others in the group. There is also a special mechanism and set of rules for amending courses. Other initiatives dedicated to enable the teacher community to "go online" include MERLOT (Multimedia Educational Resource for Learning and Online Teaching)[310] which is a free and open resource designed primarily for faculty and students in higher education.

Note that the students themselves are also active in cyberspace. There are, for instance, sites for sharing seminar papers[311], for helping doctoral students[312] or simply exchange information[313]. Furthermore, with a view to detect plagiarism of students, software has been developed to compare students' essays with Internet resources.[314]

2.4.9 Translation tools on the web

One of the assets of the Internet is, at the same time, an obstacle to seamless communication: the WWW is multi-lingual. English does not seem to be the solution: Schlegel (1998) points out that the increase of non-English documents by far outweighs the increase of English documents. Much of the information in the WWW as well as in the various E-lists is not accessible for everyone. There some remedies available:

- Only few websites offer more than one language version of their pages.[315]
- Online dictionaries offer word-to-word translations. Little scripts ("bookmarklets") allow for a simple procedure to get instantaneous translation of a marked unknown word.[316]
- Increasingly, there are online translation devices available which offer a crude full translation of a webpage text in one's own language.[317] Machine translation is a vi-

[305] <Cyberlink=735>.
[306] <Cyberlink=696>.
[307] Often cited, but seemingly not available any more at telnet://maud.cariboo.bc.ca:4000.
[308] <Cyberlink=476>.
[309] <Cyberlink=552>.
[310] <Cyberlink=94>.
[311] Hausarbeiten.de (<Cyberlink=503>).
[312] Doctoral Students (<Cyberlink=563>).
[313] Mnemopol (<Cyberlink=502>).
[314] E.g. Plagiarism.org (<Cyberlink=811>).
[315] HYPERWAVE is supporting multi-lingual websites (<Cyberlink=43>).
[316] LEO is perhaps the most frequently used of these online dictionaries with over 1,600,000 queries per day (<Cyberlink=767>).

brant research and development area and might once reach a level of accuracy and quality satisfying even academic purposes (see e.g. STOA 1999).
- More sophisticated projects embed multilingual meta-descriptions in monolingual pages (e.g. in XML). By this token, specialised search-engines may retrieve pages of various languages on the basis of a common thesaurus.[318]

2.5 Archiving

Given the tendency to shift (perhaps) all scholarly output to cyberspace, the need for a sustainable solution to the problem of loss of digital data is pressing. Here I shall discuss the technical options and difficulties, leaving aside the organisational, financial and selection aspects (see 7.3.3).

Every cyberscientist is aware of the problem of vanishing resources. There is a variety of reasons for and aspects of this problem:
- *Moving*: As already discussed above (2.1.2) resources may not vanish altogether, but may get a new address.
- *Replacement*: The Web is dynamic, much content just disappears because it is replaced by newer versions.
- *Hardware*: Storage media and the corresponding hardware (disc drives and the like) become outdated or obsolete in the sense that new hardware does not support them any more. The classical example is the old 5 ¼" inch floppy disc which was in widespread use only a few years ago but has almost vanished today. Old archives on this out-of-date storage medium are almost useless today if not migrated to at least 3 ½" diskettes (which, in turn, are already almost out of date in the advent of zip drives and re-writable CDs). Raney rightly notes that the requirement of constantly updating the archive "effectively constrain, if they do not eliminate, the possibility that an individual researcher can maintain his or her own specialized library" (1998, 4).
- *Data-formats*: Software is evolving continuously and with it the formats in which the data (documents) are produced by the software. Up to a certain point, higher (later) versions of programmes can read files produced in lower (older) versions. In general, however, this does not go on forever and there will be ever more files that cannot be read any more by up-to-date software.
- *Operating systems*: Operating systems have also evolved similarly to application software. Older software may not be executable in later operating systems. Hence, the files (although readable by the software) cannot be read because the old software cannot be installed in the new environment.
- *Life span*: Storage media do not last forever. While magnetic data carrier and microfilm have a life time of only a couple of years, and CD-ROMs of perhaps 10-15 years, premium paper combined with good storage conditions have proven to have a lifetime of a couple of hundred years (Raney 1998, 3; see also Risak 2000, 19).

