Journal of Universal Computer Science, vol. 2, no. 9 (1996), 650-658
submitted: 19/9/96, accepted: 19/9/96, appeared: 28/9/96 © Springer Pub. Co.

Professional Electronic Publishing in Hyper-G[1]

The Next Generation Publishing Solution on the Web

Klaus Schmaranz
(Graz University of Technology, Austria
kschmar@iicm.tu-graz.ac.at)

Abstract: The first part of the paper identifies disadvantages of first generation Web publishing solutions that have to be overcome for professional publication providers. Using Hyper-G for distribution of electronic documents opens the way to the first fully-integrated professional publishing solution on the Web. User and group access rights as well as billing mechanisms are integrated into the server, links are bidirectional and annotations to existing documents are possible without changing the contents of documents themselves. Hyperlinks are supported in arbitrary document types besides hypertext which opens the way to new publishing paradigms beyond the paper-based approach. Naturally, a rich set of structure and search mechanisms is provided. On this basis, a set of tools has been developed that almost fully automatically supports electronic refereeing, automatic hyperlink creation, glossary links and table of contents generation. All the data prepared on a Hyper-G server can then be simply exported to CD-ROM, allowing hybrid Web/CD-ROM publications to be created without any additional effort.

Key Words: electronic publishing, WWW, Hyper-G, HyperWave, automatic hyperlink creation, electronic refereeing

1 Motivation

Electronic publishing is today one of the booming branches on the Web. Web surfers will find lots of electronic paperware on their ride through the servers. However, most of the Web sites that serve electronic publications are run by universities and only a few are operated by publishing companies on an evaluation basis free of charge.

Having a closer look at the way electronic publishing is done today, one will mostly find HTML or PDF documents that are very similar to their paper-based counterparts. Often a search engine is provided to make location of interesting papers easier, but other benefits of publishing electronically are mostly neglected. For the reader of electronic publications, nearly no value is added compared to paper-based articles. Worse than that - considering HTML documents - the possibility of making high quality printouts for archival purposes is lost. This is surely not enough to make electronic publishing on the Web a success.

Distributing electronic documents on the Web is considered to be cheaper than publishing on paper [see Odlyzko 95], additionally turnaround time from submission of an article to the final published version is considered to be significantly shorter than for paper-based publications. As experience with J.UCS [see Section 5] shows, this is certainly not true for high quality publications that

[1] This paper was also presented at *WebNet '96: The First World Conference of the Web Society*, San Francisco, CA, October 1996

are refereed. In this case refereeing, corrections by the authors and resubmission are the most time-consuming stages of document preparation. Once the paper is ready for publication, hyperlinks have to be inserted in the electronic version, which is still done by hand in first generation systems because appropriate tools are missing. Both refereeing and hyperlink creation can be sped up when using the right tools as can be seen later.

Hyperlinks in first generation Web systems are represented by URLs (Uniform Resource Locators) [see Berners-Lee 94] that are unidirectional and embedded in documents. This approach implies that links can easily point to nowhere (the "this document has disappeared syndrome"). Using URNs (Uniform Resource Names) instead of URLs is the solution to this problem. URNs no longer point to the physical location of a document on a server, but add another level of indirection by defining unique location independent names for documents.

Embedding hyperlinks in documents has the disadvantage that only document formats designed to be hyperlinkable, such as HTML, allow navigation through cyberspace. Unfortunately, a lot of formats used for different document types were designed long before anybody knew about the Web. This includes all image formats like TIFF, GIF, JPEG, [see Murray 94], it also includes all video formats such as MPEG as well as one of the most widespread formats for professional publishing: PostScript [see Adobe 90]. The way out of this dilemma is to have a separate link database and consider links to be overlays. This automatically makes all document formats hyperlinkable because documents and links are handled separately.

Structuring and classification of documents in standard first generation Web servers has to be hand-coded using HTML documents. For this reason, most servers have poor structure and no document classification because it is too much work to provide it. This makes navigation rather difficult [see Andrews 95], because after a having followed a few hyperlinks, the readers no longer know where they are (the "lost in hyperspace" syndrome). Once a reader has found an interesting paper, it is absolutely necessary to immediately store a bookmark. If the way to the paper was longer than around 3 mouseclicks it is nearly impossible to remember the way to find it the next time it is needed.

One of the really severe problems of electronic publishing today is the lack of accounting and billing mechanisms as well as user access rights [see Maurer 94]. For this reason, electronic publications on the Web are mostly free of charge which is not the way publishing companies operate.

