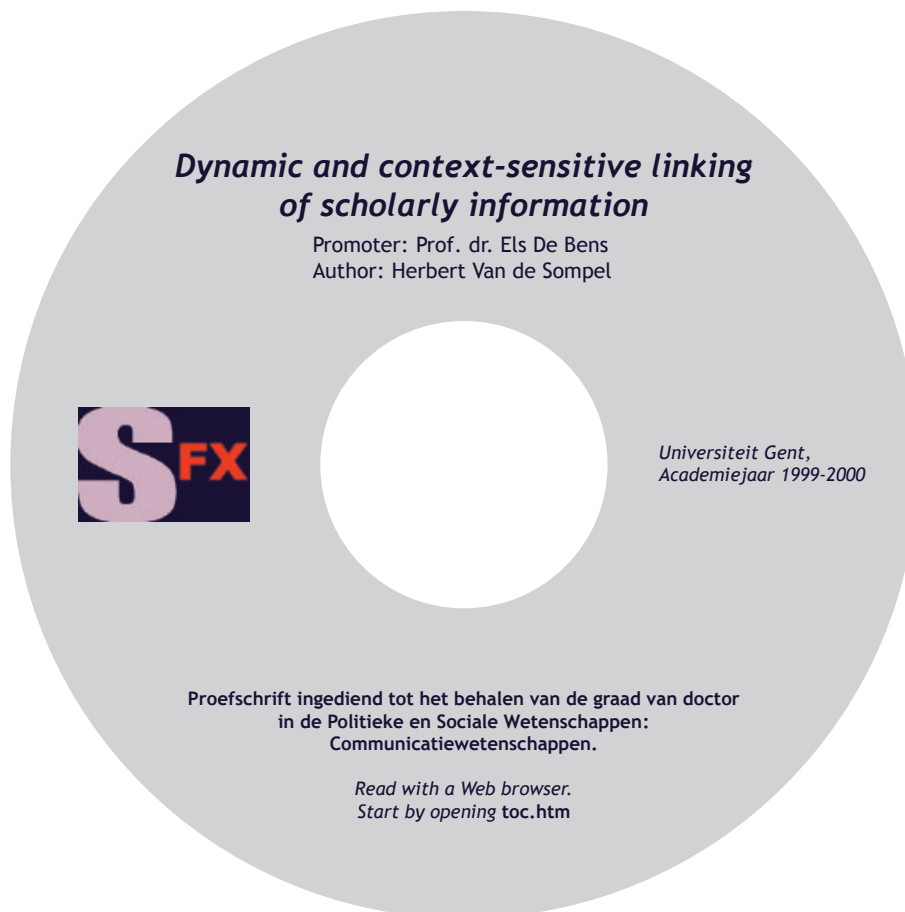


Proefschrift ingediend tot het behalen van de graad van doctor  
in de Politieke en Sociale Wetenschappen :  
Communicatiewetenschappen

# Dynamic and context-sensitive linking of scholarly information



# Dynamic and context-sensitive linking of scholarly information

## Table of Contents

<b>Introduction</b>	
<b>Acknowledgements</b>	
<b>Problem statement</b>	
<b>Experiments and Concepts</b>	
	Part 1: Initial identification of characteristics of the SFX linking solution in the Elektron experiment
	Part 2: Generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment
	Part 3: Applying SFX to integrate e-prints with the established scholarly communication system in the UPS Prototype experiment
<b>Conclusions</b>	
<b>References</b>	
<b>Appendices</b>	
	Press Releases on SFX
	Miscellaneous on SFX
	Miscellaneous on the Open Archives initiative

# Introduction

---

This thesis demonstrates that it is feasible to interlink distributed electronic scholarly information resources in a context-sensitive and dynamic manner.

This terse formulation of the linking problem addressed in the thesis is explained in detail in the Problem Statement. The latter also describes the motivations to take on the research effort. In addition, it actualizes the problem in terms of ongoing linking research and linking initiatives. The thesis achieves its goal by gradually designing and testing the Special Effects -- henceforth SFX -- linking framework over the course of three experiments. The reports on these experiments are included as separate Parts of the main section of this thesis that goes under the name Experiments and Concepts. This chronological organization highlights the evolution of the concepts that underly the solution and the reasons why they evolved. Generally, this organization helps to make the thesis accessible.

One can hardly claim that a solution is able to interlink distributed scholarly information resources when the testbed remains restricted to -- a subset of -- a single digital library collection, as was the case in the Elektron experiment described in Part 1. Nevertheless, some fundamental characteristics of the SFX linking framework are already introduced in that Part. In order to make a more credible claim, the testbed had to be extended significantly. This was done in the course of the "SFX@Ghent & SFX@LANL" experiment -- described in Part 2 -- for which the author spent six months at the Los Alamos National Laboratory. Concepts introduced in the first experiment were applied in the complex digital library collections of the University of Ghent and the Los Alamos National Laboratory. This complexity led to a crystalization and generalization of the SFX framework, that -- because of the modularity, adaptability and proven performance it achieved in the course of the experiment -- could already be considered to be adequately generic to serve as a proof of the thesis. But the conclusive proof is presented by the application of the framework -- designed in Part 2 -- in an environment that is radically different from the ones of the other experiments. In the UPS Prototype project, described in Part 3, both the technical environment and the content of the core collection can hardly be compared to those of Elektron and "SFX@Ghent & SFX@LANL". So far, mainly traditional scholarly information resources -- abstracting and indexing databases, electronic journal repositories, OPAC systems -- had been included and interconnected in the experiments. The Universal Preprint Service -- UPS -- Prototype concentrated on the integration of the subversive scholarly information resources -- the e-print archives -- with the traditional ones. The UPS Prototype project was complex and elaborate, and circumstances hindered the straightforward application of the SFX framework. Nevertheless, the project results are convincing and as such confirm the generic nature of the SFX linking framework.

The Conclusion provides an overview of the major findings of the three experiments. It identifies the major components of a context-sensitive linking solution and it presents a summary of the design and of the major characteristics of the SFX linking framework. It concludes that it is indeed feasible to interlink distributed electronic scholarly information resources in a context-sensitive and dynamic manner.

The Acknowledgements section is quite elaborate and is a good illustration of the nature of digital library research. Digital library research is not just a combination of library & information science and computer science. In order for it not to be sterile, it must take on real-life challenges, and deal with the complexities and dimensions thereof. Many such

challenges cover too many domains -- both Internet domains and areas of expertise -- to be handled by a single person. Therefore, extensive collaborations can be required to achieve convincing results. Establishing and coordinating such collaborative efforts is part of the social dimension of digital library research. Although it is an aspect that can hardly be measured on any scientific scale, it can be decisive with regard to the success, failure or credibility of a digital library research effort. As such, a sentiment of pride rather than of embarrassment comes with acknowledging so many people for their contribution to the results presented in this thesis.

Another aspect of digital library research is the amazing pace at which its research subjects evolve. Doing digital library research is akin to shooting at a fast moving target. Therefore, it should come as no surprise that important parts of this thesis have already been disseminated to the interested audience. This has been done via the publication of 'stories' in D-Lib Magazine addressing the SFX linking framework. This initiated a positive feedback loop that had constructive repercussions for the evolution of the ongoing SFX-research, as well as for initiatives -- such as the NISO Reference Linking workshops -- dealing with a related problem domain. Also, it stimulated a dialogue with important parties of the scholarly information industry that was valuable for all parties involved. The net impact of the research reported in this thesis would not have been comparable should the findings have been kept silent until their release via this document. The latter is true for both the academic results as well as for the penetration thereof into the communities that are concerned with the problem domain -- research libraries, the scholarly information industry, the digital research community.

In order to not raise any false expectations with regard to the content of this thesis, it is prudent to briefly delimit its working area. The 'scholarly information' mentioned in the title are the electronic scholarly information resources that are most commonly used for dissemination and discovery of scholarly information: abstracting and indexing databases, citation databases, repositories holding full-text articles of traditional scholarly journals, library OPAC systems and e-print archives. Occasionally even web resources that can not easily be classified with this traditional terminology make their appearance. Other electronic resources that may also be used in the course of scholarly activities -- electronic books, encyclopedia, list servers, etc. -- have not been dealt with. Also the 'context-sensitive' property mentioned in the title requires a little explanation. Roughly speaking, the grain of the context-sensitivity that is being addressed is at the level of an institutional digital library collection that is accessible by users of the institution. It does not deal with a finer grain such as the context of an individual user's personal preferences, his personal authorizations, or more technical contextual issues such as the user's physical location or access method. Consequently, the research does not step into the quicksand of authentication and authorization, although -- at a certain point -- it points to a noteworthy parallelism between the SFX approach and a research effort dealing with this important global problem.

Finally, a remark with regard to the medium on which the thesis is delivered. Static stand-alone paper is not an appropriate habitat for a thesis on linking. Therefore, the paper is included with the CD-ROM, not vice versa. The thesis on the CD-ROM can be used on any computer that is capable of running a regular web browser. The opening file -- *toc.htm* -- is located in the root directory of the CD-ROM. All sections of the thesis are mutually interlinked, as well as linked to related information on the Web. Most importantly, the CD-ROM contains movies that are the most compelling illustrations of the results of the reported experiments. A reading or evaluation of this thesis is impossible without viewing these movies. Unfortunately, to do so, Mac and UNIX users will have to turn to a Windows computer. All movies are in the *movies* subdirectory of the root directory of the CD-ROM; they are named according to the experiment they originate from: *Lanl* for the Los Alamos results of "SFX@Ghent & SFX@LANL" ; *Ghent* for

the Ghent results of the same project ; *Ups* for results of the UPS Prototype project. With a fast CD-ROM player, each movie can be started by double clicking its icon. With a slow CD-ROM player, the movies will have to be copied to a hard disk first.

Herbert Van de Sompel

Ghent, February 29th 2000

# Acknowledgements

---

There are so many people that I need to thank for their active or supportive contribution to the realization of this thesis. I have chosen to mention them -- more or less -- in the chronological order by which they crossed the path that has led to its submission. Things started with Professors Greta and Ludo Milis who motivated me to take on this important challenge. For many years, both have been very supportive of my work at the University of Ghent, and ultimately they succeeded in convincing me to apply for a grant that would allow me to embark on a year's sabbatical from my work as head of the library automation department. Professor Els De Bens enthusiastically welcomed this idea, and -- with Professor Richard Philips and Professor Guido Van Hooydonk -- supported my request for a grant with the Belgian Science Foundation. They did so with great success and as such played a crucial role in making this research effort happen. Many thanks to the Belgian Science Foundation for according me a special Ph.D. grant that allowed me to concentrate on this thesis for a year. Professor Els De Bens became my promoter; I want to thank her especially for the courage she showed in actively supporting someone with a background that is quite unusual for her faculty. Special words of gratitude are also due to Professor Guido Van Hooydonk -- my head librarian -- for his incomparable trust in letting me guide the Ghent digital library efforts for many years and for his persistence in the pursuit of an environment where novel ideas can flourish and can be experimented with. We are both intrigued by the potential of a linked scholarly information environment and had inspiring discussions on the matter.

Patrick Hochstenbach, who is on my team in the Ghent library automation team, deserves a special paragraph in these acknowledgements. There is no doubt in my mind that without his involvement there would be no thesis on SFX linking. Patrick was involved from the very first experiment and -- with me -- he has lived through all the iterations of the concepts that ultimately became the SFX framework. He made the experiments possible by creating the required software. Several discussions -- in person, via e-mail, by phone and via chat -- on the issues involved in establishing the framework have had an important impact on the directions taken.

Lieve Rottiers, also on my team in Ghent, has also been involved in SFX from the very start until the finalization of the thesis. Her most important contribution to the SFX experiments was the data gathering for the SFX-base, a less than cheerful task that she performed with an amazing precision and persistence. She also took care of the layout of the consecutive user interfaces that were used in the experiments and she has been a great help in the process of wrapping up this thesis. Many thanks also to the others on my team -- Paul Bastijns, Dominiek Decleyre, Lieven De Vos, Danny Van den Bulcke, Frank Vandepitte -- for taking such good care of the library automation activities during my absence. The other person that has witnessed the complete evolution of the SFX research is Jennifer De Beer. I met Jennifer very briefly at the occasion of a Symposium at the University of Stellenbosch (South Africa) where I presented a paper. We kept in touch via e-mail and Jennifer, a linguist by training, proofread all my writings on SFX. Many thanks to her for a job so well done.

Then, there are two people who initially made crucial contributions to making my stay possible at the Los Alamos National Laboratory (LANL); both ended up changing my life over the course of this thesis. I first met Deanna Marcum, the president of the Council on Library & Information Resources (CLIR), at an international meeting in the Slovak Republic. A few months later, Deanna came to visit me in Ghent. By that time, my plans for a Ph.D. had become solid and we discussed the possibilities of a stay at an important digital research library in the

US. Behind the scenes, Deanna started to work hard to make that happen. CLIR accorded me their only travel grant of 1999. But Deanna's impact reached much further than this. Throughout my stay in the US she has actively supported my efforts to contribute in a constructive manner to the ongoing transformation of scholarly communication. She has done so both for my reference linking work and for my work in the Open Archives initiative that has received substantial funding from CLIR and the Digital Library Federation of which CLIR is the administrative home. Words are not sufficient to thank Deanna for her trust and support. Rick Luce, the director of the LANL library also came to visit me in Ghent, on his way back from a lecture he gave at the TICER Summer School on Digital Libraries, where we both teach. EBSCO's Wim Luijendijk had convinced him that the digital library in Ghent was worth a detour. Rick was one of the first to see embryonic SFX concepts on a white board during his visit. Probably they must have looked appealing, since Rick immediately sent a positive reply when, sometime later, I cautiously sounded him out in an e-mail about a possible stay on his renowned Library Without Walls team. Once I was over at Los Alamos, Rick supported my every move. His help has been fundamental in the successful conclusion of the "SFX@Ghent & SFX@LANL" experiment, to which he dedicated substantial resources and for which he provided me with important contacts in the information industry. Rick trusted me in handling delicate discussions on the future directions of the Los Alamos e-print archive with Paul Ginsparg, to whom he had introduced me. At several occasions, Rick involved me in policy matters. Both Deanna and Rick have changed my perspective and there is no doubt that it is a change for the better.

Although the Elektron experiment was rather modest, it was also made possible thanks to the support of companies in the information industry. Especially Ex Libris (Oren Beit-Arie, Yohanan Spruch) and SilverPlatter (Denis Lynch, Jenny Walker and Andrew Wilkins) played an important role by believing in the rather vague ideas of which I tried to convince them and by making their systems interoperable with the early SFX solution. Both companies have sustained their support throughout the SFX experiments. Ex Libris has ended up believing so much in the concepts that they decided to acquire the rights to the SFX software from the University of Ghent. Also involved in the Elektron experiment were Academic Press, Swets and UMI by providing test access to their services.

The "SFX@Ghent & SFX@LANL" experiment was much more elaborate and the list of people to acknowledge for their contribution is extensive. To start with, there are all the people with whom I worked at LANL. Everyone on the Library Without Walls team (Miriam Blake, Johan Bollen, Doug Chafe, Mariella Di Giacomo, Frances Knudson, Dan Mahoney and Mark Martinez) made an important contribution to making the experiment happen even if doing so disrupted the very tight ongoing implementation schedules. I can hardly forget the collective feeling of accomplishment that came with the successful finalization of the experiment and the presentation of the results at the New Orleans 1999 Summer meeting of the American Library Association. Also at LANL, Abe Lederman needs to be mentioned for making the Science Citation Databases interoperable with SFX. Crucial to this experiment was the involvement of two more parties from the information industry, because it allowed demonstrating the bi-directional linking capability of the SFX framework. Therefore, special words of thanks for trust and cooperation go out to Mark Doyle at the American Physical Society (PROLA archive) and to Andy Stevens, Andy Townsend and Craig Van Dyck at Wiley Interscience. Every single audience to which I demonstrated the solution were left speechless when seeing a link from citations in their full-text articles into an institutional database such as Inspec or the Science Citation Database.

Even more people were involved in the UPS Prototype experiment. Paul Ginsparg at LANL,

should be mentioned first. Parallel with my work on SFX at LANL, I worked on ideas that would ultimately lead to the Open Archives initiative and its Santa Fe Convention. From the very first time we met, Paul lent me a critical but interested ear in discussions regarding future directions of his -- meanwhile legendary -- arXiv e-print system or of e-print archives in general. Paul ended up lending his name to the Call for Participation in the Open Archives initiative, and as such gave it a crucial sense of credibility. His support opened up many doors and as such also made the UPS Prototype project possible. Paul has also supported my SFX concepts and as such has enabled several cooperations. I have great respect for Paul's accomplishments and for his overwhelming integrity. My early partners in the UPS project -- Michael L. Nelson (NASA Langley Research Center) and Thomas Krichel (University of Surrey) -- deserve a very special mention, for enthusiastically jumping off the cliff with me and for being so persistent in achieving our project goals. I still find it rather incredible that we have achieved so much in only four month's time. Many thanks also to all the other researchers that actively worked on the UPS Prototype: Victor M Lyapunov at the Siberian Branch of the Russian Academy of Sciences, Kurt Maly, Mohamed Kholief, Xiaoming Liu and Mohammad Zubair at Old Dominion University and Heath O'Connell at the Stanford Linear Accelerator Centre. Many thanks also to the maintainers of the e-print archives that were involved in the Prototype, for granting us the right to use their datasets: Paul Ginsparg (arXiv.org), Stevan Harnad and Robert Tansley (CogPrints), Michael L. Nelson (NACA), Carl Lagoze (NCSTRL), Ed Fox and Anthony Atkins (NDLTD initiative), Thomas Krichel (RePEc initiative). And finally, acknowledgements also go to Richard Johnson and Alison Buckholtz at SPARC & ARL for supporting the Open Archives initiative from the very start.

