Reducing White Noise
Towards a common information garbage policy

Dr. Stefan Gradmann
Regionales Rechenzentrum der Universität Hamburg

stefan.gradmann@rrz.uni-hamburg.de
Schlüterstraße 70, D-20146 Hamburg
http://www.rrz.uni-hamburg.de

Keywords: *information overload, libraries, information filtering, information aggregation*

Abstract: *Information filtering and aggregation become increasingly vital for academic communities wishing to use productively WWW-based information infrastructures, which in turn suffer from growing information overload bearing entropic traits. New co-operative strategies implying players traditionally acting in separate sectors of scientific information organization - such as libraries, computer center and multimedia centers - may contribute to innovative information garbage policies in academic communities. Ongoing work at Hamburg University illustrate the potential shape of such future institutions for leveraging, filtering and aggregation of information that re-implement the concept of university libraries in a rapidly changing context.*

Introduction

One of the visions almost excessively quoted in current discussions of what the information infrastructure of the WWW may ultimately turn out to be is Vannevar Bush's article "As we may think" written in 1945, and thus quite some time before the advent of computers as we know them today. In fact, Bush's vision concerns the organization of information much more than the technical means and instruments used for that goal. Bush has often been quoted, because some of the technical ideas developed in that context seem to be excellent guesses at what later architectures for information automation turned out to be - and this especially concerns the famous Memex metaphor, prefiguring to some extent some of the basic technical principles underlying the information architecture of the WWW. Still, the article actually is mainly concerned with the organization of information, with modes of accumulation, aggregation and selection of information.

And the striking fact, in that respect, is Bush's primary concern, which is information selection (rather than accumulation) - long before the advent of the tremendous amount of trash and pearls we continuously are confronted with when using the information space of the World Wide Web. After discussing the rapidly increasing means for content production and accumulation Bush states

The prime action of use is selection, and here we are halting indeed. There may be millions of fine thoughts, and the account of the experience on which they are based, all encased within stone walls of acceptable architectural form; but if the scholar can get at only one a week by diligent search, his syntheses are not likely to keep up with the current scene.

- to continue, however, in a later section stating

The real heart of the matter of selection, however, goes deeper than a lag in the adoption of mechanisms by libraries, or a lack of development of devices for their use. Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing.[1]

This last statement on the crucial role of information organization in the context of efficient selection of organization leads Bush to the famous Memex metaphor and thus to the vision of a new, networked information paradigm, in which information is organized not so much in taxonomic tree structures but in networked threads of associations.

White Noise

The American writer Don Delillo has coined one of the strongest metaphors for the problem that had been indicated by Bush and that this paper is concerned with, too. One of DeLillo's most successful novels bears the title "White Noise" [2] and is essentially concerned with the entropy-like situation caused by information overload: if the probability of locating a specific bit of information is equal at any given point on the information map, and this simply, because these bits of information are overwhelmingly omnipresent and means for their filtering and for aggregating information are absent, the effect is white noise information, non-information caused by an overload of unstructured information.

DeLillo's novel was written in 1985, and thus the loss of coherence in world perception and organization caused by information overload depicted in this book is mainly associated to the television medium. Had this apocalyptic vision been written 15 years later, it could well have been about information in the World Wide Web. When describing the information ecology of the WWW in 1997, Jeff Ubois stated that "the rapidly widening gap between the amount of data in the world and the amount of attention available to process it means a growing percentage will never be looked at by a human." [3] This could have been put even more pessimistically: not only is there a risk of substantial information never actually being retrieved: the real threat is that of valuable information being submerged by heaps of irrelevant bits of information, by what David Shenk called "Data Smog" in his monograph published in 1997. This paper is about this data smog, the tendency towards entropy induced by its omnipresence, the danger of WWW information quality being reduced to the precision and specificity of white noise. But most of all, this paper is about some of the means we have or we may conceive to prevent this entropic tendency from becoming increasingly dominant and ultimately suffocating in terms of information ecology. More specifically, the contribution is thus concerned with the role of traditional players in academic institutions - such as computer centers and libraries - or newly emerging institutions - such as multi-media centers - in this context of what may be termed an academic information garbage policy.