Very often it would be interesting for server operators to have statistical information on a user session basis to be able to follow users on their way through a server and optimize their offer. As an example, one could find out that 60% of the readers from overseas are leaving the server after having followed 4 hyperlinks. Having a closer look it could turn out that the document after the 4th hyperlink contains a huge inline image that is unacceptable for an overseas data transmission and that this document has to be changed.

2 Hyper-G as a Basis for Electronic Publishing

All the problems and difficulties mentioned above have lead to the development of a Web server with a completely new underlying concept compared to first generation Web systems: Hyper-G [Maurer 96]. Hyper-G now really allows professional electronic publishing to be done on the Web and enters a new era of electronic publications in terms of usability and content.

Up to now, electronic publishing has been done either on the Web or on CD-ROM basis. Using Hyper-G one can create hybrid Web and CD-ROM publication without additional effort as has been successfully proven over the last two years with J.UCS, the Journal of Universal Computer Science by Springer Pub. Co. [see Maurer 94]. The only thing publishers have to do to produce a CD-ROM is to prepare the data on a Hyper-G server and then export the whole collection tree or parts of it.

The kernel of the Hyper-G server is an object-oriented distributed network database with a separate link database. Information structure as well as document meta-information are a basic part of the concept [see Kappe 91]. This makes it possible to present the user with a seamless world-wide structured information space across server boundaries.

Document meta-information such as author, title, keywords, creation date, modification date as well as expiry date and many more support the Web surfer in getting as much information as possible. Naturally, document meta-information is searchable in addition to the fulltext search facility. The scope of searches is user definable and can be one small part of one server or even the whole content of all servers worldwide in one single operation. Even when performing searches on multiple servers it is not necessary to know about the server addresses.

More than that: meta-information cannot only be applied to documents but also to hyperlinks! This means that links can have types, such as annotation links, inline links, also version links for documents where multiple versions exist and many more.

Hyper-G servers do not only provide read access but write access is also possible. Read and write access to documents are controlled on a user and group access rights basis and billing is integrated into the server.

This concept opens completely new perspectives of electronic publishing: having the hyperlinks in a separate link database makes every document hyperlinkable even when the document format does not allow links [see Maurer 96]. All links in Hyper-G are bidirectional, making it possible to not only follow the links pointing from one document to another but also to see links referencing a document and follow them in reverse direction. Being able to examine the neighbourhood of a paper makes it possible to find other interesting papers on the same topic that very likely are difficult if not impossible to locate if only unidirectional links were possible as is the case with first generation Web servers.

Hyperlinks in arbitrary document types such as PostScript, images, movies, 3D scenes and even sound make navigation in hyperspace easy and a structured hierarchical view of the database with location feedback helps overcoming the "lost in hyperspace" syndrome.

With this approach, electronic publications need no longer be text based with some multimedia add-ons. Instead authors can choose the document type most suitable for the topic without loosing important hypernavigation features.

As an example, a paper about new chemical structures could consist of 3D models of molecules that are clickable. The hyperlinks could then lead to spectrum images that are then linked to some additional text based explanations in e.g. PDF [see Adobe 93]. A video of an experiment, naturally again with hyperlinks to explanations, could complete the presentation.

All the documents in the example above would carry meta-information such as keywords and therefore could be easily located in a search.

Acceptance of electronic publications is highly dependent on their quality. For electronic publishing, quality does not only mean high quality contents, which can be assured by an appropriate refereeing process. Stability of electronic publications is at least as important. Technically it is easy to change electronic papers after publication but this is unacceptable. Instead, Hyper-G's annotation and versioning mechanisms can be used to alert the reader of new results or errata. In this case the paper is not changed at all, only additional information is added to the paper. Therefore all citations of the paper that existed for the original version are still valid and the reader can choose to browse annotations and newer versions of the paper on demand.

Annotations in Hyper-G are hyperlinks pointing to the document that is annotated. Since Hyper-G's links are bidirectional the reader simply follows an annotation link backwards to read the annotation. The use of URNs in a link database instead of URLs embedded in documents guarantees that the annotation links are stable. This means that an annotated document can be moved around on the server or even from one server to another without generating annotations that point to nowhere. All links that pointed to the document before are then pointing to the document at its new location.