Then, there is William Y. Arms, to whom I wish to express my most sincere gratitude. Bill has always been one of my heroes when it came to library automation. For many years, his groundbreaking ideas have been a fundamental inspiration in my work. It will come as no surprise that I felt both very honored and intimidated when we met at the occasion of the ISI Strategic Advisory Board meetings or at the TICER Summer School on Digital Libraries. I owe a lot to Bill in that he was the first person to make me aware of the fact that -- with SFX -- I was working on something really, not marginally, important. This may sound a little bizarre, but it is not when considering that I have been working from within a single perspective -- the Ghent library automation -- for over 17 years. Bill recognized that the SFX work reached beyond the Ghent borders and from that point onwards, he has actively supported me in many ways. The other person that is on everyone's short list when it comes to issues involved in library automation is Clifford Lynch. When starting to work on this thesis, I went back reading several of his early papers and could not feel other than intimidated by the far forward-looking vision expressed therein. At several occasions, I heard Cliff address large audiences, discussing complicated digital library matters with an amazing clarity. Cliff's work has always been a great inspiration to me. I met Cliff for the first time in person at the Open Archives meeting in Santa Fe, for which he had enthusiastically accepted my invitation to serve as a moderator. His involvement was crucial to the successful conclusion of the meeting. During the meeting, Cliff learnt about my SFX work by means of a demonstration of the UPS Prototype. Shortly after the meeting, he invited me to present the linking framework at one of the important meetings of the Coalition for Networked Information. Both Bill and Cliff have accepted the invitation to serve on the jury for this thesis. Again, I feel somehow intimidated by this. But most importantly, I feel very honored.

As a conclusion, I want to thank some people for irreplaceable support, not directly related to the academic effort presented in this thesis. Donna Berg at LANL, for giving my family and me such a warm welcome in New Mexico and for taking such good care of us while being there. My fellow countrymen Johan Bollen and his wife Cindy, Tom Delaye, his wife Yetunde Aregbe



and their daughter Leonie for the unforgettable time we spent together in Santa Fe. And, unfortunately illustrative of the sacrifices it takes to accomplish an effort like this, my last words of enormous gratitude go to my girlfriend Katrien and to our little Mo-man for always being there and for temporarily accepting to live with the selfish researcher I had to be.

# Problem Statement

---

## Linking

The creation of services linking related information entities is an area that is attracting an ever increasing interest in the ongoing development of the World Wide Web in general, and of research-related information systems in particular. Although most writings on electronic scientific communication have touted other benefits, such as the increase in communication speed, the possibility to exchange multimedia content and the absence of limitations on the length of research papers, currently both practice and theory point at linking services as being a major opportunity for improved communication of content. Publishers, subscription agents, researchers and libraries are all looking into ways to create added-value by linking related information entities, as such presenting the information within a broader context estimated to be relevant to the users of the information.

One of the first people to recognize this potential was Gardner. He expressed the desire to implement a hypertext structure linking scientific articles as a long-term goal of the electronic archive conceived by King and Roderer in 1978 (King and Roderer 1978), which he introduced to the psychology community more than a decade later (Gardner 1990). Hitchcock (Hitchcock et al. 1997a) relates the necessity of links to the associative modus operandi of the human mind. It comes as no surprise that both Gardner and Hitchcock refer to the historic writings by Vannevar Bush, in which he introduces the associative indexing (hypertext) Memex concept (Bush 1945).

But theoretical justification for linking information has become quite superfluous, since many practical illustrations of its importance have become available. Hitchcock attributes the explosive success of the World Wide Web to its linking possibilities (Hitchcock et al. 1997a). In the area of scholarly information, linking solutions have been introduced and have quickly become popular with their users. Initiatives by the Institute of Physics Publishing and BiomedNet spring to mind, where journal articles and their citations are being linked with the corresponding primary and secondary data. Ovid's linking in its Biomedical collection and its recently announced OpenLink toolkit, SilverPlatter's SilverLinker, Links between articles in HighWire Press, and ISI's Links in the Web of Science are other examples. The list of linking initiatives has grown rapidly, driven by expectations for a fully linked scholarly communication environment, created by these early linking-showcases.

## Linking in library solutions

### *The necessity of linking*

In the context of networked library services, the necessity to integrate secondary data, catalogues and primary information has been expressed quite some time ago (Evans et al. 1989; Van de Sompel 1991). More specifically, librarians have brought to the fore the need to link abstracting databases with library catalogues (Dempsey 1993; Boss 1993; Van de Sompel 1993; de Haas 1994 ; Dempsey 1995); catalogues with primary information (Van de Sompel & Van Hooydonk 1994); abstracting databases with full-text primary information (Arms 1993). These specific linking notions have evolved towards a concept of connecting all the available information, in order to come to a fully interlinked information environment (Van de Sompel 1997b). Lynch puts it this way (Lynch 1997):

*"Over time, the set of necessary linkages will expand to include not only A&I databases to*

*primary content and serials holdings and serials holdings to primary content (or, more precisely, to navigational systems for cover-to-cover content of journals, including material not in the scope for the A&I databases), but also from (monographic) catalog bibliographic records to primary content (or to finding aids that assist in the navigation of large collections of primary content) and to secondary materials such as book reviews."*

The omnipresence of the World Wide Web has raised users' expectations in this regard. When using a library solution, the expectations of a net-traveler are inspired by his hyperlinked Web-experiences. To such a user, it is not comprehensible that secondary sources, catalogues and primary sources, that are logically related, are not functionally linked (Van de Sompel 1997a).

Once implemented, such library link services become popular with the target audience and turn out to be an important aspect of integrated library services. There are indications of a strong correlation between this satisfaction and the introduction of linked electronic services. Caswell has shown this regarding the link between A&I databases and library catalogues (Caswell et al. 1995). Users' reactions to the linking experiments in the Open Journals project -- where article citations and A&I databases have been linked -- were very positive overall (Hitchcock et al. 1998b). In a survey of library users at the Los Alamos National Laboratory 30 percent of the customers were 'delighted' and the majority of the remainder 'satisfied' with the highly linked library service (Weislogel 1998). Public presentations of the evolving SFX linking service described in this thesis -- held on several occasions between November 1998 and January 2000 -- led to very positive feedback from the audience, again emphasizing the desire of users to work in a fully linked environment.

This necessity of linking information can be seen as a contemporary version of early attempts to create an integrated information environment that can easily be navigated. Some author's -- see for instance (Rayward 1994) -- trace the roots of these attempts back as far as the Belgian librarian Paul Otlet (1868-1944) and consider his life's work on the universal bibliography (Otlet 1898) as an early illustration of hypertextual navigation of scholarly information.

## **The actual situation**

### ***Static and dynamic linking approaches***

Linking mechanisms that are in use or are being developed in the scholarly information environment, can be categorized as static or dynamic. This categorization is correlated with the architectural set-up of the information collection:

- ***Static linking:*** Lately, most initiatives -- initiated by both commercial and non-commercial authorities -- have used a static linking concept. Links between information entities are computed in advance using batch processes and are held in a linking database. Typically, the processes use SICI-related information to detect relationships. Static links are used in initiatives like IOP's HyperCite, BioMednet's Bundled Links, Ovid's Biomedical Collection, ISI's Web of Science (Atkins 1999) and many other commercial linking frameworks as well as in advanced electronic library services like the Los Alamos Library Without Walls (Knudson et al. 1997; Luce 1998) and Tilburg's and Bielefeld's environments.

Records in such a database of static links describe relations between information entities that are available in the controlled environment. Static links are foolproof in the sense that following a pre-computed link will most certainly lead to the desired target. When considering solutions where bi-directional linking -- from now on called interlinking -- is

the aim, building the linking solution requires the availability of all data that needs to be interlinked under the control of the authority creating the environment. Such an authority can be a single party, in which case it has to create a self-supporting collection in order to be able to interlink it. This is the true for the examples mentioned above. Alternatively, such an authority can be a group of cooperating parties, in which case interoperability agreements between the distributed resources under control of each of the parties need to be established. This is the case for the DOI-based CrossRef initiative for reference linking (Spilka 1999b).

- *Dynamic linking*: Some interesting initiatives have started from a decentralized concept, where not all of the data that is required to build an interlinked information environment can be under the control of the authority creating the environment. As such, "a priori" computation of the links is not feasible, and linking must be done in a dynamic way, computing the links for an actual information entity "on the fly". Of special interest in this area is the work by the Multimedia Research Group of the University of Southampton, who have extensively published very valuable information on their ongoing linking implementations and experiments (Carr et al. 1995; Hitchcock et al. 1997a; Hitchcock et al. 1997b; Hitchcock et al. 1998a; Hitchcock et al. 1998b).

### ***Closed and open linking frameworks***

The frameworks that have been introduced so far feed links based on the collection that the provider of the links -- henceforth referred to as the authority -- has within its reach, and leave no room for adaptation to the environment where the links are consumed. The linking frameworks can be called "closed" or not context-sensitive. The following considerations apply for the closed linking approaches:

- *Dictated linking*: the linking solutions basically start from a presumption that includes a dictate about the target of a link. Linking from a record in an abstracting database leads to the corresponding full-text, and linking from a citation in a paper leads to a bibliographic description in a predefined database.
- *Limited range of linking*: many of the linking solutions are limited to the sphere of influence of the authority, being its collection.
- *Linking bypasses the local environment*: links are being delivered from the authority directly to the end user. The local institution where the links are used has no means to act upon the link.

## **Discussion**

The limitations related to closed linking frameworks cause serious problems. Most environments where links are consumed are hybrid libraries, made up of OPAC systems, abstracting databases, e-journals and e-editions as well as web-services. Some of the latter can hardly be classified using traditional library jargon. In this environment, a wide range of services -- that go beyond the initial aims or the possibilities of the authority -- can be delivered by creatively using the available information. The combination of an information unit that a user considers to be of interest and the entire collection that is accessible in the actual environment in which he operates can lead to the provision of a wide range of *extended services* for that information unit.

The authority can not anticipate the diversity of information that is available in the local environment. Thus, in order to deliver links that deal with the full richness of the information environment, the authority can not just autonomously define the target(s) of a link. Rather, linking should be seen as influenced by the environment where the link will be used. It should

reflect a combination of the authorities' and the consuming institutions' intentions, ultimately even the users' goals.

Although these considerations apply to both commercial and non-commercial authorities, the hindrance resulting from closed linking frameworks is most significant with commercial services that follow a strategy of vertical integration that restrict the freedom to combine information from different vendors in the same environment. In a consortium environment some libraries rely on the hosting authority for all their library services, making the local environment the same as the authority's. As such, integration can fully be dealt with by the authority. But in some consortia, participating libraries may host some information locally that is not relevant to the entire consortium, but still want it to be integrated with the whole. The concrete examples below illustrate the problem. Most apply to commercial services:

- The consuming institution might not be willing to present a link leading to a pay-per-view service, out of principle or because it holds a local copy of the paper (Bide 1997; Hellman 1998).
- The consuming institution might want to present alternative or additional link targets within its accessible environment. For instance:
  - IOP's link from a citation in an IOP published paper, to the corresponding Inspec abstract is an important service. But, the Inspec database might be available in the local environment, and the consuming institution might prefer to redirect users to the local copy, because it is linked to a local document delivery service.
  - It has been predicted that it will take about 20 years until 90% of the references in journal articles will be to papers that are in electronic form (Bide 1997 cites Norman Paskin). Thus, a link-to-holdings from a citation in a paper is an important service that institutions might want to supply in addition to the link to the abstract intended by the authority.
  - When a user's attention is drawn by a citation included in a journal paper or one found in an abstracting database, viewing the corresponding full-text might not be the only concern. The user might want to get an indication of the quality of the cited journal before deciding to read the full-text (Wang 1999). Or the user might want to look up the author's background as an alternative method of quality control. The citation might originate from a special issue on the user's actual research topic, and as such the whole table of contents of the cited issue might be relevant. The user might want to get in touch with an author of the cited publication, via e-mail.
  - A link from a citation in a journal article to the corresponding full-text should lead to the full-text in the repository that is part of the user's digital library collection. This may be (one of) the publisher's repository(ies), that of an intermediary or a local storage system in the user's institution. This problem is meanwhile known as the appropriate copy problem (Caplan & Arms 1999). If the user's institution does not hold a subscription to the journal in which the article was published, it may be elegant not to provide a link.
  - When the user has located a book in the OPAC, an abstract or book review might be welcome.
  - An authority might host only electronic secondary and primary data for a library consortium, while each of the institutions run their own Integrated Library Systems. In this case, the link-to-holdings facility depends on the local environment where it is being used.

The mainstream of the current linking approaches excludes the involvement of the consuming institution that is required to implement such services. The context of the environment in which

the information is consumed is being ignored, because the information is interlinked in a de-facto manner. The partial approaches that exist to this problem require institutions to deliver linking information to authorities running the information resources. This is done in batch procedures. Given the amount of resources controlled by different authorities that libraries deal with and given the amount of libraries that would need to deliver linking information in this manner, serious problems of scale can be predicted for such an approach (Caplan & Arms 1999).

Given the requirement to control the information collection, in order to be able to statically interlink the information, the centralized commercial solutions are restricted by the sphere of influence of the information authority. Therefore, the creation of a fully interlinked information environment -- that would result in a true one-stop shop -- would require either an information monopoly or extensive partnerships. Logical behavior by companies in the information industry tends to prevent a monopoly from happening. While the -- appealing -- CrossRef cross-publisher interlinking initiative (Spilka 1999b) and the Astrophysics Data System linking framework illustrate that linking partnerships are feasible, they do remain limited to a finite amount of cooperating parties. Again, competition in the information industry tends to prevent a one-size-fits-all solution. As such, a certain degree of dynamic linking will always be required, if only in order to cross the boundaries of a specific initiative.

In the non-commercial arena, the systems that make up digital library environments can be under local control, as is typically the case with OPAC and some secondary data systems. Increasingly, systems are under technical control of an external authority, such as a database vendor, a subscription agent, a publisher, and another library. The non-commercial parties -- libraries and consortia -- are in a good position to build integrated services, since they are not copyright owners. As such, they are neutral enough to potentially receive a green light from a wide variety of information vendors, to integrate and interlink their data-collections. Because of the increasingly distributed nature of the digital library collections, static interlinking solutions -- that require the local availability of all data -- will most probably be excluded. In addition, it is very hard to imagine how a wide range of context-sensitive services mentioned before could at all be provided in a static manner. Hence, in digital library environments, linking tends toward a dynamic approach.

## **The SFX problem statement**

From the above, it can be concluded that there is a genuine need for a dynamic, context-sensitive linking framework to interconnect distributed scholarly information resources. By reporting on the Special Effects -- SFX -- research, this thesis will show that such a context-sensitive dynamic linking framework is feasible and it will also identify its major characteristics.

But first, a terse formulation of the problem statement of the SFX research is presented, as well as some initial considerations on the problems in designing a solution.

The problem statement of the SFX research can be expressed in a comprehensive manner by building on the terminology that was used in the context of the meetings and the subsequent reports and publications on reference linking organized by the Digital Library Federation (DLF), the National Information Standards Organization (NISO), the National Federation of Abstracting and Indexing Services (NFAIS), and the Society for Scholarly Publishing (SSP) (Caplan 1999a; Caplan 1999b; Caplan & Arms 1999; Needleman 1999).

The generic statement of the reference linking problem, as defined by the working group on

reference linking was (Caplan 1999a; Caplan & Arms 1999):

*Given the information in a standard citation, how does one get to the thing to which it refers?*

However, the working group concentrated on a specific variation on this:

*Given the information in a citation to a journal article, how does a user get from the citation to an appropriate copy of the article?*

The SFX research also addresses these problems, but only as an instance of a more general problem that can be formulated as:

*Given bibliographic metadata, how does one present relevant extended services for it?*

### ***Bibliographic metadata as a starting point***

Clearly, the SFX research is not only concerned about information in a standard citation. Its starting point is bibliographic metadata in general. As such, information entities originating from typical scholarly resources such as records from abstracting & indexing databases, OPAC systems and e-print archives can be used as a starting point in the SFX problem statement. This is also the case for citations to both journal articles and books found in journal articles or books. But -- in principle -- even fractional bibliographic metadata such as an author's name taken from an e-mail message could be a valid starting point in the SFX problem statement.

### ***Extended services as a goal***

A similar generalization holds for the target of the problem statement since the SFX research is not only concerned about linking to the full-text that corresponds to a citation in a journal article. It aims at the presentation of a variety of extended services for whichever metadata is used as a starting point. Extended services are services that present an information entity in a digital library -- defined as the link-source -- in the context of the entire information environment. For instance, for a given link-source record from an abstracting & indexing database, extended services can -- amongst others -- be the presentation of:

- The full-text of the paper that is abstracted in the link-source;
- A record abstracting the same publication taken from another abstracting & indexing database;
- Citation information corresponding with the link-source;
- Library holdings for the journal in which the article described by the link-source appeared.

### ***Relevant extended services as a goal***

The word relevant is of particular importance in the SFX problem statement. It actually refers to two separate notions:

- Relevance of a service in the context of the user's digital library collection, in other words, context-sensitiveness of a service;
- Relevance of a service as opposed to irrelevant in every context. For instance, it is always irrelevant to provide a book review service if the link-source refers to a journal article.

## **Initial design considerations regarding a solution**

As shown before, due to the increasingly distributed nature of the information collection at hand, a dynamic linking approach or at least some combination of static and dynamic linking is the most realistic path leading towards a solution to the problem statement. Moreover, the desire to act upon information units that are being provided by an authority -- in order to be able to

deliver context-sensitive extended services -- calls for an open linking framework that is not in place. As such, the following are important challenges in designing a solution to the problem:

- *Grabbing a link-source item:* in order to be able to present locally defined links for a certain information unit (the link-source) originating from an authority, it is necessary to identify, capture and analyze the unit in the local environment first. When source systems are under local control, the required system enhancements can be dealt with internally, using ad-hoc techniques. When source systems are under external control, grabbing the link-source can become a very cumbersome task. Complex proxying and parsing solutions have been introduced to deal with this problem (Hitchcock et al. 1997a). Eventually, both situations should be handled via the same generic open linking framework. But in the absence of it, finding techniques to grab link-sources presents a major challenge in dynamic linking solutions.
- *Link verification:* inherent to dynamic linking approaches is the uncertainty regarding the success of a link that has been created on-the-fly. Depending on the protocol supported by the linked system, links can be verified before delivery or not.
- *Data-processing delay:* the dynamic approach to linking causes processing delays when servicing links. Lynch anticipated this problem for link verification in a distributed environment (Lynch 1997) and designers of the Open Journals Project (Hitchcock et al. 1997a) have confirmed the problem in the operational context of citation linking. Later, delay in response times was mentioned as one of the few criticisms by users of the Open Journals test system (Hitchcock et al. 1998b). In digital library environments, the number of information units that are being transferred daily can be very high. For each of these units, delivery of extended services will introduce certain delays. Therefore, in the design of a linking solution, processing delay must be an important concern.
- *Locally hosted linking service:* the multitude of heterogeneous information systems that should be interlinked, calls for a linking service that can be shared amongst systems (Carr et al. 1995; Pearl 1989). Such a linking service provides a look up in a database where data items are interpreted as links. Since the consuming institution is in the unique position to know its complete interlinkable collection, it should host and (co)-feed the linking service. The early Ghent linking experiments confirmed the necessity for a linking service in an empirical manner. These experiments required link-specific enhancements to be made to each of the systems where links originated. It was anticipated that such an approach would soon lead to a maintenance overhead of system enhancements.
- *Link-to-services:* in order to be able to link into a system, it must provide a link-to-service, that can be addressed using a published link-to-syntax. For instance, most of the actual Integrated Library Systems provide a syntax for a link-to-holding facility. Linking into secondary services, such as A&I or citation databases, has little been dealt with so far, and it comes as no surprise that real linking services are rare in that area. PubMed's Entrez link-to solution is a very noteworthy exception. Increasingly, primary publishers and intermediates have made available genuine link-to-services, that can be used when jumping from A&I services or OPACs into their full-text collections:
  - Academic Press <http://www.idealibrary.com/help/links.jsp>
  - American Physical Society <http://publish.aps.org/linkfaq.html>
  - SwetsNet <http://www.swets.nl/press/may982.html>
  - Elsevier ScienceDirect  
[http://www.sciencedirect.com/science/page/static/splash\\_pr9.html](http://www.sciencedirect.com/science/page/static/splash_pr9.html)
  - UMI SiteBuilder <http://www.umi.com/builder>



But with many publishers that have online content, such services are still not supported. Careful examination of their URL structures may lead to insights that can help when trying to link into their collections. Still, there is no overall uniformity in the approaches taken, and linking can become very complicated due to authentication issues, the level(s) of the links that can be created (journal level, publication year level, volume level, issue level, article level), the information required to create the links etc... Again, a generic framework, accepted by the scholarly publishing community would be most welcome. The S-Link-S initiative (Hellman 1998) should be seen as a feasible proposal.