The sad fact is, that the logic of building WWW-based information services, until now has remained mostly cumulative, the sheer amount of accumulated information having been considered a quality in itself during a long period. This cumulative logic was a reaction to the traditional information paradigm, in which the generation and accumulation of information was a tedious process, which went along with strong mechanisms for information filtering and aggregation, some of these deeply anchored in academic culture. Sparseness of information resources and the difficulties of accessing them physically further decreased dangers of information overload in this traditional information economy. With this past situation in mind, any infrastructure granting easy, fast access to huge amounts of information was of course essentially perceived as eliminating restrictions and thus a value in itself.

The information explosion within the WWW thus has been acclaimed for its sheer quantitative aspects for quite some time, as if the pure doubling of bytes available on the Web was equivalent of a proportional growth in informational value. However, this tremendous amount of easily accessible electronic information comes with only poor mechanisms for information filtering and aggregation.

Approaches based on the use of search engine are a good example for such shortcomings, the following example (Figure 1) is meant to illustrate the almost uncontrollable precision rate combined with only seemingly impressive recall obtained when searching for the term ‘virtual library’ via the meta-engine SuperSeek:

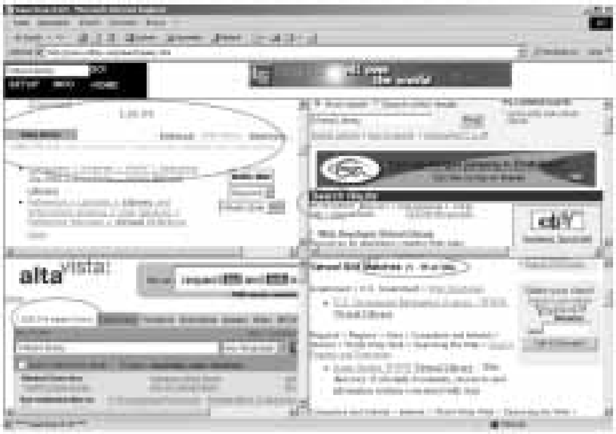

Figure 1

These results are questionable without doubt: while the highly selective Yahoo service announces 336 matches, InfoSeek already reports 64.118 of them and Lycos as well as AltaVista announce truly impressive figures with 1.407.776 web sites and 1.630.740 pages respectively. Even though these discrepancies may partially be due to the different functional paradigms underlying the search engines it is close to impossible to judge the actual relevance of the ‘information’ thus retrieved. The lack of transparence of the respective ranking algorithms clearly doesn't contribute to a more focused picture either: in fact, when comparing the top ten results reported back from each of these information services the user ends up with 35 different site indications and none of these contained in all four top ten sets (and only one reported back by at least three services, another four figuring in at least two top ten result sets).

This example should make clear that WWW based information services and the tools used for retrieving these certainly are doing an impressive job in terms of data accumulation but that they are rather poor aggregators of information.

Such pure accumulation of information does not serve research as such. Information overload, bearing entropic traits, even tends to harm serious research work, and it certainly does so, once individual researchers are submerged by heaps of information without a real chance to determine the actual relevance of any given bit of this information stream.

Even though actually tackling this problem clearly is far beyond the reach of any individual institution and technical solutions therefore need to be found in a more global context, academic institutions - as all other users of WWW information services - nevertheless are in need of consistent strategies for locally dealing with white noise information. The following sections of this paper provide some examples for elements of such strategies together with indications concerning the players needed to implement them.

Network noise reduction: the role of computing centers

A first level of white noise reduction - within the network layer - will only be briefly considered in this context, simply because it is well known ground for university computer centers: such computer centers traditionally are concerned with opening and securing information channels, and in this context also traditionally have put to work mechanisms for filtering evident information spam and other unwanted information. Institution-wide policies for e-mail use, installation and configuration of proxies and firewalls are examples of such means for low-granularity noise reduction. The main reason this level is mentioned even though is the fact that any strategy intended to seriously tackle the problem of information overload within a university will have to be based upon the networking layer typically maintained by the computing center and thus calls for active participation of this player.