As has been mentioned above, refereeing is one of the very time consuming processes of publishing. Using annotations, electronic refereeing can be performed easily: papers are inserted in the Hyper-G server with read access only for the referees. The referees then comment the papers using the annotation mechanism. If desired, annotations can also be made readable for the author, so the author is able to react immediately on the referees' comments. More than that - the author himself could also annotate the referees' comments to clarify misunderstandings. Of course the author as well as the referees remain anonymous [Maurer 95]. Saving the time taken sending documents back and forth between referees, authors and editors, as well as being able to do corrections in papers while refereeing is still in progress can shorten the refereeing process significantly.

Amongst other structuring elements, Hyper-G supports the concept of clusters. A cluster contains several documents that are related to each other and therefore should be viewed together. In the example given earlier in this paper, the 3D molecular models could be clustered together with an explanatory text. In this case the user would get the 3D model in one window together with the explanatory text in another window.

Clusters are also used to serve multilingual documents. Documents in different languages are put together and readers then only get the documents matching their language preferences. In first generation Web systems the only possibile way to have multilingual documents was to let the user choose the language on the entry page and then follow different paths through the server for different languages. This approach caused a lot of work for server operators and readers had no chance to change the language on the fly. With Hyper-G only one path through the server has to be maintained and the readers can switch back and forth between multiple languages whenever they like.

3 Professional Tools for Publishers

So far the discussion has been about the technical design of Hyper-G and the resulting possibilities arising when using it for distribution of electronic documents. But that's not really all: these features can also be helpful for internal purposes such as collecting all the versions of a paper from the original submission over intermediate corrected versions to the final published paper. The complete refereeing process including all different versions of papers together with the referees' comments can be kept on the same server that is used for distribution. Supported by Hyper-G's access control mechanisms, the publisher defines what a reader can see. As an example, subscribers would only see the published versions of papers, while referees would also see intermediate versions of the paper they are currently refereeing. The editor in chief and the responsible staff would have access to all the information including refereeing forms and internal information. Referees would not have the rights to change a paper they are refereeing, but they would have the rights to make annotations. If desiredi, the author could get the right to also annotate the referees' annotations to clarify misunderstandings - naturally both referees and author remaining anonymous.

For the whole process from administration to final document preparation, a set of tools and fill-out forms has been developed that can be adopted according to the publishers' needs. The following paragraphs deal with some of the more sophisticated tools that have been developed to provide a cost effective way for publishers to add value to electronic papers.

As mentioned above, it is desirable to keep older versions of documents for archival and maintenance purposes. For this reason a special versioning tool was developed that uses Hyper-G's cluster mechanism: different versions of a document are clustered together and the reader can switch back and forth between different versions on the fly. Whenever a new version of a document is inserted a special link migration tool parses all outgoing hyperlinks in the older document and inserts them at the appropriate location in the new document. Only outgoing links are considered during migration because incoming links should not be touched - they are normally comments for exactly the version of the paper they are pointing to.

Considering a paper one will find a lot of so-called vocative links. A vocative link is a textual pointer to a location such as "see also page nn". In scientific papers one will normally find a references section with pointers to other publications - again a typical example of vocative links.

These considerations about vocative links have lead to the development of an automatic vocative link creation tool. The tool is based on a so called "Vocative Link Creation Language" (VLCL) that was designed to support the description of contexts in documents and to find potential hyperlinks in it. To spare the reader from too many details, here is a small example:

Consider a piece of text with the phrase *...details can be found in [Moser 95]...*. The tool will identify this phrase to be a vocative hyperlink leading to the references section and insert a link to the references here. Parsing the references it finds the entry *[Moser 95] Moser J., "The Art of PostScript programming", available at http://www.iicm.edu/app, Dec. 95.* The URL will be identified by the tool and a link to the according document will be inserted. Since VLCL is a programming language the behaviour can even be controlled to the extent that the link to the references is not created but leads directly to the location instead.

If this sounds complicated - VLCL was designed to describe such rules in a very compact and high level manner. Normally VLCL programs having the functionality described above are not longer than around 30 – 40 lines. Furthermore they normally have to be written only once because journals each have their own well-defined citation rules that do not change much over time. In addition, citation rules do not vary much between journals so that an existing VLCL program can be slightly modified and will then suit the needs of another journal.