- *Licensing and consortia*: the presentation of links to end-users is dependent on licensing and subscription boundaries that apply within the collection. In a consortium environment, where different parties have access to different information sources via the same service, this can turn interlinking of the sources into a quite complex matter.

# Experiments and Concepts

<b>Part 1</b>	Initial identification of characteristics of the SFX linking solution in the Elektron experiment
<b>Part 2</b>	Generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment
<b>Part 3</b>	Applying SFX to integrate e-prints with the established scholarly communication system in the UPS Prototype experiment

# Part 1: Initial identification of characteristics of the SFX linking solution in the Elektron experiment

---

## Introduction

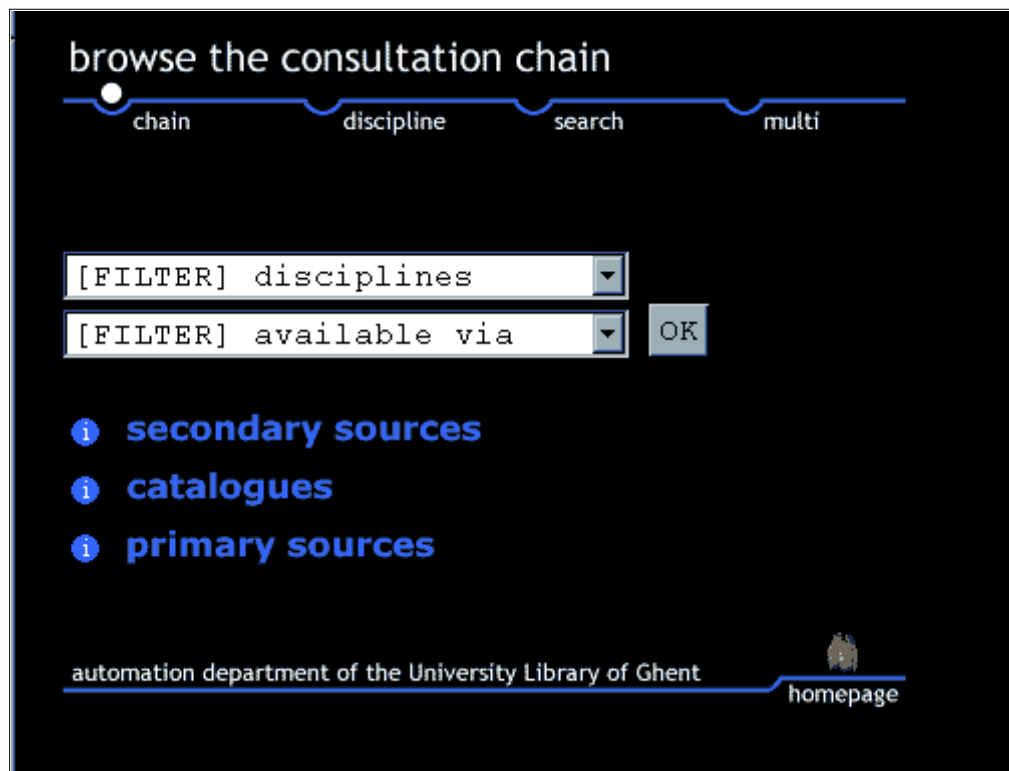
This first Part gives a description of the approach towards the creation of extended services in a digital library environment as taken in the Elektron experiment, which is the first of three SFX experiments being reported on in this thesis. In order to explain the SFX-concepts in a comprehensive way, the discussion starts with a description of the context in which the experiment was conducted. Thereafter, the basics of the SFX-approach are explained, in combination with implementation details of the Elektron SFX-linking experiment. Elektron was the name of a modest digital library collaboration between the Universities of Ghent, Louvain and Antwerp.

The Elektron experiment has been conducted within the boundaries of a subset of the digital library collection of the University of Ghent (Belgium), and has involved a limited amount of parties from the information industry. There was active involvement of the Ghent library automation team. The experiment was conducted between August 1998 and November 1998.

## The working environment for the Elektron SFX experiment

The University of Ghent subscribes to a wide variety of electronic information services. They include SilverPlatter's Electronic Reference Library solution (ERL) and ExLibris's Aleph 500 Integrated Library System both of which are important local building blocks. The ERL server hosts a wide variety of mainly abstracting & indexing data (70+ Gb), while the Aleph system hosts the local catalog (500,000+ bibliographic records). ISI's Web of Science has been added in the course of 1999, but was not part of the Elektron linking experiment. The environment also provides access to a collection of about 300 e-editions of scientific journals that are available without additional charge as part of the institutional paper-based subscription. Amongst those, the Springer, Wiley, HighWire, Institute of Physics and American Physical Society collections are the most noteworthy. For the Elektron SFX-experiment described below, temporary access to the Academic Press, the UMI Business Periodicals Online and the Blackwell Science collections was granted.

The environment is presented to end-users via a web-based menu-system called the Executive Lounge, which is an easy-to-use interface to the database of databases (Figure 1). The Executive Lounge menu items point at both the traditional library related sources (typically networked databases and full-text collections) and a limited number of websites with academic relevance. Upon a user's request, menu items can be presented in different views: by data-type (secondary sources i.e. abstracting & indexing databases, catalogs, primary sources); by discipline (humanities, medicine, engineering, etc.); via a menu item search screen; via a display presenting only menu items that can be searched simultaneously. For instance, in the data-type view, the menu-header "secondary sources" gives access to Current Contents as well as to the major Internet search engines. A reference to most of the e-Lib subject-based gateways will be found under the same header. The menu-header "catalogues" points at several important Belgian library catalogs, as well as at a catalog of electronic journals and important Internet bookstores. The menu-header "primary sources" points at established publishers' e-editions as well as at a selection of free Internet e-journals.



**Figure 1: the Executive Lounge interface**

## **Pre SFX-linking experience**

The Ghent library automation group has been actively involved in reference linking for several years:

- Several early papers identified the need to integrate a variety of electronic library resources in general (Van de Sompel 1991) and to link between abstracting & indexing data, catalogs and primary data in particular (Van de Sompel 1993; Van de Sompel & Van Hooydonk 1994).
- A link-to-holdings between the SilverPlatter ERL and the Aleph 500 system was created as soon as the Aleph system went into production [June 1997]. The implementation of this link was a basic requirement, expressed explicitly in the tender for a new library system [1995]. The decision to acquire a new library system was strongly inspired by the desire for integration. Eventually, the link-to-holdings implementation led to the general availability of a link-to-holdings feature in SilverPlatter's WebSPIRS release 4 and in the Aleph 500 system [1998].
- A link between the Inspec database on SilverPlatter's ERL and the IEEE electronic library collection has been implemented on behalf of the IMEC engineering research institute. This was a joint effort of the Belgian SilverPlatter distributor IVS, IMEC, and the Ghent library-automation group [fourth quarter 1997].
- Experiments have been conducted linking from SilverPlatter's ERL databases to the full-text collection available via SwetsNet [mid 1997]. This led both to the availability of a general link-to-syntax for SwetsNet and the inclusion of SwetsNet in SilverPlatter's SilverLinker solution [end 1997] (Hamilton 1998).
- Experiments have been conducted to link from SilverPlatter's ABI/Inform to UMI's Business Periodicals Online collection hosted on the ProQuest Direct service [fourth quarter 1998]. These experiments have been facilitated by the availability of the UMI SiteBuilder link-to-syntax.

## Concepts introduced in the course of the Elektron SFX experiment

As explained in the Problem Statement, the fundamental aim of the SFX linking research is to develop a framework that enables the presentation of an information entity of a digital library collection -- from now on referred to as the link-source -- in the context of the complete collection. This means that SFX wants to provide service links in a context-sensitive manner: the target(s) of a link is (are) seen as a combination of the information providers' and the libraries' intentions. Moreover, in order to achieve its goal, SFX realistically has to provide such links in a dynamic manner: it can not rely on a static database of links that are computed in advance, because that would require full control of the whole information environment that is to be interlinked.

An overview of the design introduced in the Elektron experiment is shown in Figure 2 and is expanded in the following sections.

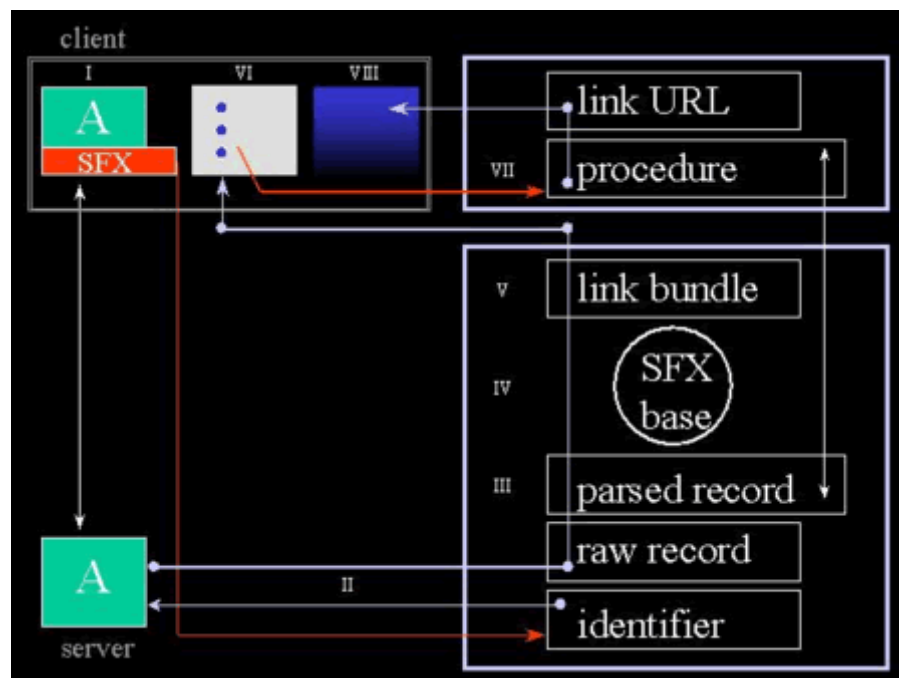


Figure 2: the SFX mechanism

### *SFX, linking from ... to ...*

Link-sources in the digital library environment can be records from OPAC systems, from abstracting and indexing databases (A&I), the bibliographic information of a full-text paper as well as each of its citations. The targets of the links reside in the same type of systems.

The Elektron SFX experiment has concentrated on the link-sources and the link-targets that are shown in Table 1. This set of service-links has been chosen as a research domain because it contains service-types that have hardly been investigated. More importantly, this choice restricts the problem to systems where the link-sources are under local control. Although this choice might seem to limit the scope of the research, it has allowed the development of solutions to grab link-sources other than proxying, and to concentrate on other aspects of linking that are equally important.

SOURCE				
A & I databases	yes	yes	yes	yes
OPAC	yes	yes	yes	yes
primary collection	no	no	no	no
other web info	no	no	no	no
	A & I database	OPAC	primary collection	other web info
TARGET				

**Table 1: SFX linking from-to**

### ***The SFX linking service***

The SFX server is introduced in the environment (see the rectangle at the bottom-right of Figure 2). It is a linking service that provides a look up in a database where data items are interpreted as links and that is to be shared by all other systems in the environment (Carr et al. 1995; Pearl 1989).

### ***The Colli: a collection of anticipated conceptual links***

Static linking solutions are not considered in SFX-linking. Therefore, the SFX linking service will not build on a database containing hardwired links between the data that is involved. Instead, the notion of a collection of anticipated conceptual links (Colli) that the library wants to make available to its users, is introduced. The organization of the Colli is based on the feasibility of actually creating the link at some further stage in the process (i.e. existence of a link-to-service) and in anticipation of users' expectations. Each of the conceptual links is introduced in order to provide a certain service that is thought to be valuable for users of the system.

Each of the anticipated links in the Colli is accorded a name that corresponds to a procedure designed to resolve the link-to-syntax using parameters extracted from the link-source. The links that have been introduced for the Elektron experiment, are shown in Table 2.

There are 3 links to OPAC systems that are important for interlibrary loan: the Ghent Aleph 500 system, the Belgian union catalogue of serials, and the Dobis/Libis system at the University of Louvain. There are links to abstracting & indexing databases, such as L-BIP, which is intended to look for the record in Books in Print that corresponds to the link-source. L-ULRICH is similar, with links into Ulrich's Serials Directory. L-JCR looks up Journal Citation Reports data for the link-source; thus it provides the user with ISI's notion of the quality of the referenced journal. L-CC is intended to bring up the table of contents (including abstracts) from the Current Contents database, for the issue of the journal that is referred to by the link-source. There are several links to primary information collections, whose names are self-explanatory. Finally, the L-AMAZON link leaves the typical academic information environment, and searches for the book referred to by the link-source, in order to present the user with book reviews and ordering information.

All conceptual links and related procedures are seen as being independent of:

- The nature of the link-source: Abstract and Indexing record, OPAC record, citation in full-text paper, bibliographic data of a full-text document.
- The physical location of the system where the link-source originates from: local or remote.

the COLLI		
type	link name	links to
to OPAC systems	L-ALEPH	University of Ghent OPAC
	L-ANTILOPE	Belgian union catalogue of serials
	L-LIBIS	University of Louvain OPAC
to A & I databases	L-BIP	Books in Print
	L-ULRICH	Ulrich's International Periodicals Directory
	L-JCR	ISI's Journal Citation Reports
	L-CC	ISI's Current Contents
to primary information	L-SWETS	SwetsNet collection
	L-SPRINGER	Springer full-text collection
	L-ACADEMIC	Academic Press full-text collection
	L-BPO	UMI Business Periodicals Online collection
to others	L-AMAZON	Amazon.com online bookstore

**Table 2: the Colli in the Elektron SFX-experiment**

### ***The SFX-button: just-in-time linking***

SFX takes a "just-in-time" instead of a "just-in-case" approach to linking. When information is presented to the user, potential link-sources are marked with an SFX button. As a means of reducing delays, no links are computed until requested by the user. For each link-source, an identifier is hidden behind an SFX-button (see I in Figure 2). This identifier holds the following information:

- ID of the server from which the link-source originates
- database ID of the database where the link-source originates
- unique record ID of the link-source within that database
- the SFX server process that is executed by clicking this button

A user must explicitly request links for a link-source by clicking the SFX-button. Clicking transfers the identifier to the local target that uses it to fetch the link-source into its environment (see II in Figure 2). The ID of the server not only gives information on its location, but also on the protocol to be used to grab the link-source. In the case of OPAC or Abstracting and Indexing databases, this might be Z39.50. But it might also be a Lightweight Directory Assistance Protocol (LDAP) look up, a Handle resolution, or an http link. Next, the document is parsed into a generic format and essential parameters are extracted (see III in Figure 2). All information is kept at the server-side, in relation to the users' session-ID. The system is now ready to start the next phase in the process: the conceptual verification of potential links from the Colli.

The "just-in-time" approach, requiring an explicit user action to request links, seems to be justified by the following:

- The link-to-holdings feature connecting Abstracting and Indexing databases with the local OPAC in the Ghent environment also requires an explicit user action. Logs show that the holdings button is being used for approximately 3.3 % of the records that are being transferred, meaning

that the link remains idle for 96.7 % of the records. Similar statistics can be expected in the broader context of SFX-linking, if only because end-user searches in large databases are typically done with a low accuracy (Bates 1998) and because links for records that look irrelevant to a user are not likely to be followed. As such, "just-in-time" linking can dramatically reduce response times by only going through the required overhead, when necessary.

- Since SFX intends to serve a bundle of links to the user for each link-source, a "just-in-case" approach, instantly feeding all links for each link-source, would inevitably lead to user interface problems.
- The explicit user action identifies records that the user considers relevant. The accumulation of such information -- in combination with search strategies -- can, in the long-term, lead to a database that supports a recommendation system.

SFX-linking approaches the problem of grabbing the link-source by introducing a clickable identifier, containing a small data record, for each link-source. The technique is identical for all systems involved. It is recognized that the implementation of this solution was simplified by the fact that the originating servers used in the experiments were under local control. Both providers of the local systems in Ghent -- ExLibris and SilverPlatter -- have enabled its straightforward realization. Still, the concept is quite generic, and could also be implemented with systems under remote control, to create a general purpose "just-in-time" linking solution.

For instance, in the case of the Open Journal Project, journal papers are proxied, their citations are parsed and citation-links are inserted on-the-fly before delivery to the user. Many of the complexities involved with linking from these citations could be postponed to a later phase in the process, by initially only identifying link-sources in the HTML or PDF documents and inserting, respectively, SFX-anchors or SFX-named-destinations as unique link-source IDs. Storing the enhanced document in the server environment and simultaneously sending it to the user would create a set-up in which the link-source could be retrieved and processed only upon the user's request.