This fact is one of the major reasons for the choice that has been made at Hamburg University to primarily locate efforts in the central field of semantic noise reduction within the computer center in order to create a maximal potential of synergies among the different players tied together within the common networking infrastructure.

Semantic noise reduction

The core field of noise reduction is the semantic information level, since actual relevance of information required for focusing, filtering and aggregation of information can only be determined on this level. The examples given below illustrate possible lines of action in this area and indicate two of the potential players that have to interact in order to reduce information overload (or also prevent overload from being generated).

The role of Libraries

Apart from acquiring content and making this content available for their users libraries always systematically been concerned with content selection and thus with semantic focusing of information systems. This concern with content selection was never exclusively motivated by the sparseness of material resources: a prominent objective of this activity always has been to distinguish potentially relevant items from clearly irrelevant information and furthermore to aggregate potentially relevant material by means of subject indexing, classification and the creation of bibliographies. One of the most important activities in this respect was the cataloguing of information items, in other words: generation of metadata. Libraries thus have a long tradition in fine-granularity noise reduction and content focusing - but in the world of printed documents only, with hardly any expertise as to the only just emerging techniques that are relevant for electronic content.

However, the traditional strengths and competences of libraries can be made to contribute efficiently to the goal of noise reduction in the changed information paradigm of networked electronic information resources provided these libraries extend their expertise with new techniques of work and can be made to co-operate intensely with the new players in this rapidly changing infrastructure.

One of the traditional areas of librarian competence most heavily affected by this need is the process of metadata generation. Cataloguing as traditionally practiced in libraries is a time-consuming and expensive activity even in its original context of printed publications. This process becomes completely inappropriate and impossible to sustain with the advent of networked electronic information resources: the traditional cataloguing approach has not the slightest chance to catch up with these rapidly proliferating bits of information and even if this was theoretically possible no institution could pay the price of such an attempt.

On the other hand, metadata are an excellent antidote for white noise information, among the most efficient means for enhancing retrieval precision and thus reducing information overload potential. Unfortunately, metadata - as far as these are already available - are seldom used by popular search engines, mainly because of past abuse of HTML meta tags by commercial players wishing to ensure a prominent position of their pages in the result set ranking algorithms of these search engines. New, sustainable strategies for generating metadata are thus required, and means have to be sought to massively increase the use of meta-information in the WWW infrastructure. One way to do so is to massively involve the producers of electronic information in the process of metadata generation, which itself must be substantially simplified in comparison with traditional bibliographic standards and at same time ensure a minimum of coherence and consistency in metadata output by producing meta-information complying to an institutional quality policy and which is certified by this institution.

At Hamburg University, we try to put such concepts to work in two ways, both of them heavily involving our libraries but - unlike in the former cataloguing models - in close co-operation with our scientific staff:

- from 2001 on, at least the top-level WWW-pages produced within the university will contain metadata complying to the Dublin Core (DC) standard, and these metadata will be identifiable as 'institutional' metadata generated under our university's responsibility
- metadata production complying to such standards by authors of scientific work is planned to be a systematic requirement both for submitting resources to the internal, intranet-based document workflow as for output in the digital university press infrastructure we are currently setting up.

However, such approaches - even if they may substantially improve the information ecology - only apply to 'syntactic' metadata (the equivalents of bibliographic descriptions) and do not solve the problem of semantic aggregation (assignment of subject indexing terms and/or keywords as well as classification), which cannot be done intellectually/manually for the multitude of resources concerned. Furthermore, the two steps sketched above only concern resources we produce ourselves at Hamburg University: external information resources integrated into our systems do not necessarily comply to our internal standards: they may contain metadata complying to different - even superior - standards or even no metadata at all.