A different tool for automatic hyperlink creation deals with glossaries. A glossary in Hyper-G is defined as an arbitrary collection of explanatory documents that are classified by their titles and keywords. The glossary link creation tool accepts arbitrarily many glossary collections and automatically interlinks items in papers with glossary entries. The links created get the special type *glossary* so that the reader can turn them on and off seperately when needed. The glossary link creation tool works at the moment for HTML documents; PDF and PostScript support is under development.

Feedback about the behaviour of readers is necessary to improve the quality of a server. Standard statistics tools today are able to count the number of accesses to a document and give information about the location of the reader. Due to Hyper-G's user session concept, a lot more information can be extracted from the logfiles. For this reason a specialized statistics tool was developed that is also able to trace the readers' way through the server from the beginning to the end of a session.

Using the session-oriented statistics, the information provider can see the "typical" path of a reader through the server and find out about specific problems that arise. As an example, the server operator could see that the majority of users are quitting a session when downloading a certain document. In this case the information provider could examine the document and perhaps find out that it contains a huge inline image that takes too much time to be downloaded. Another situation could be that link references are misunderstood and the readers get a page they don't want.

It can also be analysed whether the structure on the server is easy to handle for the reader. Items that are often searched instead of accessed directly are very likely not easy enough to reach. Parts of the structure that are never accessed also alert the operator that something must be wrong there. Analysing access and path statistics carefully and inserting appropriate links or restructuring the server accordingly helps a lot to ensure proper quality.

4 Making Life Easier for the Reader

Up to now discussion has been about how to add as much information as possible to documents. One of the points was to interlink documents to the maximum extent possible by adding inter- and intra-document links as well as glossary hyperlinks and more. Using the tools described above this can be done with minimum effort and stability of hyperlinks is guaranteed by Hyper-G. But there is a negative aspect too: too many hyperlinks also means too many color changes which disturbs the reading flow.

As has been discussed earlier Hyper-G supports different kinds of hyperlinks: annotations, glossary links, referential hyperlinks, inline links, texture links and many more. To avoid too many disturbing color changes, links of different types can be turned off and on seperately when needed. Another possibility that is only implemented in Harmony at the moment is to show different kinds of links in different colors (Harmony is one of the very specialized Hyper-G authoring tools). Utilizing this feature, the reader would know about the type of a link even before following it.

As has been mentioned before grouping of documents in clusters is one of the basic concepts in Hyper-G. One of the applications of clusters is it to provide multilingual versions of the same document. Naturally it is normally not possible to translate every single document on a server to several different languages. On the other hand it is not too much effort to translate neuralgic and mostly static parts of the data such as glossaries. In reality, if there are multilingual glossaries existing it normally does not even help too much to translate the papers as well.

Consider an international journal: the language of papers will most likely be English and readers will understand English. The only problem is that some very topic-specific words will be found in the papers, in which case readers can use the glossary to look up unknown words. If the readers were not native English speakers and the glossary also contained a difficult explanation they might be forced to follow some more links between glossary entries until they finally understood the meaning. For this group of readers, a translation of the first entry would very likely help a lot. And there are certainly a high number of non-native English speakers reading English papers.

An all-time-important aspect on the Web is transmission speed. Long distance data transmissions on a slow transmission rate are annoying and in the worst case, electronic publications will not be accepted by the reader for the lack of availability. There are two ways out of this dilemma that can be done simultaneously: providing documents of different size and quality as well as mirroring and caching documents.

For providing documents of different size again Hyper-G's cluster concept can be used. For instance, a cluster could hold several verions of the same document: one in HTML format with small inline images for quick browsing, a second, better version could also be HTML but with high quality images and finally a third, the professional version, could be PostScript with 600 dpi resolution including 600dpi true color images. The reader decides which level of quality to get once per session and the rest is done automatically throughout the whole session. Of course, readers can also change their choice on the fly.

More than that - consider the case of an HTML document with different sets of inline images - small ones with poor quality and huge professional ones. On standard first generation Web servers two different versions of the HTML document must be provided with different links. On Hyper-G links do not point directly to the images but instead to the clusters! In this case the HTML document only has to be prepared once and the inline images automatically change according to the readers' choices, even on the fly if desired!

The other well known solution to overcome the transmission speed problem is it to cache and mirror documents and of course Hyper-G has the standard proxy functionality implemented. Besides that, a sophisticated document replication mechanism [see Kappe 95] is implemented in Hyper-G.

Document replication in Hyper-G is understood to be a special kind of mirroring with the mirrored documents knowing about the original. This is possible by giving the documents a replica identification that matches the global object Id of the original document. The global object Id of documents in Hyper-G is a world wide unique 64 bit identifier.