Proxying should be considered to be the hard way to grab the link-source. There are definitely scaling problems related to such an approach when dealing with a highly distributed collection. But more importantly, proxying does not require the involvement of the authority. At a first glance, this may seem to be an attractive feature. It would allow to build interlinking solutions without requiring any interoperability agreements. Still, not having the slightest involvement of the authority is synonym to not having any guarantees regarding the longevity of a solution. In order to build a sustainable solution, a certain level of cooperation with the authority is desirable. It can lead to more straightforward solutions to grab the link-source. One can imagine a situation where the authority inserts the required identifier along with the appropriate address of an institutional SFX-server on a subscription basis. Although this may sound like wishful thinking, such a possibility is almost inherent to the DOI concept, on the condition that the resolution of a DOI can be redirected to a local target rather than to the DOI handle system. In that case, an institutional SFX-server could retrieve the link-source corresponding with the DOI-identifier from a DOI directory.

### ***Conceptual verification of links from the Colli via the SFX-base***

Since there are no "a priori" computed links in this environment, there is no initial certainty on the relevance of a specific conceptual link from the Colli for a specific link-source. Meanwhile, that link-source resides in a parsed format in the server's environment (see III in Figure 2). In order to prevent irrelevant links from being presented to the user, the SFX-base is introduced (see IV in Figure 2). The SFX-base describes the relationship between the conceptual links from the Colli and the parameter values of link-sources for which the conceptual links are valid. Matching parameters of a link-source with the SFX-base filters out irrelevant links. The matching process fulfills a conceptual verification for each of the links from the Colli. Once a link has been selected in this process, it will be included in the bundle of links that will be presented to the user (see V in Figure 2).

It should be emphasized that this selection does not guarantee the success involved in following the link, at a later stage. The conceptual verification minimizes the amount of predictable failures. For

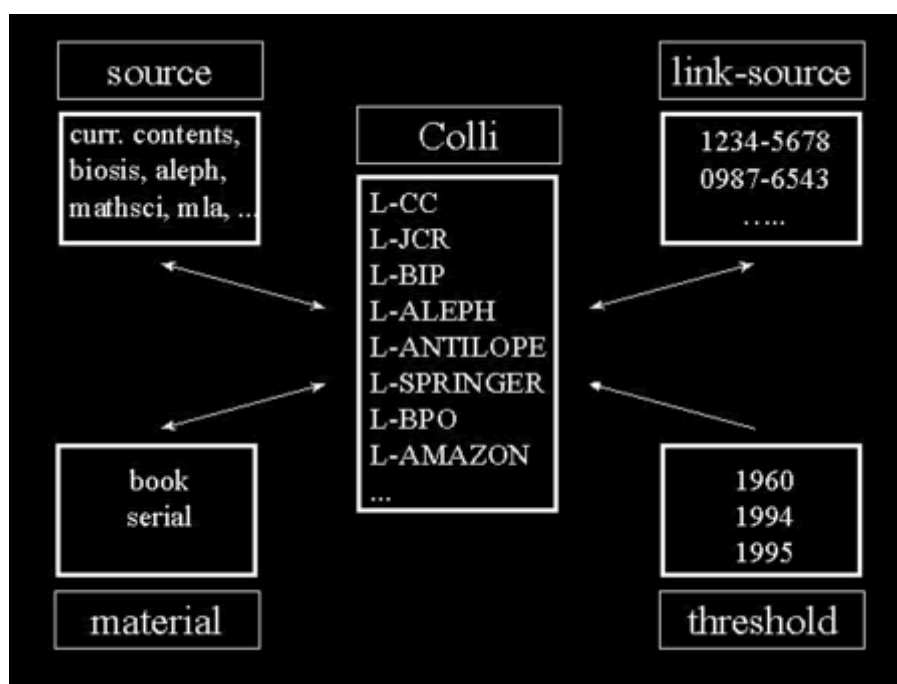


instance, when the active document refers to a journal article, the anticipated link to Amazon.com will be filtered out. When the active document originates from the Current Contents database or when the journal referred to by the active document is not indexed in Current Contents, the L-CC link will receive a negative flag. A link to Springer will only be selected when the active document refers to a paper published in a Springer journal with a publication year that makes electronic availability near to certain.

A limited number of parameters have been defined for the SFX-base of the Elektron experiment:

- Material type of the document described by the link-source: restricted to books and serials;
- ISSN number of the document described by the link-source;
- Threshold for the publication year of the document described by the link-source, beyond which a certain link becomes relevant;
- ID of the database where the link-source originates: in the Elektron implementation these are the names of databases installed on local systems. Extension of the service to link-sources from primary collections, would introduce additional IDs, probably referring to publishers' collections and/or ISSN numbers.

This information is brought together in a relational database, with the Colli as a central table (Figure 3). In addition to the described parameters, a link-type table is added to the SFX-base, which allows the presentation of the relevant links in a structured way, corresponding to the classification made in Table 2, reflecting the organization of the database of databases in the Executive Lounge menu-system (Figure 1).



**Figure 3: the SFX-base**

A simplified overview of the contents of the SFX-base used in the Elektron experiment is given in Table 3.

link name	material type	threshold year	source dbase id	ISSN
L-ALEPH	all	all	all except ALEPH	all
L-ANTILOPE	serials	all	all except ANTILOPE	all
L-LIBIS	all	all	all except LIBIS	all
L-BIP	books	> 1970	all except BIP	none
L-ULRICH	serials	all	all except ULRICH	all
L-JCR	serials	all	all	Only journals evaluated in JCR
L-CC	serials	>= 1996	all except CC	only journals abstracted in CC
L-SWETS	serials	>= 1997	all	ISSN numbers of Blackwell journals
L-SPRINGER	serials	>= 1997	all	ISSN numbers of Springer journals
L-ACADEMIC	serials	>= 1996	all	ISSN numbers of Academic Press journals
L-BPO	serials	>= 1997	all	ISSN numbers of UMI BPO journals
L-AMAZON	books	> 1970	all	none

**Table 3: content of the SFX-base**

It is obvious that the design of the SFX-base requires fine-tuning in order to become a production system, but for a first experimental set-up, a certain roughness has been tolerated:

- The "threshold year" parameter gives cause for some degree of uncertainty:
  - The threshold values for L-BIP and L-AMAZON are quite arbitrary.
  - The one for L-CC is more exact, since it reflects the starting year of the Current Contents database that is available for look up. Still, the 1996 issues of Current Contents can contain records referring to papers published in 1995. Using the proposed threshold filters out the L-CC link for those records. Bringing the threshold down to 1995 would conceptually select all records with a publication year starting in 1995, the majority of which would not be covered by the available collection of Current Contents.
  - The threshold values for the full-text links refer to the earliest publication year for which the linked publisher has online content available. However, in many cases the electronic starting point varies for different journals of the same publisher. This calls for the introduction of a new table connected to the ISSN table, containing information on

publication year, volume, and issue of the first electronic edition.

- For now, the SFX-base has only been fed with information on databases and journals with current subscriptions. In order to come to a more generic design that would be applicable in a consortium environment, there is definitely a need to include subscription information in the design, both in connection with the database-ID and the ISSN table. This might call for integration with the serials module of the Integrated Library Systems that are involved.
- Since the scope of databases changes over time, limiting links into abstracting & indexing databases to link-sources that have ISSN numbers of journals indexed in those databases, without involving a time-concept, is not fully waterproof.

### ***The SFX-screen: a bundle of unresolved, functionally unverified links***

The result of the conceptual verification process is a buffer of potential link-names, corresponding to links from the Colli that are relevant for the current link-source. The link-names in the buffer are organized in accordance with the classification shown in Table 2, and delivered to the user in a separate browser window (see VI in Figure 2 ; see Figure 5 , Figure 7 and Figure 9). Following the same argument that led to justification of just-in-time linking, at this stage links are still not resolved. The potential links are sent to the user, with the link procedure names as parameters. A server based link-resolution process that will be activated when the user chooses to follow a certain link will use these parameters. At that point, the essential information from the actual document is retrieved from the copy of the link-source that is held at the server's side. Next, this information is fed to the chosen procedure in order to resolve the link (see VII in Figure 2). Finally, the user is redirected to the appropriate location (see VIII in Figure 2 ; Figure 6, Figure 8 and Figure 10).

As a consequence of this approach, the links in the SFX-screen are not functionally verified, and following them may lead to empty results. This design option is subject to some considerations:

- Many of the links presented in the SFX-screen should be interpreted as alternative search features, rather than foolproof links. Hence, using them is subject to all the characteristics of searching, including empty result sets, abundant result sets, serendipity, etc.
- Conceptual verification of links has preceded the current phase, as such minimizing the amount of irrelevant links. Functional verification of each link would not only cause significant delays, it might even turn out to be impossible, when not supported by the linked system.

Exploiting this approach and properly designing the procedures to resolve the links can lead to features that are appealing instead of frustrating to end-users, as can be seen from the following examples:

- The procedures try to resolve links as accurately as possible, given the number of parameters that are available in the link-source. Typically, linking from a record in an abstracting database enables the extraction of ISSN, publication year, volume, issue and page information. Depending on the accuracy of the link-to-syntax provided by the primary publisher's system, this can lead directly to the full-text of the referred paper. This is rarely the case, since the best that most link-to-syntaxes enable is linking to the appropriate table of contents, from which a link to the full-text can be chosen. This is not necessarily a disadvantage, since it brings some serendipity into the mechanism.
- Quite frequently, not all of the parameters defined for a linking procedure are available. Either it is impossible to extract them from the link-source, or the data is just not there. Typically, as is the case in many libraries in Europe, in the Ghent environment OPAC records do not contain volume and issue information for serials. Therefore, the procedures have been designed to make the best use of the information that is available and to lead the user as close as possible to the goal intended by the chosen link. If issue information is missing, the procedure tries to construct a URL to the level of a volume; if volume is missing too, a URL to the publication year of the reference will be generated; if that is missing too, a link to all electronically available years for the cited journal is the solution. This approach has been turned into an appealing interface feature. For those procedures that require more than just ISSN or ISBN information for full

resolution, the remaining parameters are displayed in editable text boxes when the link is presented to the user. Thus, the user can change or complete the form. This feature is especially relevant when working from the OPAC, linking into Current Contents or a full-text collection (Figure 5 and Figure 6).

- UMI's SiteBuilder link-to syntax does not allow use of publication year, volume nor issue as search terms, and therefore the L-BPO procedure has been designed to search for a combination of an ISSN number and title words instead. Although this might be seen as far from optimal linking, it can lead to pleasant surprises in the search results.

## Illustrations of project results

The concrete results of the project are illustrated by means of screendumps that show how a user requests extended services when navigating the digital library collection available in the context of the Elektron SFX experiment as well as the type of services that are being provided by the SFX-server.

Figure 4 shows a link-source from the Aleph 500 catalogue system of the University of Ghent, describing a serial. At the bottom of the record, the Special Effects button that is dynamically inserted is visible. Figure 5 shows the services that become available when the user chooses to click the button. The services are organized in the same way as menu items in the Executive Lounge are. Under the item "secondary sources", a service linking into Current Contents is presented, which has survived the conceptual verification phase, because the SFX-base has revealed that the serial is indexed in Current Contents. Since the OPAC record does not contain volume or issue information, the service is presented with empty fill-out boxes. Figure 6 shows that the following this service-link, resulted in the appropriate results in Current Contents. Also available under the same menu-heading are services linking into the Journal Citation Reports and Ulrich's Serials Directory. There are also services linking into other catalogues: one links to Antilope -- the Belgian Union Catalogue of Serials --, the other to the OPAC system of the University of Louvain. No full-text links are available for the given link-source. Figure 7 shows the SFX menu-screen generated dynamically upon request of a user, for a link-source -- again originating from the Aleph 500 system -- which describes a book. Services pointing to Books in Print, Amazon.com and to the Louvain OPAC system are visible. Interestingly, a link to Antilope is not provided, since Antilope does not contain book descriptions. Figure 8 shows that the user chose to follow the Amazon link. Figure 9 shows the SFX-menu presented to a user after clicking the Special Effects button for a record originating from the EconLit database, describing a journal article. Services similar to the ones shown in Figure 5 are available. In addition, a link to full-text at Academic Press is now available. Figure 10 shows that the user preferred to check the Journal Citation Reports to get an indication about the quality of the cited journal.