Currently, we therefore are preparing a project supposed to deliver tools for the use by librarians and scientific staff in two ways:

detection, extraction and - if needed - on the fly conversion of metadata present in external document resources for ingesting these together with the document resources themselves into our information system

automated generation of 'semantic' metadata (keywords, lexical clusters) using lexicon-based approaches for semantic extraction, aggregation and filtering combined with morphological normalization techniques in order to produce controlled vocabulary associated to documents ingested in our repositories as well as automatically generated abstracting information.

In preparing this project, we build heavily on existing know-how and work already done, especially in the domain of metadata extraction, by such institutions as UKOLN ('DC-DOT') or OCLC. Partners with strong experience in language engineering will supply the linguistic techniques needed for semantic aggregation procedures. Furthermore, ongoing national and international projects and initiatives are closely monitored, and be it only to do as little work locally as possible. Examples of such external initiatives are the German Carmen project or the Open Archives Initiative, but also current work being done by scientific publishers. Our specific task is to pull these elements together and to define a consistent information policy for our university integrating these and other technological approaches into an institution-wide strategy for white noise reduction and for preventing information overload.

The Centre for Media Competence

Institutional players such as libraries and computer centres thus can do a lot to reduce white noise in information services, but these efforts of very moderate use only, as long as the users of such services are not aware of the problem and have not been taught to use and generate information resources efficiently themselves. The building of a centre for media competence currently under way at Hamburg University - among other concerns - is a reaction to this aspect of the problem. To some extent, the centre for media competence is conceived as a 'traditional' multi-media centre providing the relevant infrastructure for use and production of multi-media resources and for archiving and preserving multi-media content.

A very strong emphasis, however, is put on building competence within the academic user community and thus enabling these students and searchers to make efficient use of multi-media resources in their learning and teaching work with specific emphasis on the building of efficient environments for multi-media based tele-teaching and tele-learning. Multimedia can do a lot of harm in adding tremendous volumes of white noise to the information ecology of an academic institution. It would therefore not have been sufficient to conceive our multi-media centre as a mere institution for generation an accumulation of multimedia content: the aspect of user education is seen as a key factor for such an institution to make a useful contribution to the university's information infrastructure instead of just setting up yet another powerful means for information pollution.

Making Strategies converge

Other players are part of the game, even though their role is not explicitly identified here. First among these are students and the scientific staff of our university, our primary customers. In order to efficiently serve these, in terms of reducing information overload in this case, different actors have to join their forces and make their strategies converge.

This fact may well be illustrated having a look at another problem we intend to tackle in a project currently under preparation and which is concerned with methods for generic user authentication in various application contexts.

Authentication is one of the essential prerequisites for making Digital Library (DL) resources available: both in terms of resource and of user identification and authentication. Before providing access to document resources, any DL application needs to check the identity of the requester and/or his organizational affiliation and eventually match this information with corresponding data stored inside the DL application. Likewise - and especially in the case of electronic resources which are subject to various changes of document status and rapidly proliferating - requesters may wish a proof of document authenticity of the resource they are asking for and which they eventually have to pay for.

Students and academic staff are confronted with multiple (and rapidly proliferating) authentication methods and instances in this respect: not only do they have to identify within their own institutional context, but also authentication is often required anew by external document providers linked to the DL environment. Finally, the scope of the problem is further increased in a hybrid university context with authentication requirements added by typically at least the local library automation system and eventually additional, newly emerging instances (such as campus management systems for administration of students and of exams).

This proliferation of authentication contexts involving differing methods and context specific data (passwords etc.) to be kept in mind, all of these relating to one given individual, is creating a very specific brand of white noise in the area of authentication information and the means to tackle this, once again, is information aggregation, which is the core task of the infrastructure we are setting up for providing uniform authentication services.

The technical solution proposed is based upon the idea of integrating all authentication information pertaining to one given individual person or institution within a single entry for this entity as part of a LDAP based directory service. The directory service models the organizational tree of the university starting from the "root" organization on top via linked sub-organizations down to the level of individual users.