[317] For instance, the web search-engine GOOGLE offers online translation of the pages retrieved (so far, in three languages; <Cyberlink=760>).

[318] See, for instance, the Cross-Language Evaluation Forum (CLEF) which supports interlingual information retrieval in global digital library applications (<Cyberlink=874>).

- A special problem is the *integrity* of the stored file. We can distinguish between the integrity of appearance, integrity of content of the information and internal integrity and coherence (e.g. if modules are not stored locally, but in databases) (Kircz 2001, 3). Related problems provide multimedia applications that often consist of various elements scattered around the Internet. The archiving process would need to secure that all elements are stored together.

A reliable and convincing solution (or set of solutions) for the archiving of the digital output of cyberscience is essential. Many are reluctant – and rightly so – to entrust their works to a system which would not be able to guarantee its permanence. The following technical solutions are being discussed:

- *Mirror sites* and other forms of *back-ups*: Backup copies (which can replace the current copies in case of damage) and mirror sites (that is additional servers at different places with an identical configuration and storage) are being advocated. This is mainly a solution in the framework of data security that is for preserving the current versions in a functioning environment. Ginsparg speaks of "a global backup system resistant to localized database corruption and/or loss of network connectivity" (1996, 7).
- *Good old paper*: While not questioning E-publishing as a whole, many argue that the archiving problem cannot be solved completely with digital means. Hence, long-lived paper copies at more than one site (but not hundreds) should be stored (Varmus/et al. 1999). For instance, each yearly volume of the E-journal JAIR[319] is published in hardcopy by a commercial publisher (Wellman/Minton 1998, 8) in order to relieve concern about the permanence of the journal.
- *Static archival storage*: The data could be stored in very durable material. For instance, the Long Now Project[320] is planning to engrave the data in a microscopic scripture on silicon disks, hence the data remain readable for the human eye (with a microscope).
- *Dynamic storage*: The data is being continuously copied from old data carriers to new data carriers. By this token, you can at the same time both update to later software versions and to longer lasting state-of-the-art storage media. However, even digital copying is a tricky business: there is a (very) little, but nevertheless realistic chance that once in a while small copying errors occur. With no additional (and time consuming and hence expensive) re-check, such errors slip the attention of the storing institution and may accumulate over time in a way that the data is not readable any longer.
- *Non-proprietary data formats*: The use of Unicode (an extension of ASCII) and standardised mark-up languages (in particular of SGML or XML) should solve the problems produced by storing data in the many proprietary formats, codes and tags. Equally, the usage of different languages, national special characters as well as the intermingling of content, structure and presentation/layout could be managed (Risak 2000, 22).
- *Software emulation*: The presentation format can not be stored without reference to the software used, in particular with regard to video sequences etc. One possible way out of this dilemma is the project to store, next to the information items themselves, also the software programs used, including the operating systems. The plan is to emu-

[319] <Cyberlink=457>.
[320] <Cyberlink=304>.

late[321] the soft- and hardware on future computers with a view to restore the same look and feel as in the original (Kircz 2001, 6, referring to Rothenberg).[322] The latest development in this respect is the idea of a Universal Virtual Computer (UVC), currently developed in the IBM laboratories.[323]

- *Periodic Web-harvesting*: Documents will be harvested at certain time intervals to produce snapshots of the entire web space (or of specific parts thereof).[324] The idea is that you can make a sort of time-travel and browse through historic web spaces as if it was today. This mainly solves the problem of replacement; finding particular documents is, however, as difficult in the historic as in the present web space. A similar route is based on the data that are "spidered" and indexed by search-engines. If stored, these data can also be used to access old versions of web documents.[325]
- *Dedicated archiving servers*: Documents worth being archived could be sent to dedicated archiving servers (central or decentral, but interconnected[326]) which take over the responsibility to secure continuous access.