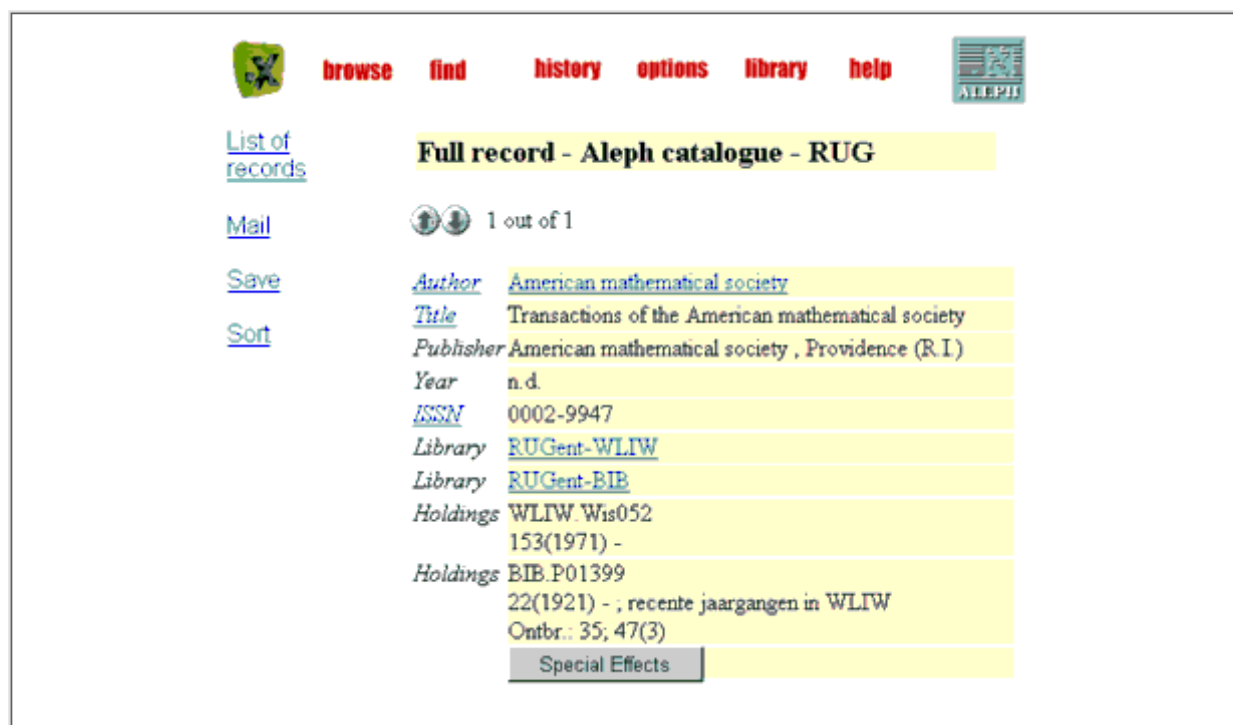


Figure 4: an OPAC serials record as a link-source

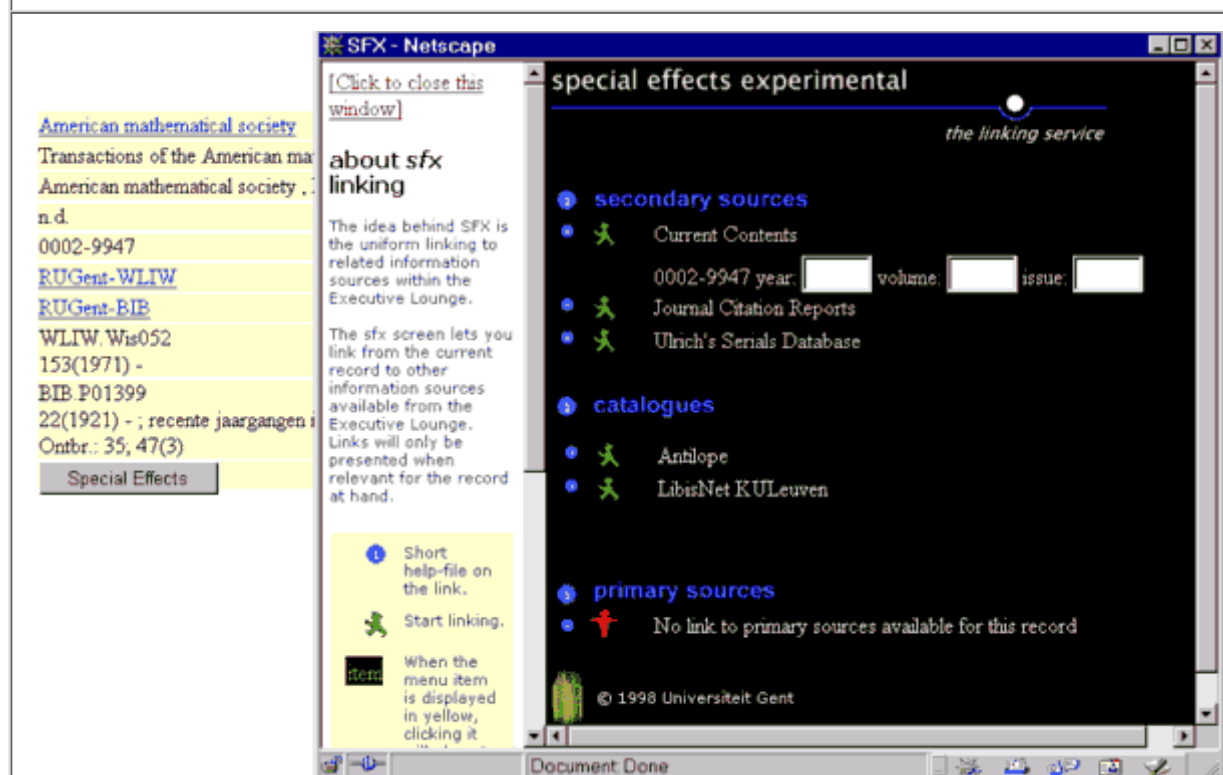


Figure 5: the SFX screen for the OPAC serials record



Figure 6: a SFX-link to Current Contents for the OPAC serials record

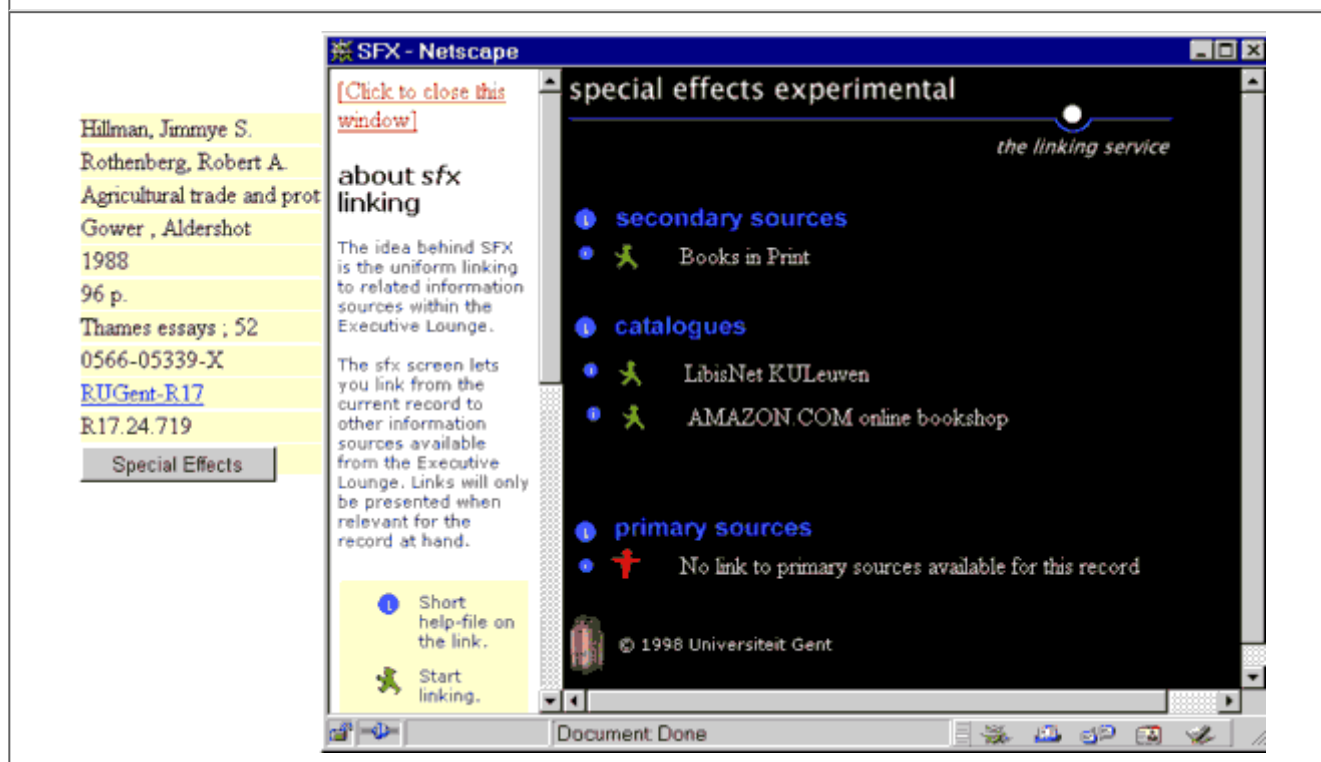


Figure 7: the SFX screen for an OPAC book record



Figure 8: the SFX-link to Amazon.com followed for the OPAC book record

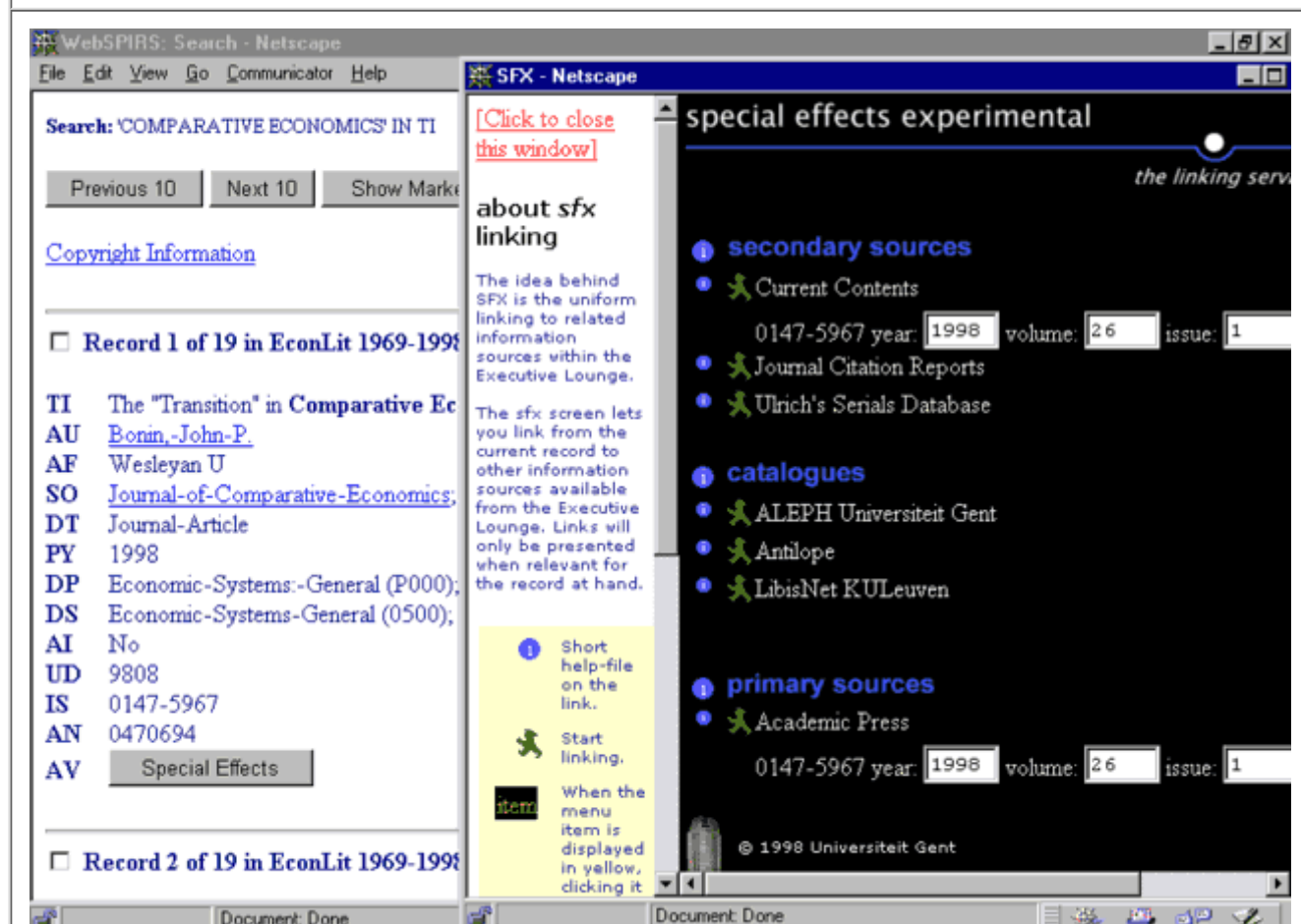
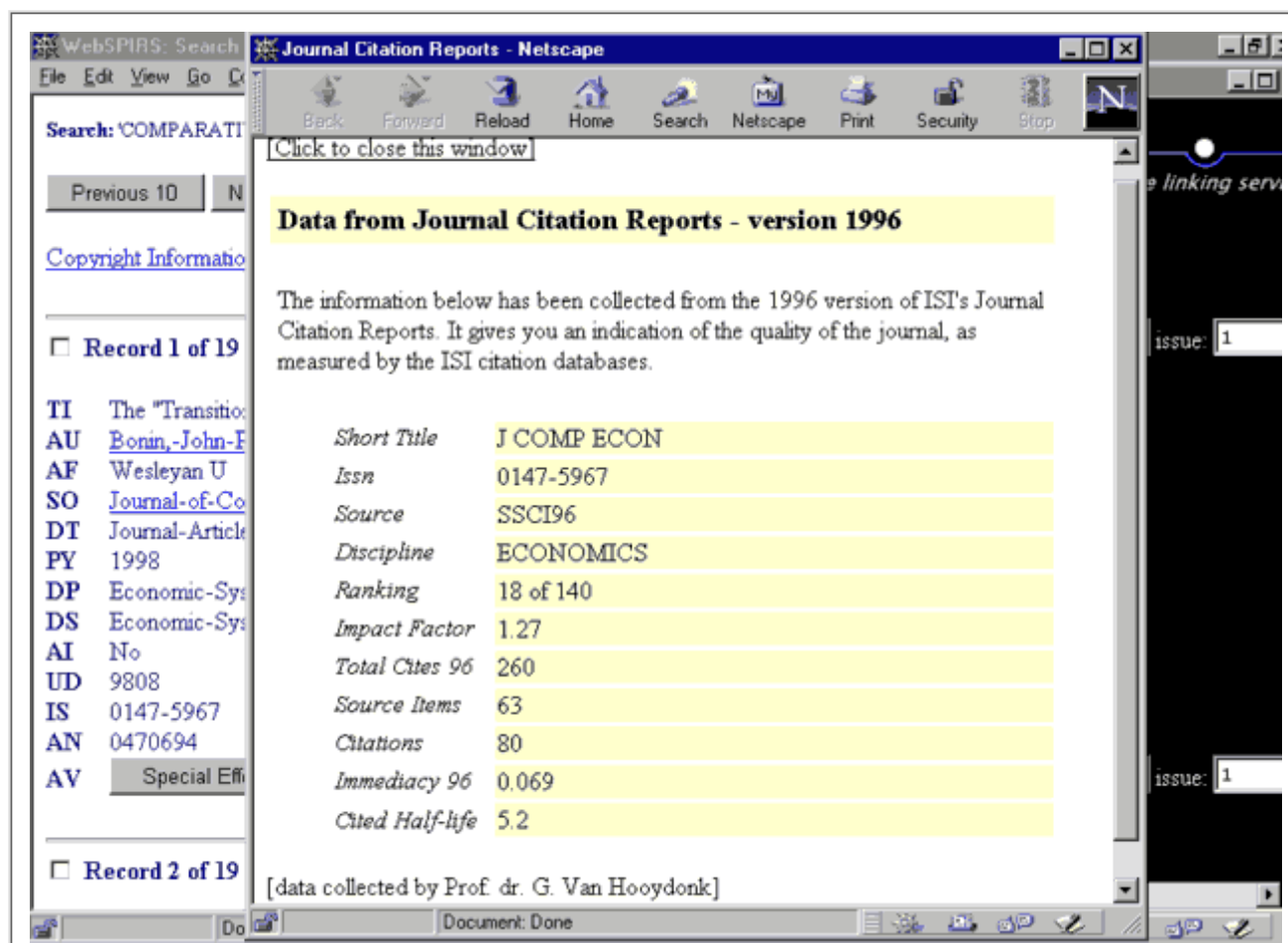


Figure 9: the SFX screen for a record from the EconLit database





**Figure 10: the SFX-link to Journal Citation Reports followed for the record from EconLit**

## Intermediate conclusions

Minimally, the Elektron experiment has justified the claim for an open linking framework that provides links as a combination of the information providers' and the hybrid libraries' aims. The Elektron SFX experiment has also provided a preliminary demonstration of the feasibility to deliver extended services in a digital library environment in a dynamic way, via the SFX approach. Given the fact that all resources from which link-sources originated were part of the local environment where the experiment was conducted, there is no convincing proof of the feasibility of an open, context-sensitive linking framework: the context was restricted to the local authority and as such was linking was inherently context-sensitive. Still, the experiment provided interesting insights. They will be summarized in the remainder of this section.

### Grabbing the link-source

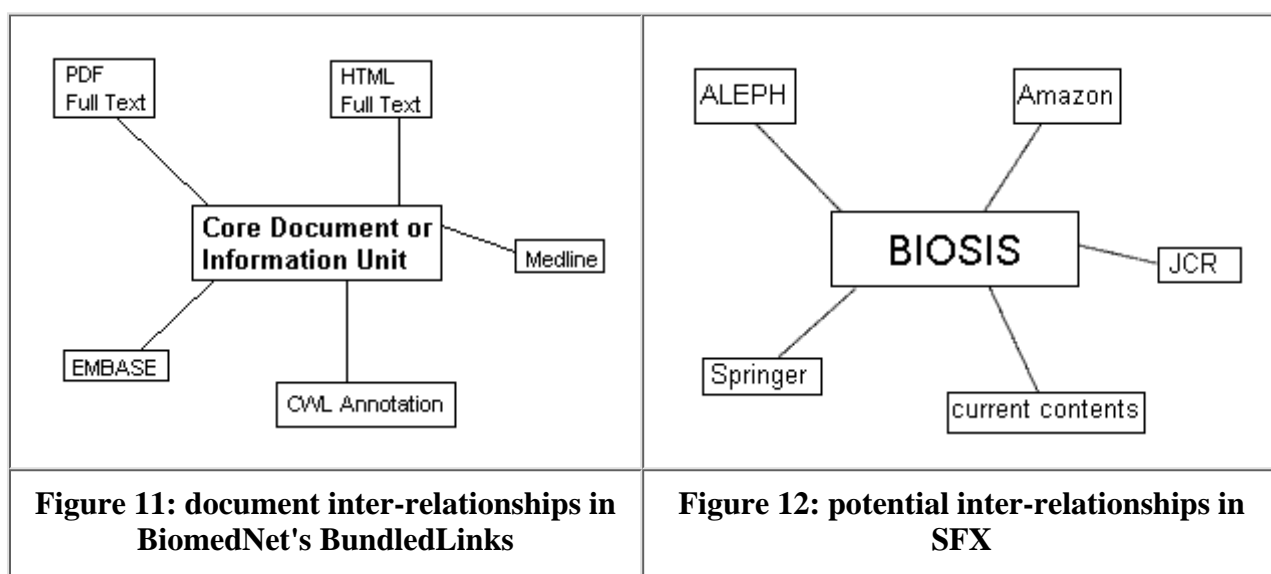
Even if the Elektron SFX experiment was restricted to the local environment, it did introduce a mechanism to grab the link-source that offers promises with regard to the feasibility of a generic approach to open linking. This mechanism builds on the introduction of the SFX-button that transports an identifier of a link-source to the SFX-server. This enables the SFX-server to fetch the link-source from its origin database. As such, the link-source itself is transported from the origin database -- all of which were situated in the local environment for the Elektron experiment -- to the local SFX linking server. Disregarding the context of the Elektron experiment, this approach presents an interesting alternative to proxying solutions as used in the Open Journals Project. Such solutions have both the attractive and unattractive consequences of the non-involvement of the authorities in the deployed linking framework. In order to have adequate guarantees regarding the longevity of a solution, their involvement is desirable and may lead to a generic approach accepted by the information industry.



Related to the introduction of the SFX solution to grab the link-source is the just-in-time approach to the provision of service links. Service-links are only provided upon explicit request of the user, who actually starts the process to grab the link-source by doing so. This just-in-time approach dramatically reduces the amount of time required to deliver service links, because links are not computed for records in which the user is not interested.

### ***Dynamic linking via the SFX linking server***

The SFX server presents a solution to interlink the available information entities in a digital library environment, without requiring "a priori" computation of links from the available data. The solution uses concepts drawn from the domain of linking services, without being one in the strict sense. In SFX, the notion of a database containing bundles of links in which each record represents an inter-relationship between documents -- as used in BiomedNet's BundledLinks (Hitchcock et al. 1997b) (Figure 11) -- is replaced by a concept of potential inter-relationships between documents, expressed at the level of the databases from which they originate (Figure 12). The "a priori" computation of links -- as done in self-supporting environments such as BiomedNet -- is replaced by the "a posteriori" conceptual verification of links via the SFX-base, without any further functional verification. This results in a level of verification that lies between no verification, which is achieved when adding links blindly, and on the on-the-fly verification of links for every link-source (if that would be possible). The former requires little computing overhead but offers poor service; the latter offers perfect service, but causes significant delays (Hitchcock et al. 1997a). The proposed design achieves a balance between the extremes, through the introduction of the SFX-base that exploits know-how about the actual digital library collection in order to reduce both the amount of potential dead links and the required computing time. The presentation of unresolved service links in the SFX menu-screen and spreading the total required processing time over different phases further reduces potential delays. The more the SFX-base is fine-tuned, the more the risk of dead links can be reduced. In the Elektron experiment, the design of the SFX-base was rough and it has been fed "manually". There is clearly a need for more fine-tuning of the design, and for automated procedures to feed the SFX-base.



Interpreting the SFX solution as a searching aid, a navigational aid or as a provider of extended services rather than as a linking service in the traditional sense helps to justify the lack of complete verification that can be expected from traditional linking services. Moreover, such an interpretation can lead to the inclusion of other types of links in the Colli, such as:

- Links that redirect the actual search term to resources related to the one from which the link-source originates;
- Links that use other information from the link-source rather than SICI-related ones.

Extended service links can hardly be delivered into information resources that do not provide and

support a link-to-service. Link-to-services exist for some primary collections but are rare for abstracting & indexing databases. In order to be able to exploit the full richness of the digital library collection, each information resource should come with a link-to-service. Furthermore, if such link-to-services are conceived of as adhering to some generic framework -- such as the S-Link-S framework proposed by Eric Hellman (Hellman 1998) -- the implementation of certain components of SFX-like software would become much more straightforward.

# Part 2: Generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment

---

## Introduction

Part 1 described the Elektron experiment in which a novel mechanism was introduced to grab the link-source. It was shown that the mechanism worked successfully in the local environment where the experiment was conducted. It was noted that the mechanism also offered a promise with regard to the feasibility of a generic approach to transfer a link-source from its origin resource to the SFX-server. Still, the lack of resources in the experiment that were controlled by external authorities prevented an adequate proof regarding the feasibility to build an open linking framework based on the proposed mechanism. Part 1 also showed that the SFX-server -- a special kind of linking service -- was able to dynamically deliver a bundle of relevant extended services to an end-user who explicitly requested them. Still, its design was rough and not appropriate to be used in production environments. Part 2 gives a description of a generalized approach towards the creation of context-sensitive extended services in a digital library environment as developed in the course of the "SFX@Ghent and SFX@LANL" experiment. It identifies crucial and generic components of an open, dynamic linking framework and it describes the details of their implementation in the SFX linking framework. The discussion starts with the introduction of new concepts that are crucial to fully understand the generalized solution.

The "SFX@Ghent and SFX@LANL" experiment -- henceforth referred to as Ghent&LANL -- has been conducted in the complex digital library environments of the Research Library of the Los Alamos National Laboratory (New Mexico, US) and of the University of Ghent (Belgium). There was extensive cooperation from several parties in the information industry. There was active involvement of the library automation teams of Ghent and Los Alamos. The experiment was conducted between February 1999 and June 1999.