This directory service can be accessed by duly authorized applications in order to extract the information required to authenticate the entity (person or institution) within this application's specific context. The basic, simplified relation between the directory service and two given external application in terms of the underlying data model is sketched in the following diagram (figure 2), where application 1 and 2 share basic identification data contained in the generic LDAP record and further use their respective, specific data segments within this same record

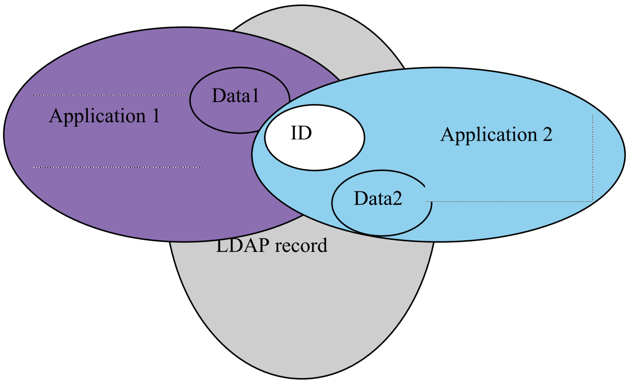

Figure 2

In order for this basic model to work the directory service must provide a query and data transfer interface and make selected information available to the external instances regarding its internal data organization (common elements which can be accessed without restriction vs. application specific elements accessible for instances with specific rights only).

However, this basic model needs substantial enhancements, and be it only for security reasons: LDAP in itself implements a public, open directory infrastructure and has to be complemented with strong security features in order to ensure that only registered client applications may access the service and that these can only access the common elements plus those of specific relevance for them. Public/private key based encryption techniques for communication between the directory service and the client applications will thus be used on top of an SSL transport layer. Furthermore, certification will be required for the directory server and for the client applications, which in turn must require certification for their respective users accessing directory information via these application clients. A certification agency (CA) thus must be added to this infrastructure.

The infrastructure created this way can further be used to implement document authentication procedures, as well. Other examples for content aggregation and filtering strategies could have been supplied, too, such as our ongoing efforts for building a distributed e-publishing environment with an electronic University Press output component coupled with networked methods for peer reviewing and quality control or the collaborative large scale efforts currently made for metadata generation related to quality content in the WWW within the CORC project. In all these cases, in order to successfully build the relevant infrastructure components and subsequently tie these together in operational workflow models, active participation of almost all relevant players within a given academic institution is strictly required and their respective technical and functional strategies must be made to converge to make common use of such components as the directory service mentioned above or to comply to the related security architecture.

More generally thus, structural convergence is not only just an option in the field of information organization and aggregation, but will turn out to be vital. Multi media centers will increasingly be in need of a semantic focus, libraries will be unable to apprehend, leave alone develop, relevant information filtering techniques for electronic content that can no longer be aggregated using pure intellectual/human means and computer centers will be confronted with the need to stretch beyond mere implementation of basic information protocols.

This joint effort for reducing information overload contributing to the building of the "semantic web" on the one hand will benefit the respective university's scientific community that may have more demanding tasks to serve than to deal with white noise from the Internet. But on the other hand it also induces institutional convergence among the players involved and is thus likely to prepare the ground for new institutional models transcending current barriers between information organization agents within universities. This contribution also was meant to illustrate the shapes such future institutions for leveraging, filtering and aggregation of information may take, whatever name they finally may adopt: computer centers, multimedia centers, libraries - or probably some new term yet to be coined.

Bibliography

1 - Bush, Vannevar: As we may think. In: The Atlantic Monthly 176 (1945),1, pp. 101-108

2 - DeLillo, Don: White Noise. New York: Viking, 1985. ISBN 0-670-80373-1

3 - Ubois, Jeff: Casting an Information Net. In: Upside Today 1998 http://www.upside.-com/texis/mvm/story?id=34ce6fdb0