2.6 Outlook

As we have seen, large parts of the working environment of researchers have already shifted, or are about to shift, to the virtual world. Many scholarly activities are, at least in principle, apt to be performed in this new environment. In what form this cyber-environment will be of importance for academics is above all a matter of task suitability and of organisational and cultural factors (as will be discussed in later chapters). In this technically oriented chapter, I will conclude with a few observations about the crucial technical factors.

Despite many promises it can hardly be denied that there is also a considerable downside of computer use. As mentioned above, both hardware and software are (still?) not very reliable. While the computer chips in many appliances of our daily work (such as cars, television sets or telephones) rarely fail to work (if at all), it is quite likely that an average computer user experiences "crashes" in various forms. Programmes "freeze", computers need to be "re-booted" (that is restarted), data get lost (often unnoticed at first with the consequence that it becomes ever more difficult to restore), servers are "down", so that their services (access to data, software) are not available when needed, etc. This seems to be the price for ever more sophisticated tools being developed rapidly: intensive testing and increasing reliability often come last in the considerations of companies dominating the market.

[321] This means that the programmers write special software that provides for an "environment" in whatever hardware old programmes can still be loaded. Emulation means that you are able to let an old ATARI programme run on a state-of-the art PENTIUM PC.

[322] The Rothenberg report on emulation is to be found here: <Cyberlink=463>.

[323] <Cyberlink=831>.

[324] E.g. the Austrian project AOLA, based on Scandinavian and US concepts, <Cyberlink=439>.

[325] The most prominent solution is the Internet Archive since 1996: its so-called WAYBACKMACHINE allows users to browse through older versions of webpages (<Cyberlink=779>).

[326] The Open Archives Initiative (OAI) aims at securing interoperability of such archives (<Cyberlink=60>).

The continuous flow of new software versions and, hence, the constant need to upgrade and to learn once again how the new release functions is a second serious practical problem. Every computer user has already been forced (probably not only once) to search for previously often used buttons, features, procedures etc. in the new version of a software. We have to constantly relearn the differences between old – reasonably well functioning – systems and newer versions – that have been increasingly sophisticated and complex. Certainly, the continuous improvement of user interfaces (i.e. the "surfaces" of any software, the pre-set ways to interact with it, be it via keyboards, mouse-clicks or voice-control) is a noble goal. It is also important in the sense that we cannot expect cumbersome software to be used regularly *if* there are (non-cyber) alternatives. In addition, the attractiveness of the newly added features may be key in convincing people to shift from traditional to cyber-ways of doing research. However, the need to continuously adapt and learn will probably persist and not everyone welcomes this (in 5.1 and 11.2.3.2 I shall specifically address this constant need to learn).

These two problems (lack of reliability and software dynamics) account for the increasing multiple dependencies on functioning hard- and software. This can bring about inefficiencies and loss of precious research time. An optimistic view highlights that the technical problems will all be solved soon. One argument is that incremental changes in scholarly communication produce new demands and feed back to the development of the technology (cf. 1.2.3.4) so that the limiting technical factors will lose importance over time. While there is certainly some truth in it, the probably more realistic pessimistic view is that there always will, at the same time, be new technical problems. The level of problems may persist as we are dealing with highly complex systems. Hence, the level of inefficiencies in the use of the sophisticated new tools may not decrease, after all. Only if the gain in efficiency in those times where the soft- and hardware works well is large enough to balance any losses from short-term malfunctioning can we, in total, speak of efficiency gains. The balance sheet may look different in various aspects of cyberscience. It seems, however, unthinkable (and there is no author supporting) that a negative balance should result in all or many dimensions of scholarly cyber-work. (I shall come back to this issue when I discuss possible gains in research efficiency and productivity in 4.3.2.)