## Global and local relevance of extended services

In light of the results of the Elektron SFX-experiment and in order to come to a generalization of the SFX linking framework, it is interesting to further reflect on the Problem Statement of the SFX research:

*Given bibliographic metadata, how does one present relevant extended services for it?*

Here, the adjective *relevant* is of particular importance in the notion *relevant extended services*. Actually it has two meanings: relevance as a global notion and relevance as a local notion. In order to explain this, the following types of extended services are considered:

- *full\_text*: a service providing the full-text that is referred to by a link-source;
- *review*: a service showing a book review for the item referred to by a link-source;
- *abstract*: a service that provides the abstract from an abstracting & indexing database for a link-source.

**Relevance as a global notion** must be interpreted as being opposed to irrelevant in every context. Certain aspects of extended services are independent of the context of an individual collection; they actually apply on a global level:

- *full\_text*: If the publication year of an article is equal to or higher than that of the first electronic issue of the journal in which the article was published, a *full\_text* service has global relevance. On the other hand, it never makes sense to present a *full\_text* service for a link-source referring to a paper in a journal if the publication year of the paper is lower than the publication year of the first issue of the journal for which full-text is globally available. As such, the publication year of the first electronic issue is a constraint of global significance to the *full\_text* service.
- *review*: It is always irrelevant to present a book *review* service if the link-source refers to a journal article. But, if the link-source describes a book, such a *review* service is globally relevant. In this case, the material type is a constraint of global significance for the *review* service.
- *abstract*: A constraint of global significance rules the relevance of an *abstract* service that looks up the

abstract of a citation to a journal article in a particular abstracting & indexing database. Such a service is globally relevant if the journal in which the article is published is actually indexed in that abstracting & indexing database and is globally irrelevant otherwise.

**Relevance as a local notion**, refers to the fact that other aspects of extended services are dependent on the boundaries of a certain digital library collection. It refers to the context-sensitiveness of service-links. Local relevance has two manifestations:

- Relevance related to the content of a local collection:

While certain services are relevant in a global sense, they can become irrelevant if the digital library collection does not contain the information resource(s) required to implement them. Even if a *full-text* service is globally relevant for a certain link-source, it might be considered to be irrelevant in the context of a certain digital library collection if the journal referred to by the link-source is not part of that collection. In the same way, an *abstract* service pointing to a particular abstracting & indexing database for a given link-source can be globally relevant, as described above. Still, such a service is of no local relevance if the user's digital library does not provide access to an implementation of that particular database, while it can be of local relevance if the digital library does.

- Relevance related to the implementation of a local collection:

The relevance of extended services will also depend on the technical implementation of the information resource(s) required to create the services. When a *full-text* service is globally relevant -- an electronic edition of an article exists -- as well as relevant in relation to the content of a certain collection -- the users of the digital library are authorized to access the electronic edition -- it can be regarded inappropriate to let the *full-text* service link to a full-text instance at a publisher's site, when the digital library holds an instance in its local storage. In the DLF reference linking discussion, this issue was given the name of "the Harvard problem" or "the appropriate copy problem" (Caplan 1999a). Similar problems occur in the broader scope of extended services. For instance, as shown before, an *abstract* service can be globally relevant -- the journal in which an article was published is abstracted in a particular abstracting & indexing database -- as well as relevant in relation to the content of the collection -- the local digital library does provide access to the particular database. Still, the service might be irrelevant in relation to the implementation, if the actual implementation of the database does not support a mechanism to link into it using the parameters required to do an *abstract* look-up.

## Systems supportive of selective resolution

Both considerations regarding the local relevance of extended services emphasize the need for open linking solutions that take the context of the local collection into account when links are presented to a user.

When addressing the Harvard problem, the DLF reference linking discussions have referred to open linking solutions as systems supportive of selective resolution (Caplan & Arms 1999). From the above, it can be seen that the problem of local relevance of extended services is actually a generalization of certain aspects of the Harvard problem. As such, when a framework is able to present an approach to deal with the broader problem, the approach will also contain valuable elements to address the narrower Harvard problem.

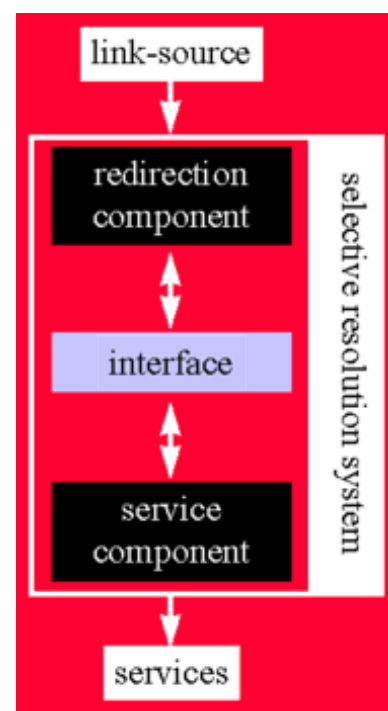


Figure 1: Systems supportive of selective resolution

In relation to the Harvard problem, Caplan and Arms divide systems that support selective resolution into two categories:

- Systems with a non-local location database, to which institutions provide a profile describing their full-text collection. The profile controls the selection of links returned to users of that profile.
- Systems with an institutional location database describing the local full-text collection and a global location database as a fall back. In addition to that, there is a mechanism to pass resolution requests to the local resolver first and in the event of a local full-text instance not being found there, to the global resolver.

This categorization can further be generalized by:

- Broadening the scope of the services to be provided beyond the restriction to the full-text, taking into account all kinds of extended services.
- Identifying the crucial components of systems supportive of selective resolution (see Figure 1):
  - The **redirection component** that brings metadata of the link-source, for which extended services are requested from the information resource to which the link-source belongs, to the service component. The redirection mechanism addresses the problem that has been referred to as grabbing the link-source in the Problem Statement and for which a partial solution has been presented in Part 1 that describes the Elektron experiment.
  - The **service component** that takes metadata from whichever information resource in the digital library collection as an input, delivering extended services as an output. The service component is an extension of the location database referred to by Caplan and Arms. The SFX server introduced in Part 1 is such a service component.
- Recognizing that the order of the redirection is subject to variation:
  - Redirection of the link-source metadata to the local service component first, using a central service component as a means to complete the set of services that can be presented.
  - Redirection of the link-source metadata to the central service component, whose default services can be overwritten and/or completed after communication with the local service component.

CATEGORY			
Category 1		central	central
Category 2	a	central & local	local => central
	b	central & local	central => local
Category 3		local	local
		SERVICE COMPONENT	REDIRECTION ORDER

**Table 1: categorization of systems supportive of selective resolution**

The resulting categorization is represented in Table 1, where 3 main categories of systems supporting selective resolution are shown, based on the nature of the service component and the redirection order:

- Category 1 only has a central service component and hence a central redirection mechanism. To some extent, this is the category under which the NCBI LinkOut solution resides. Still, since that solution is tied in with the PubMed database and cannot be used in connection with other resources, it can hardly be seen as a real service component in the sense described earlier.
- Category 2 has both a central and a local service component, both of which contribute to the presentation of the services. Also, there is some form of communication between both. For this Category, it is possible to imagine both approaches regarding the redirection order mentioned above.
- Category 3 builds purely on a local service component and hence also needs a local redirection mechanism. The SFX implementations of both the Elektron and the "SFX@Ghent & SFX@LANL" experiment (see

below) fall within this Category.

## The Ghent&LANL experiment

In the Ghent&LANL experiment, the Library Without Walls team of the Research Library at the Los Alamos National Laboratory (LANL) and the Automation Department of the Central Library at the University of Ghent have cooperated to illustrate the feasibility of the SFX approach as a means to provide context-sensitive extended services in a realistic and complex information environment.

- The information environment in which Ghent&LANL has been conducted is dramatically different from the one of the first Elektron SFX experiment. To illustrate this, Table 2 presents an overview of the information resources used in Ghent&LANL. The rows show the names of the information resources used in the experiment, the columns refer to the digital library collection. For each resource/collection combination the table indicates:
  - The Type of resource: OPAC system, abstracting & indexing database (A&I), full-text collection (FTXT) or web-service (WWW);
  - The Authority running the resource;
  - Whether or not the resource is used as a Source within the digital library collection. If it is a Source, information entities from the resource can be link-sources for which extended services can be requested. If a resource is a Source, the authority running it has made it interoperable with the SFX redirection mechanism. Such systems are henceforth called SFX-aware;
  - Whether or not the resource is used as a Target within the digital library collection. If it is a Target, the resource is used to be linked into in order to provide extended services. If a resource is a Target, a link-to syntax has been developed by the authority running the resource, in order to allow for it to be the Target of dynamic SFX-links.

RESOURCE	GHENT			LANL		
	Type	Authority	Source	Target	Authority	Source Target
Advance	OPAC	-	-	-	LANL	yes yes
Aleph 500	OPAC	Ghent	yes	yes	-	- -
Amazon.com	WWW	Amazon	no	yes	Amazon	no yes
Antilope	OPAC	UA	no	yes	-	- -
APS PROLA	FTXT	APS	yes	yes	APS	yes yes
the arXiv	FTXT	LANL	yes	yes	LANL	yes yes
BIOSIS	A&I	Ghent	yes	no	LANL	yes no
Books in Print	A&I	Ghent	yes	yes	Ghent	yes yes
Compendex	A&I	Ghent	yes	no	LANL	yes no
Current Contents	A&I	Ghent	yes	yes	Ghent	yes yes
EconLit	A&I	Ghent	yes	no	-	- -
Genome base	A&I	NCBI	no	yes	NCBI	no yes
Inspec	A&I	-	-	-	LANL	yes no

		SP	no	yes	SP	no	yes
Ulrich's	A&I	Ghent	yes	yes	-	-	-
LiSa	A&I	Ghent	yes	yes	-	-	-
MathSci	A&I	Ghent	yes	no	-	-	-
Medline	A&I	Ghent	yes	no	-	-	-
		NCBI	no	yes	NCBI	no	yes
SciSearch	A&I	LANL	yes	yes	LANL	yes	yes
ScienceServer	FTXT	LANL	no	yes	LANL	no	yes
Various	FTXT	various	no	yes	various	no	yes
Wiley InterScience	FTXT	Wiley	yes	yes	Wiley	yes	yes

**Table 2: information resources in Ghent&LANL**

Some considerations regarding Table 2:

- As can be seen, some resources are available in both digital library collections. Still, some run on different technical implementations in both environments. This is the case for BIOSIS and Compendex, which in Ghent run on a SilverPlatter ERL platform, while LANL -- at the time of the experiment -- used a Geac Advance implementation.
- For the purpose of this experiment, Ghent and LANL share some of their resources. Ghent makes its SilverPlatter ERL version of Books in Print and Current Contents available for LANL, whereas LANL opens access to its Topic implementation of the ISI Science Citation Index (SciSearch) and its ScienceServer storing the full-text of all Elsevier journals.
- Ghent uses two Medline versions: a locally stored ERL version as Source and the NCBI PubMed version as Target. Similarly, LANL uses two Inspec versions: the local Geac Advance implementation as a Source and an ERL implementation run by SilverPlatter in Boston as a Target. Time constraints that prevented the development of appropriate link-to syntaxes for the local versions are the reason for this peculiarity.
- Of special importance is the fact that some journals from the Wiley InterScience collection as well as the complete PROLA archive of the American Physical Society are made SFX-aware for the purpose of this experiment (Halstead 1999; Spilka 1999a). Both Ghent and LANL can use citations in the full-text of these repositories as link-sources for SFX requests. Also, in the course of this experiment Wiley has implemented a link-to syntax that will be brought into production early 2000, as a result of Ghent&LANL (see <http://mddb.wiley.com/instructions.html>). For the PROLA archive such a link-to syntax was already available (see <http://publish.aps.org/linkfaq.html>).
- Some resource names require a little more explanation. Aleph 500 is the Ghent Integrated Library System, Advance is the one for LANL, while Antilope is the Belgian Union Catalogue of Serials run by the University of Antwerp. The header "various" refers to a variety of full-text repositories to which dynamic links are available in this experiment. This is -- amongst others -- the case for Academic Press, Company of Biologists, HighWire, Springer, American Chemical Society, etc. The arXiv is the Topic implementation of the Ginsparg arXiv e-print repository, developed by the Library Without Walls team of the LANL Library. It has also been made SFX-aware. ScienceServer is the local full-text repository of the LANL Library that stores all Elsevier journals. A link-to-syntax is also available for it (see <http://journals.ohiolink.edu:8088/cgi-bin/sciserv.pl>).
- As can be seen from careful exploration of Table 2, from the point of view of each digital library collection, the SFX-aware information resources are highly distributed. Some resources are run by the institutional

library automation team while others are run remotely, actually by three external authorities. From the point of view of Ghent these external authorities are LANL, Wiley and the American Physical Society; from the point of view of LANL, they are Ghent, Wiley and the American Physical Society.

From the above, it can be concluded that given the amount of resources and technologies that is involved, and given their distributed nature and the availability of multiple SFX service components, Ghent&LANL is a very realistic experiment.

## **The need for a generalization of the SFX components**

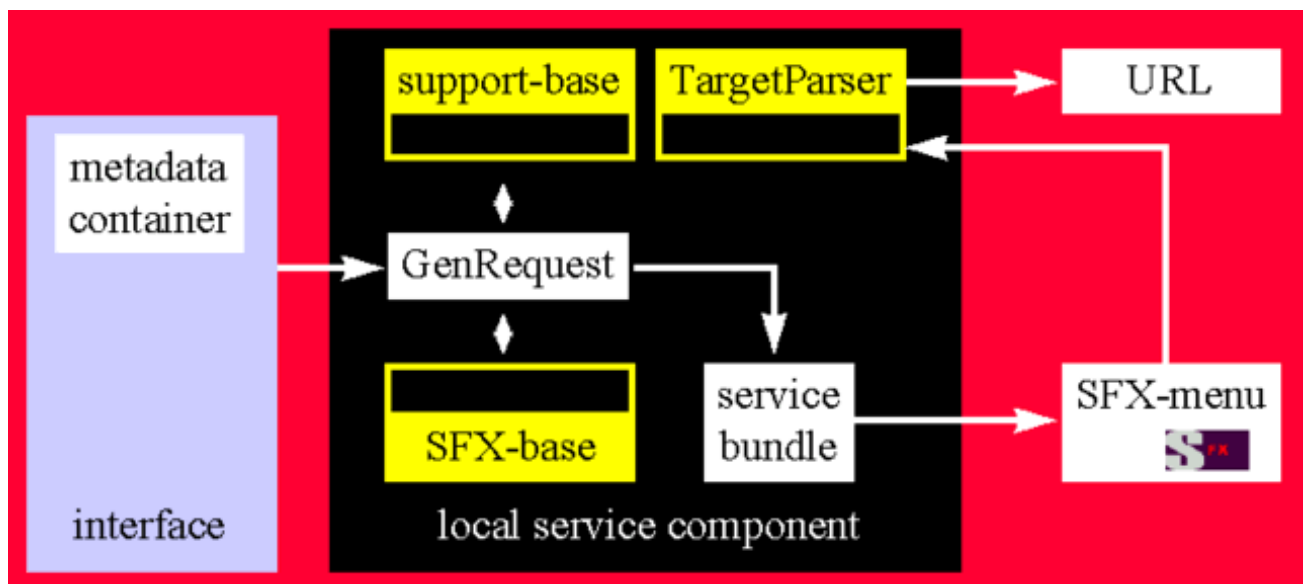
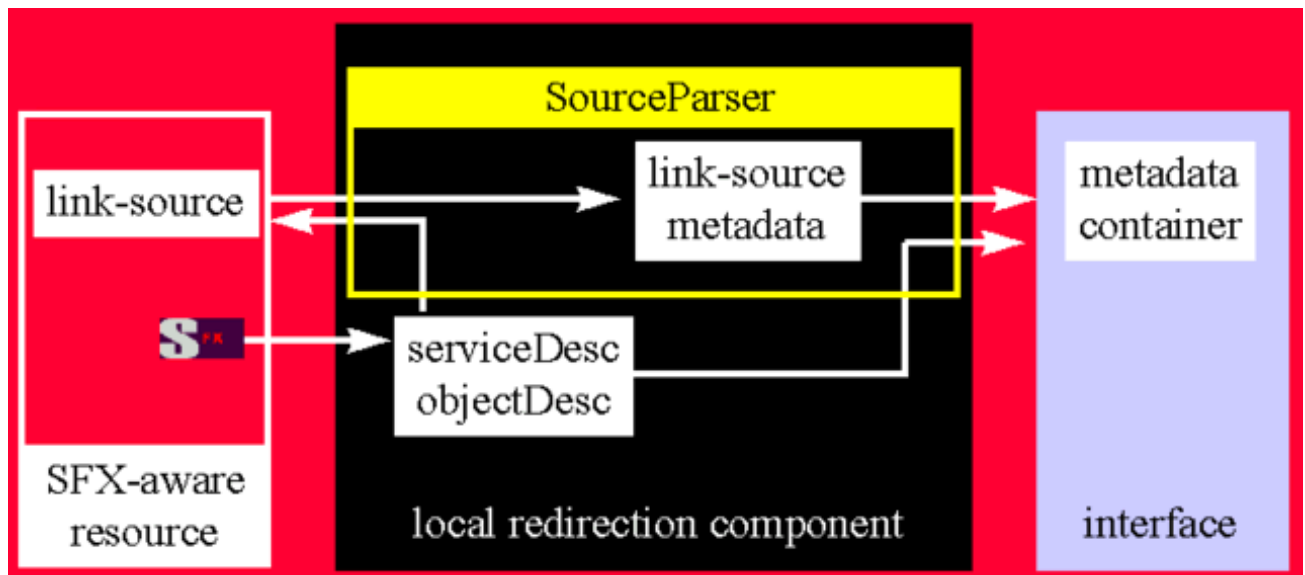
Although the fundamental concepts of the SFX framework -- dynamic linking, just-in time linking, conceptual services and the mechanism for local redirection (see Part 1) -- have been left untouched for the Ghent&LANL experiment, the nature of its working environment and its goals have led to a strong generalization of the components on which the framework builds. The main impulses that inspired such a generalization and that distinguish the Ghent&LANL project from the Elektron experiment are:

- The extension of the digital library collection in which SFX was being tested beyond a well-controlled sub-collection of one institution. In Ghent&LANL, SFX is introduced in the realistic, complex and dissimilar digital libraries of two autonomous institutions each running their local SFX components;
- The extension of the scope of data for which extended services can be requested beyond the internally stored collections. Link-sources in Ghent&LANL also originate from resources held by external authorities;
- The extension of the datatypes for which extended services can be requested beyond abstracting & indexing databases and OPAC systems. Link-sources in Ghent&LANL can also be citations in journal articles;
- The accommodation of extended services linking into target resources, based on metadata in general, not only SICI-related metadata;
- The need for high transportability of the SFX solution between the digital library environments involved.