The effect is that documents can be mirrored to the local site without users having to know about it. If they try to access one of the original documents on the remote server or even one of the documents on another mirror site they automatically get the replicated document from the local site instead of doing a long distance data transfer.

5 Current Electronic Publications With Hyper-G Technology

- The first electronic journal based on Hyper-G was J.UCS - the Journal of Universal Computer Science by Springer Pub. Co. It is a monthly journal covering all knowledge areas of computer science. In addition to the Web version, a yearly CD-ROM and printed version are provided by Springer. Papers in J.UCS appear in two parallel formats: hypertext and hyperlinked PostScript, PDF is planned for 97. See [http://www.iicm.edu/jucs] - the master J.UCS server - for more details and information about mirrors.
- Academic Press decided to distribute JNCA - the Journal for Network and Computer Applications (former JMCA and JMA) - electronically using Hyper-G starting in January 96. JNCA can be found at [http://www.iicm.edu/jnca], mirrors are in preparation.
- One of the most reputable journals in physics, FBS - Few Body Systems - also by Springer - started regular electronic service in January 1995 using Hyper-G. Find FBS at [http://www.iicm.edu/fbs].
- The German bible of Data structures - Datenstrukturen by Ottmann and Widmeyer - is published electronically on a Hyper-G server using hyperlinked PostScript. At the moment the German version is available at [http://www.iicm.edu/datenstrukturen].
- Since this paper excessively deals with Hyper-G - naturally also the "Hyper-G bible" called "HyperWave - The Next Generation Web Solution" published by Addison Wesley is available electronically at [http://www.iicm.edu/hgbook] (Note: HyperWave is the product name of the software whereas Hyper-G is the name for the underlying technology). Check out [http://www.iicm.edu/hyperwave] for detailed information about Hyper-G technology and the HyperWave product line.
- In additional to the Hyper-G bible, Addison-Wesley decided to publish some 30 books electronically on the Web using Hyper-G. Some of them are already under preparation. As soon as they are released they will be found in the IICM electronic library [http://www.iicm.edu/electronic_library].
- One of the most comprehensive German encyclopedias - Meyer's Lexikon - is electronically available via Hyper-G on an n-user license basis - have a quick look at [http://www.iicm.edu/ref.m10] if there is something you always wanted to know.

References

[Adobe 90] Adobe Systems Inc.: "PostScript Language Reference Manual", 2nd Edition, Addison-Wesley (1990)

[Adobe 93] Adobe Systems Inc.: "Portable Document Format Reference Manual", Addison-Wesley (1993)

[Andrews 95] Andrews K., Kappe F., Maurer H. and Schmaranz K.: "On Second Generation Network Hypermedia Systems", Proc. ED-MEDIA '95, (1995), 69-74.

[Berners-Lee et al 94] Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. and Secred, A.: "The World-Wide Web." Communications of the ACM 37,8 (1994), 76-82.

[Kappe 91] Kappe F., Maurer H. and Tomek I.: "Hyper-G - Specification of Requirements", Proc. Conference on Intelligent Systems (CIS) '91, (1991), 257-272

[Kappe 95] Kappe F.: "A Scalable Architecture for Maintaining Referential Integrity in Distributed Information Systems", J.UCS 1,2, (1995), 84-104.

[Maurer 94] Maurer H. and Schmaranz K.: "J.UCS - The Next Generation in Electronic Journal Publishing", Computer Networks and ISDN Systems, Computer Networks for Research in Europe, Vol. 26 Suppl. 2, 3, (1994), 63-69.

[Maurer 95] Maurer H. and Schmaranz K.: "J.UCS and Extensions as Paradigm for Electronic Publishing.", Proceedings DAGS'95, Boston Massachusetts, (1995).

[Maurer 96] Maurer H. ed.: "HyperWave - The Next Generation Web Solution", Addison-Wesley, (1996).

[Murray 94] Murray J., D. and van Ryper W.: "Encyclopedia of Graphics File Formats", O'Reilly and Associates, Inc. (1994).

[Odlyzko 95] Odlyzko, A., M.: "Tragic Loss or Good Riddance? The impending demise of traditional scholarly journals." In "Electronic Publishing Confronts Academia: The Agenda for the Year 2000," Robin P. Peek and Gregory B. Newby, eds., MIT Press/ASIS monograph, MIT Press (1995).