The redesign of the SFX solution for Ghent&LANL leads to an architecture with a clear separation between the redirection component and the service component. Both components obviously interoperate in order to achieve a functional system. But the redirection component can potentially operate in an environment with non-SFX service components, while the SFX service component can equally function with another redirection mechanism, as long as that supports delivery of link-source metadata to the SFX service component. Several functional building blocks in both components have also been generalized in order to address the problems that arise from the complexity of the Ghent&LANL environment. The overall approach of the generalized solution is shown in Figure 2 and will be explained in more detail in the remainder of this Part.

Information resources that can interoperate with SFX -- SFX-aware systems -- insert an SFX-button for each link-source in the result set of a query. The just-in time approach of SFX requires the user to click such an SFX-button when requesting extended services for a specific link-source record. In response to this click, the local SFX redirection component will fetch link-source metadata -- usually -- from the origin resource using whichever protocol it takes to do so. Next, link-source metadata as well as information on its origin will be converted into an interfacing format. At this point, the local redirection mechanism has fulfilled its task and is able to deliver this information in a consistent representation to the local SFX service component.





**Figure 2: the local redirection and service components of the generalized SFX solution**

The first task of the local service component is to parse the information, handed over by the local redirection component, into a normalized internal representation object. During this process, the original content can be enhanced and/or augmented. The resulting information object is then fed into the SFX evaluation process in which it will be compared to the SFX-database. As shown in Part 1, the SFX-database is a special kind of linking database. Unlike traditional linking services, it does not contain any static links between "documents" (records/citations/full-text/etc.) of a collection. Rather, it contains a collection of conceptual services that express potential inter-relationships between documents at the level of the resource from which they originate. The SFX evaluation process determines the relevance of each of these conceptual services using the -- lack of -- content in the information object. Next, the resulting bundle of relevant services is sent back to the user in the SFX-menu-screen. A services from the bundle will only be resolved into a URL, when the user decides to use it. Then, the user will be redirected to that URL. This is consistent with the just-in-time approach of SFX.

## The SFX mechanism for local redirection

The task of the local redirection mechanism is to transport link-source metadata to the local redirection component, that interfaces with the local service component. In order to be able to interoperate with the SFX redirection mechanism, information resources need to be enhanced by the authorities running them in order to make them SFX-aware. The aim of this is to create the ability for information resources to insert an SFX-button targeted at the local redirection component for each link-source in the result set of a query into the resource. In

the context of Ghent&LANL, the following are important considerations with this regard:

- a. Many information resources that are involved in the experiment are also used in normal production at the very same time. This means that they are also approached by users that do not have access to an SFX service component. In order to prevent such a user from seeing an irrelevant SFX-button, an SFX-aware resource must be able to recognize whether the user has access to an SFX service component or not. Based on that information, the resource can insert an SFX-button or not.
- b. Some information resources are approached by users from both digital library environments, hence with access to different SFX service components. An SFX-aware resource must be able to target the SFX-button at the appropriate local redirection component, in order for it to be able to deliver the link-source metadata from the origin information resource to the doorstep of the appropriate service component. This means that an SFX-aware resource must be able to parameterize the target of an SFX-button.
- c. Upon receipt of a request for extended services from a user, the local redirection component must be able to fetch the link-source metadata from its origin resource. This means that the local redirection component has to be informed about the origin and the identity of the link-source in order to be able to take the appropriate steps. Given the amount, distribution and diversity of the SFX-aware resources in Ghent&LANL, a consistent manner to communicate such information to the local service components is required.
- d. Link-source metadata must be fetched from a wide variety of distributed information resources that support different access protocols. In addition, such resources will respond by sending link-source metadata formatted according to different metadata schemes. In order for the local redirection component to be able to interface in a generic manner with the local service component, a unique metadata interchange format is desirable.

As will be shown, in the detailed description below, these issues are approached by:

- For (a) and (b): the CookiePusher mechanism;
- For (c): the consistent SFX-URL structure;
- For (d): the SourceParser solution.

### ***Making information resources SFX-aware***

The authorities running information resources need to enhance their systems in order to make them SFX-aware. The complexity of the Ghent&LANL environment has urged for a thoughtful exploration of ways to make resources SFX-aware, since only approaches that minimize the overhead in doing so for the authorities running the resources can be acceptable and workable. In the current implementation of the SFX redirection mechanism, they have to do this by:

- Installing the CookiePusher script delivered by the project manager of Ghent&LANL;
- Inserting and hyperlinking SFX-buttons for link-sources using a URL that complies to a predefined format.

#### **The CookiePusher**

The CookiePusher script is a pragmatic solution introduced to dynamically notify an information resource about the existence and location of a local SFX redirection component in the environment of the user consulting the resource. The underlying idea is that an information resource could at any time access the location of a local redirection component, if its URL were written as a cookie in the browser of the user consulting the resource. The availability of this URL is essential, since the resource must be able to dynamically target the SFX-button at the appropriate local component. However, for reasons of security and privacy, such browser cookies can maximally be read within the Internet domain of the server that has set the cookie (see Shishir 1996 pages 203-204). As such, it is impossible to set a cookie in such a way that it can be read by all information systems in a digital library collection when it consists of resources distributed over several domains, typically resources that are local and remote to the user's institution.

In order to solve this problem, the first step in connecting to a resource is to request a server in the domain of the information resource to create an HTTP cookie. This detour is called the CookiePusher. The very simple CookiePusher script is installed in the domain of the information resource that has to be made SFX-aware. Rather than connecting immediately to the desired URL in the information resource, a connection is made to the

resource's CookiePusher first, sending values for the two parameters of the CookiePusher script:

- SFX\_location: the URL of the local redirection component of the SFX solution;
- Redirect: the desired URL in the resource.

Upon receipt of these parameters, the CookiePusher will first read the URL of the local redirection component and will use it to set a cookie in the user's browser. Since the CookiePusher is in the domain of the resource, that cookie will be readable by the resource. Next, the CookiePusher will redirect the user to the desired URL in the resource.

As such, once the CookiePusher has been installed for a resource, the URL to connect to that resource will be changed to:

**CookiePusher\_URL?SFX\_location = local\_SFX& Redirect = service\_URL**

Where

- **CookiePusher\_URL** is the URL of the CookiePusher script;
- **local\_SFX** is the URL of the local SFX redirection component;
- **service\_URL** is the desired URL in the information resource as used under normal -- non-SFX -- conditions. Such a URL can point at the initial search screen for an abstracting & indexing database, it can be a URL linking to an article at a publishers site, etc.;
- **local\_SFX** and **service\_URL** are URL-encoded.

For instance:

```
http://publish.aps.org/edaccess/prolatest/cookiepusher?  
SFX_location=http%3A%2F%2Fisiserv.rug.ac.be%2Fcgi-bin%2Fsfx%2Fbin%2Fmenu.cgi  
& Redirect=http%3A%2F%2Fpublish.aps.org%2Fedaccess  
%2Fprolatest%2Ftext%2FPRD%2Fv52%2Fil%2Fp15_1
```

is the URL used to connect to an item in the APS/PROLA domain. The APS/PROLA CookiePusher will read the location of the local redirection component from the SFX\_location parameter and will use this to set a cookie named local\_SFX with value:

```
http%3A%2F%2Fisiserv.rug.ac.be%2Fcgi-bin%2Fsfx%2Fbin%2Fmenu.cgi
```

which is the encoded location of the Ghent local SFX redirection component. Next, it will redirect the user to the desired location in the APS/PROLA:

```
http://publish.aps.org/edaccess/prolatest/text/PRD/v52/il/p15_1
```

From now on, at any point in the consultation, APS/PROLA will be able to read this cookie and use it to target -- in this case -- the Ghent redirection component.

#### The consistent SFX-URL structure hyperlinking the SFX-button

The essence of the detour made via the CookiePusher is the ability it creates for an information resource to know at any point whether or not the consulting user has access to a selective resolution system and, if so, what the location of its redirection component is. Based on that information, the resource can dynamically decide whether or not to insert an SFX-button for search results and if it does, which redirection component to target with the SFX-button. In order to make the many systems involved in the Ghent&LANL experiment interoperable with SFX, authorities running the systems have been asked to make the URL targeted by the SFX-button -- the SFX-URL -- compliant to the following format:

<b>GENERAL</b>	target?serviceDesc& objectDesc
<b>DETAILED</b>	local_SFX?vendorId=<theVendor>&databaseId=<theBase>&objectDesc=<theIdentifier>

**Table 3: the syntax of the SFX-URL**

In Table 3:

- target is the URL of the local redirection component of the SFX solution;
- **serviceDesc** uniquely defines the origin resource. It contains information on the vendor of the resource and

on the resource itself. It is of the form:


**vendorId=<theVendor>&databaseId=<theBase>.**

serviceDesc information will play a crucial role at later stages of the SFX local redirection mechanism, as well as in the SFX-base which is central to the SFX service component.

- **objectDesc** contains information that relates to the identity of the link source. Its syntax and content is extremely flexible and it will be defined by the authority running the resource, making it dependent on the vendor and his database implementation. objectDesc typically contains the unique record identifier for a link-source in its origin resource. Alternatively or in addition to that, it can contain SICI-like metadata. In some cases, it can even contain all metadata of the link-source.
- The parameter values **<theVendor>**, **<theBase>** and **<theIdentifier>** are URL-encoded.

Figure 3 to Figure 6 show examples of link-sources taken from Sources in the Ghent and/or LANL collections, mentioning their SFX-URL. For reasons of readability, the parameter values are not shown as being URL-encoded. Rather, it is mentioned that parts should be URL-encoded by enclosing them in a URLEncode function.

**❑ Record 2 of 33 in Biological Abstracts 1999/01-1999/03**

TI Long term outcome of patients with **hairy cell leukemia** treated with pentostatin.  
AU [Ribeiro-Patricia](#); [Bouaffia-Fadhela](#); [Peaud-Pierre-Yves](#); [Blanc-Michel](#); [Salles-Bruno](#)  
AD {a} Serv. Hematol., Cent. Hosp. Lyon-Sud, 165 Chemin du Grand Revoyet, 69495  
SO [Cancer](#)-Jan. 1, 1999; 85 (1) 65-71..  
PY 1999  
DT Article-  
IS 0008-543X  
LA English  
AI Y  
ST Hominidae-; Primates-, Mammalia-, Vertebrata-, Chordata-, Animalia-  
RN 53910-25-1: PENTOSTATIN  
AN 199900063465  
UD 19990223 .  
AV 

SFX-URL for this link-source, pointing at the Ghent local redirection component:

[http://isiserv.rug.ac.be/cgi-bin/sfx/bin/menu.cgi? vendorId=ERL&databaseId=BX  
&objectDesc=URLEncode\(BX02 A:199900063465 I:0008-543X V:00085 S:000001 P:000065 Y:1999\)](http://isiserv.rug.ac.be/cgi-bin/sfx/bin/menu.cgi?vendorId=ERL&databaseId=BX&objectDesc=URLEncode(BX02 A:199900063465 I:0008-543X V:00085 S:000001 P:000065 Y:1999))

In the serviceDesc part of the URL, ERL refers to the SilverPlatter ERL implementation of BIOSIS, while BX is the family name of BIOSIS databases in the ERL environment. The objectDesc component contains several information elements in a tagged and fixed length representation. BX02 is the volume of the BIOSIS database where the link-source originates, while 199900063465 is the accession number, a unique record number of the link-source in BIOSIS. Other elements in the objectDesc are ISSN number, volume, issue, starting page and publication year.

**Figure 3: a link-source from the Ghent ERL implementation of BIOSIS and its SFX-URL**

Record 3 of 91

Mark

Full Record

Article:



[Article](#)

Title:

[The identification of cDNAs that affect the \*\*mitosis\*\*-to-\*\*interphase\*\* transition in Schizosaccharomyces pombe, including sbp1, which encodes a sp1p-GTP-binding protein.](#)

Author:

[HE, XIANGWEI](#) ; [HAYASHI, NAOYUKI](#) ; [WALCOTT, NATHAN G.](#) ; [AZUMA, YOSHIO](#) ; [PATTERSON, THOMAS E.](#) ; [BISCHOFF, F. RALF](#) ; [NISHIMOTO, TAKEHARU](#) ; [SHELLEY, SHELLEY](#)

Source:

[Genetics, Feb. 1998; v.148, no.2, p.645-656.](#)

Other Links:



Location

Holdings

MAIN

Holdings: v.109- (1985- ) ; Missing: v. 134 no. 4 (1993) ; Last rec'd: VOL.152 NO.2 / J issue on display ; Shelved as: GENETICS.

WWW

<http://www.genetics.org/> ; Holdings: v.148, no.1- (Jan.1998- ) ; Abstracts and table of contents: v.147, no.4 (Feb.1980-Dec.1997)

SFX-URL for this link-source, pointing at the LANL local redirection component:

[http://vole.lanl.gov/cgi-bin/sfx/bin/menu.cgi? vendorId=ADVANCE&databaseId=Biosis  
&objectDesc= URLencode\(fetchId=21179970&objectId=  
PREV199800135979&SICI=0016-6731\(1998\)148:2<645:TIOCTA>2.0.TX;2-P\)](http://vole.lanl.gov/cgi-bin/sfx/bin/menu.cgi?vendorId=ADVANCE&databaseId=Biosis&objectDesc=URLencode(fetchId=21179970&objectId=PREV199800135979&SICI=0016-6731(1998)148:2<645:TIOCTA>2.0.TX;2-P))

The serviceDesc part of this URL is self-explanatory. The objectDesc component is tagged and fields can have variable lengths. The fetchId is the unique number of the link-source in the LANL implementation of BIOSIS, while the part of objectId after "PREV" is the BIOSIS accession number which is comparable to the A field in the SilverPlatter objectDesc of Figure 3. The SICI part contains a SICI for the link-source, from which ISSN, volume, issue, pagination and publication year can be derived.

**Figure 4: a link-source from the LANL Advance implementation of BIOSIS and its SFX-URL**

WebSPIRS: Search - Netscape

Special Effects: University of Ghent - Netscape

Article Abstract - Netscape

WILEY  
**InterScience®**

PERSONAL HOMEPAGE JOURNAL FINDER SEARCH HELP CONTACT US LOGOUT

ALL JOURNALS PREVIOUS ARTICLE NEXT ARTICLE

**Article Abstract**

**CANCER**

Online ISSN: 1097-0142 Print ISSN: 0008-543X

**Cancer**  
Volume 85, Issue 1, 1999. Pages: 65-71

ADD HOT ARTICLE

**Original Article**

**Long term outcome of patients with hairy cell leukemia treated with pentostatin**

Patricia Ribeiro, M.D.<sup>1</sup>, Fadhela Bouaffia, M.D.<sup>1</sup>, Pierre-Yves Peaud, M.D.<sup>2</sup>, Michel Blanc

**References**

- 1 Saven A, Piro L. Treatment of hairy cell leukemia. *Blood* 1992; **79**: 1111-20. [Medline](#) SFX
- 2 Jaiyesimi I, Kantarjian H, Estey E. Advances in therapy for hairy cell leukemia. A review. *Cancer* 1993; **72**: 5-16. [Medline](#) SFX
- 3 Saven A, Piro L. The newer purine analogues for the treatment of hairy-cell leukemia. *N Engl J Med* 1994; **330**: 691-7. [Medline](#) SFX

SFX-URL for the third reference as a link-source, pointing at the Ghent local redirection component:

<http://isiserv.rug.ac.be/cgi-bin/sfx/bin/menu.cgi?vendorId=Wiley&databaseId=WIS>

& objectDesc= URLEncode(TYPE=JCIT& SNM=Saven&FNM=A&SNM=Piro&FNM=L&ATL=The newer purine analogues for the treatment of hairy-cell leukemia.&JTL=N Engl J Med &PYR=1994&VID=330&PPF=691&PPL=7)

The serviceDesc component now refers to the Wiley InterScience collection. The objectDesc is tagged and starts with an indication on the material type of the reference -- journal citation in this case -- followed by a tagged repetition of the full citation.

**Figure 5: a link-source from Wiley InterScience and its SFX-URL**

Preprints Retrieval Results - Netscape

File Edit View Go Communicator Help

Los Alamos

NATIONAL LABORATORY

Research Library

Preprints

New Search

Up

Comments







Mark All

List Marks

Help

Preprints Retrieval Results

23 out of 100245 records matched the query below. 23 records displayed:  
(fractal <in> tisuaab <and> year >= 1991)<and>year <= 1999<and>physics <in> archcode

Marks	Score	Title, Author, Eprint	
<input type="checkbox"/>	0.97	 	<a href="#">Integers and Fractions</a> Diptiman Sen (Indian Institute of Science, Bangalore, India) physics/9811004 (03 Nov 1998)
<input type="checkbox"/>	0.96	 	<a href="#">Irrational Numbers of Constant Type --- A New Characterization</a> Manash Mukherjee and Gunther Karner physics/9706009 (04 Jun 1997)
<input type="checkbox"/>	0.95	 	<a href="#">Adaptive Ising Model and Bacterial Chemotaxis</a> Yu Shi physics/9901013 (28 Jan 1999)

SFX-URL for the first link-source in the above result screen, pointing at the LANL local redirection component:

http://vole.lanl.gov/cgi-bin/sfx/bin/menu.cgi?vendorId=LANLTopic& databaseId=arXiv

&objectDesc= URLEncode(fetchId=phys-9811004&objectId=physics/9811004)

The serviceDesc refers to the LANL Topic implementation of the Ginsparg e-print archive. The fetchId is the unique key for the record in that implementation, while the -- very similar -- objectId is the unique record number in Ginsparg's implementation of the archive. No further metadata is available in the objectDesc.

Figure 6: a link-source from the arXiv and its SFX-URL

### Fetching link-source metadata from an SFX-aware information resource with SourceParsers

The CookiePusher mechanism enables a resource to insert an SFX-button for each of the link-sources that are transferred to a user consulting the resource. The structure of the SFX-URL targeted by these SFX-buttons has been made consistent across resources to be of the form `target?serviceDesc&objectDesc`. When a user requests extended services by clicking such an SFX-button, a request is sent to his local SFX redirection component, which will receive `serviceDesc` and `objectDesc` values as parameters for the target script. The local component holds a collection of `SourceParser` scripts with names corresponding to valid `serviceDesc`'s (see Table 4). Having analyzed the `serviceDesc` information, the target script will launch the appropriate `SourceParser`. This `serviceDesc`-specific `SourceParser` uniquely implements three distinct functions:

- The interpretation of the information contained in the `objectDesc` parameter based upon the syntax defined by the vendor (see examples in Figure 3 to Figure 6);
- The mechanism to fetch the link-source from its origin resource based on its origin and on the content of its `objectDesc`. Table 4 shows those fetch mechanisms for the examples of Figure 3 to Figure 6. As can be seen, no real fetching is required for the Wiley citations, since these are completely transferred in the `objectDesc` part of the SFX-URL. The same technique is used for citations in the PROLA archive. Both the Ghent and



LANL BIOSIS implementations deliver some -- SICI related -- metadata in the objectDesc. But since several extended services that SFX aims to deliver require additional metadata, a fetch is required in order to obtain more complete information. Since the objectDesc for the arXiv only contains an identifier, a fetch is definitely required;

- The conversion of the fetched link-source metadata, that is expressed in the metadata scheme supported by the authority running the origin resource, into a metadata container compliant with the scheme of the unique metadata interchange format. This metadata container is the interface between the local redirection component and the local service component.

RESOURCE	serviceDesc		SourceParser	Fetch protocol	Fetch key
the arXiv	LANLTopic	arXiv	S::LANLTopic:arXiv	HTTP	fetchId
BIOSIS	ERL	BX	S::ERL::BX	Z39.50	A
BIOSIS	ADVANCE	Biosis	S::ADVANCE::Biosis	Z39.50	fetchId
Wiley	Wiley	WIS	S::Wiley::WIS	none	none

**Table 4: Some SFX-aware resources with their serviceDesc, Fetch protocol and Fetch key**

## The SFX service component

The task of the local SFX service component starts at the point where the local redirection mechanism hands over the metadata container that contains, in a consistent representation:

- link-source metadata that became available through the local redirection mechanism;
- information on the origin of the link-source, basically serviceDesc information.

It is the task of the SFX service component to deliver extended services based on this information. The following are important considerations regarding the SFX service component in Ghent&LANL:

- a. The amount and quality of link-source metadata that becomes available in the metadata container is dependent on the type of resource from which its link-source originated and on the amount of information that the authority running the origin resource allows and/or supports to be fetched. In some cases such metadata can be corrupt or lack information that is essential for the SFX evaluation process to adequately perform its task;
- b. The SFX service component must be easily transportable between different digital library environments and remain easily manageable;
- c. The SFX service component must ultimately deliver service links on a just-in-time basis.

As can be seen from a detailed description of the SFX service component, these problems have been approached by:

- For (a): the GenericRequest object;
- For (b): a generalization of the implementation of the SFX-database, that explicitly reflects the notion of global and local relevance of conceptual services as well as the notion of global and local Thresholds;
- For (c): the TargetParser solution.

### **The GenericRequest object**

The service component will take the metadata container delivered by the local redirection mechanism as input and turn it into a normalized internal representation, called the GenericRequest object. Table 5 shows a representation of the GenericRequest object for the third citation in Figure 5. The GenericRequest object is an intelligent object, that is able to self-check the validity of its information elements based on pre-configured rules. It can also augment/enhance its content using information from a supporting database. For instance, the citation of Figure 5 does not contain an ISSN number nor a journal title, but rather an abbreviated journal title. In this case, the GenericRequest object augments its content, by adding the missing information via communication with the supporting database. Obviously, the GenericRequest object also contains a normalized version of the link-source



metadata, as well as information about its origin.

At the time of the experiment, interoperability between the SFX local service component and non-SFX local redirection mechanisms was not an issue, since none were existing. As such, for reasons of simplicity, the metadata scheme of the GenericRequest object has fulfilled the role of interfacing metadata scheme between the local redirection and the local service component in Ghent&LANL.

```
<perldata>
<hash>
<item key="rec$vendorId">Wiley</item>
<item key="rec$databaseId">WIS</item>
<item key="rec$dbId">Wiley::WIS</item>
<item key="objectType">JOURNAL</item>
<item key="@abbrevTitle">
<array>
<item key="0">N ENGL J MED</item>
</array>
</item>
<item key="journalTitle">NEW ENGLAND
JOURNAL OF MEDICINE</item>
<item key="ISSN">0028-4793</item>
<item key="year">1994</item>
<item key="volume">330</item>
<item key="startPage">691</item>
<item key="endPage">7</item>
<item key="@authLast">
<array>
<item key="0">Saven</item>
<item key="1">Piro</item>
</array>
</item>
<item key="@authInit">
<array>
<item key="0">A</item>
<item key="1">L</item>
</array>
</item>
<item key="articleTitle">The newer purine
analogues for the treatment of
hairy-cell leukemia.</item>
</hash>
</perldata>
```

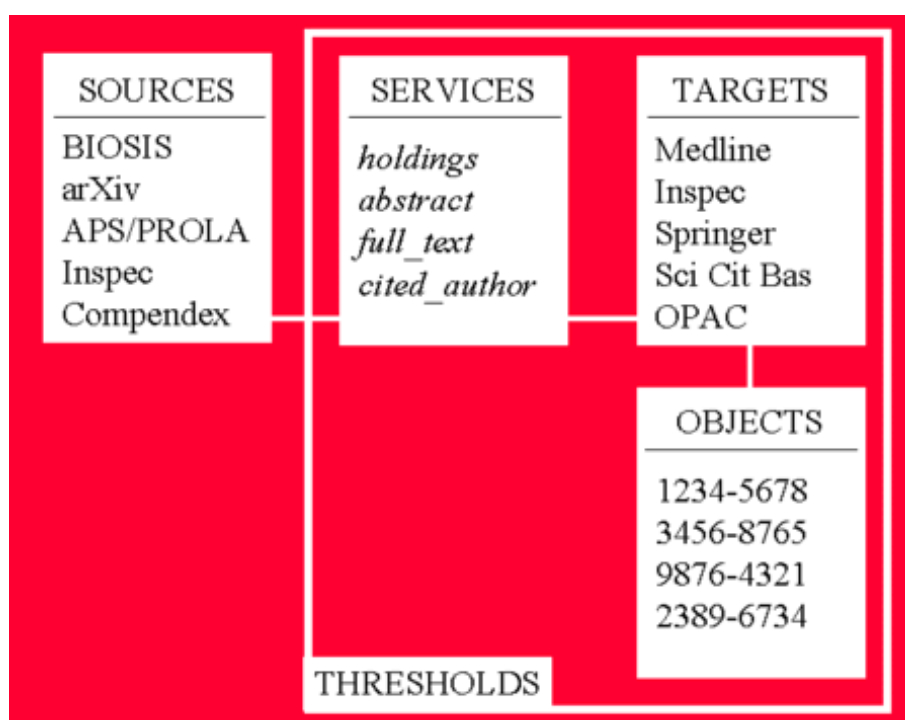
**Table 5: Representation of an augmented GenericRequest object for the link-source of Figure 5**

### ***The SFX linking service and the SFX-base***

As a result of the above, an instance of the GenericRequest object for the link-source for which extended services have been requested has become available to the SFX service component. It will be the task of this component to deliver the extended services to the user who has requested them. In this sense, the SFX service component is a linking service that, given a certain input "document", outputs "documents" related to the input. The SFX linking service is special, however, since it does not store static relationships between individual documents. Rather, it stores relationships between the resources from which the documents originate. In SFX, these relationships are called conceptual services and they are stored in the SFX-base. The SFX evaluation process will determine the relevance of each of these conceptual services based upon the information and origin of a link-source.

The requirement imposed on the Ghent&LANL implementation of the SFX service component to be easily transportable between different digital library environments has led to an important generalization of the design

of the SFX-base. This has been achieved by explicitly incorporating the notion of global and local relevance of services in the implementation. A synthesized representation of the lay-out of the Ghent&LANL SFX-base is given in Figure 7.



**Figure 7: Simplified lay-out of the SFX-base**

### Splitting the Colli table

As in the Elektron version of the SFX-base, the Source table contains the information resources that can be origins for link-sources. They are SFX-aware resources. In the Elektron version, the Colli contained conceptual services, directly coupled with the Target resources. (see Table 2 in Part 1). Such a set-up was not adequately generic and, in the Gent&LANL design, this Colli has been split. One table has kept the name Colli, the other has been named the Target table. The Target table contains those resources into which linking is possible. The Colli table that connects the Source and Target tables now expresses the type of service that relates Source with Target resources. Table 6 shows the type of services implemented in Ghent&LANL.

COLLI SERVICES	FUNCTION
<i>abstract</i>	look-up of abstract information in an abstracting & indexing database for the item represented by the GenericRequest object
<i>author</i>	look-up of references by an author of the item represented by the GenericRequest object in an abstracting & indexing database
<i>cited_author</i>	look-up of citations to work by an author mentioned in the GenericRequest object
<i>cited_reference</i>	look-up of works citing the item represented by the GenericRequest object
<i>full_text</i>	link to the full-text of the item represented by the GenericRequest object
<i>genome</i>	look-up of sequence information found in the GenericRequest object
<i>holding</i>	holdings look-up in an OPAC system for the item represented by the GenericRequest object
<i>review</i>	look-up of a book review for then item represented by the GenericRequest object

**Table 6: Services in the Colli and their function**

### Taking advantage of the global relevance of conceptual services

It is not a coincidence that the resources shown as Source and/or Target carry their globally common names rather than those of their local implementations in Ghent or LANL. This is actually a reflection of the conclusion that services relating Source and Target resources have global relevance. It is globally relevant to deliver an *abstract* service that -- given a link-source from BIOSIS -- shows the corresponding abstract from Medline. Such a conceptual service can be imagined regardless of the implementations of each of these resources in a specific digital library. Therefore, the Ghent&LANL SFX-base expresses the relationships between Sources and Targets at the level of global relevance: there is an *abstract* service connecting BIOSIS and Medline, regardless of their local implementations. A very limited number of examples of how such services of global relevance connect Source and Target is shown in Table 7.

COLLI		
SOURCE	SERVICE	TARGET
APS/PROLA	<i>abstract</i>	Inspec
the arXiv	<i>author</i>	Inspec
BIOSIS	<i>abstract</i>	Medline
BIOSIS	<i>genome</i>	Genome Base
Current Contents	<i>abstract</i>	LiSa
EconLit	<i>review</i>	Books in Print
Inspec	<i>full_text</i>	Springer
Wiley	<i>abstract</i>	Medline
Wiley	<i>cited_reference</i>	Science Cit. Base

**Table 7: Examples of global service relationships between Sources and Targets**

### Localization of services of global relevance

While the services shown in Table 7 are of global relevance, they do not take into account issues of relevance in relation to the local digital library collection. This localization of services of global relevance is achieved by:

- The introduction of fields referring to the local implementations, next to the globally common names.

As shown in Table 8 and Table 9, a key reflecting the serviceDesc values of the local implementations of resources -- found in the rec\$dbId field of the GenericRequest object -- is added next to the global common name of the Sources. In the same way, at the Target side, the name of a local TargetParser is added next to the global name of which the local Target is an implementation. The TargetParser procedure implements the link-to syntax into the local implementation of the Target resource. It can be seen from Table 8 and Table 9 that Ghent and LANL use a different SourceParser for BIOSIS, which reflects that they have a different implementation. However, they share a TargetParser to provide the *abstract* service into Medline, since both have chosen the PubMed implementation as a Target to achieve this.

- Deactivating services of global relevance when they are not of local relevance.

When the Source or Target resource required to implement a certain service is not available in the digital library collection, when the local implementation of the Target resource does not support the link mechanism required to implement the service, or when local librarians decide the service to be of no use to their end-users, its flag will be set to inactive. The service will no longer be taken into account in the SFX evaluation process deciding on the local relevance of conceptual services. In Table 8 this is the case for

services with Inspec as a Source since Ghent does not have an Inspec implementation in its collection. In Table 9, this is the case for services with LiSa as a Target, since LANL does not have access to a LiSa implementation.

SOURCE		COLLI	TARGET	
local	global		global	local
S::APS::PROLA	APS/PROLA	<i>abstract</i>	Inspec	T::ERL::IN
S::LANLTopic::arXiv	the arXiv	<i>author</i>	Inspec	T::ERL::IN
<b>S::ERL::BX</b>	<b>BIOSIS</b>	<i>abstract</i>	<b>Medline</b>	<b>T::NCBI::PubMed</b>
<b>S::ERL::BX</b>	<b>BIOSIS</b>	<i>genome</i>	<b>Genome Base</b>	<b>T::NCBI::Genome</b>
S::ERL::CCO	Current Contents	<i>abstract</i>	LiSa	T::ERL::LI
S::ERL::EC	EconLit	<i>review</i>	Books in Print	T::ERL::BOIP
<b>inactive</b>	Inspec	<i>full_text</i>	Springer	T::Springer::LINK
S::Wiley::WIS	Wiley	<i>abstract</i>	Medline	T::NCBI::PubMed
S::Wiley::WIS	Wiley	<i>cited_reference</i>	Science Cit. Base	T::CIC15:SciSearch

**Table 8: Localization of services from Table 7 for Ghent**

SOURCE		COLLI	TARGET	
local	global		global	local
S::APS::PROLA	APS/PROLA	<i>abstract</i>	Inspec	T::ERL::IN
S::LANLTopic::arXiv	the arXiv	<i>author</i>	Inspec	T::ERL::IN
<b>S::Advance::Biosis</b>	<b>BIOSIS</b>	<i>abstract</i>	<b>Medline</b>	<b>T::NCBI::PubMed</b>
<b>S::Advance::Biosis</b>	<b>BIOSIS</b>	<i>genome</i>	<b>Genome Base</b>	<b>T::NCBI::Genome</b>
S::ERL::CCO	Current Contents	<i>abstract</i>	LiSa	<b>inactive</b>
<b>inactive</b>	EconLit	<i>review</i>	Books in Print	T::ERL::BOIP
S::Advance::Inspec	Inspec	<i>full_text</i>	Springer LINK	T::Springer::LINK
S::Wiley::WIS	Wiley	<i>abstract</i>	Medline	T::NCBI::PubMed
S::Wiley::WIS	Wiley	<i>cited_reference</i>	Science Cit. Base	T::CIC15:SciSearch

**Table 9: Localization of services from Table 7 for LANL**

### Global and local Thresholds

The relationships between Source and Target resources expressed by a service connection in the Colli is made subject to restrictions called Thresholds. These Thresholds are the way to fine-tune conceptual services in order to minimize the presentation of services that are considered not to be appropriate to be presented. In order to illustrate this concept, two types of Thresholds are described:

- Thresholds expressed in terms of boundaries for the metadata elements that make up the GenericRequest object structure. Technically, these Thresholds are expressed as conditional statements using field names of the GenericRequest object. Such Thresholds are in many cases very simple, but they can as well be scripts of whichever degree of complexity. For instance:
  - *cited\_author*: In order for a *cited\_author* service to be relevant, the lowest Threshold that has to be passed is the existence of an author name in the GenericRequest object. Such a Threshold could be expressed as `$GenericRequestObject->need('authLast')`.
  - *book\_review*: A *book\_review* service is only relevant for link-sources that describe books: `$GenericRequestObject->need('objectType', 'eq', 'BOOK')`.
  - *genome*: A *genome* service is only relevant if the link-source contains genome sequence identifiers: `$GenericRequestObject->need('genID')`
  - *abstract*: The Threshold for an *abstract* service might express that the link-source should describe a journal article and should at least have year, volume and issue information:  
`$GenericRequestObject->need('objectType', 'eq', 'JOURNAL') &&`  
`$GenericRequestObject->need('year') && $GenericRequestObject->need('volume') &&`  
`$GenericRequestObject->need('issue')`.
- objectLookup Thresholds: The *abstract* service is clearly also subject to another type of boundary requiring that the Target resource into which the abstract service intends to link, actually abstracts the journal in which the item referred to by the GenericRequest object was published. This requirement explains the existence of the Objects table in Figure 7 and of a special objectLookup threshold. This type of Threshold will also be required to determine whether a journal in which the item referred to by link-source was published is part of a specific full-text repository in order to decide on the relevance of a *full\_text* service into the repository.

Just as with the conceptual services, there is a global and a local component to these Thresholds. The global objectLookup Threshold for a *full\_text* service linking into the Springer full-text collection, will learn whether a certain journal is a Springer e-journal or not. The local part of this Threshold will learn whether the journal is part of the actual digital library collection. In the same context, a global Threshold can express the fact that a journal is available in electronic form since 1996, while the local component might show that the local subscription only starts in 1998. In the same way, the BIOSIS-*abstract*-Medline service is subject to a global objectLookup Threshold. But there is also a global Threshold expressing that the publication year of the GenericRequest object has to be greater than 1965, reflecting the full coverage of Medline. Still, the local Threshold component for this service might be set to a more recent year if the local Medline implementation stores less data.

### **The SFX evaluation process**

In order to present extended services for a given GenericRequest object the SFX evaluation process will determine the relevance of each of the conceptual services stored in the SFX-base using the content, or lack thereof, in the GenericRequest object. There are two phases to this evaluation process.

#### Phase 1: Selection of active services with the origin resource of the GenericRequest object as a Source

The interface between the redirection component and the service component delivers both link-source metadata and information on the origin of the link-source. The latter is stored in the `rec$dbId` field of the GenericRequest object that is created by the service component. During the evaluation phase, the value of this field becomes the key for a lookup in the local component of the Source table of the SFX-base. The global common name of the resource is detected there, next to this key which refers to the local implementation of a resource. This global name is now connected via services of the Colli to various global names of Target resources, as already shown in Table 7. Hence, the result of this lookup is a bundle of services that might be relevant for the current GenericRequest object, as judged upon by its origin. Inactivation of certain services during the localization of the SFX-base guarantees that the resulting bundle already reflects the local digital library situation.

In Table 8 and Table 9 the mechanism is shown in bold for a GenericRequest object representing an item originating from the Ghent implementation of BIOSIS. Its `rec$dbId` value is `ERL::BX`. In this phase of the evaluation process, `S::ERL::BX` -- a prefix `S` is added as a means to refer to Source -- becomes the key for a look-up in the local component of the Source table. There, BIOSIS is detected as the global common name of the

resource. Several services are leading out from BIOSIS into Target resources. For instance, *abstract* connects BIOSIS with Medline & *genome* connects BIOSIS with Genome Base. These are the services that remain as potentially relevant.

#### Phase 2: Filtering out selected active services by comparing the content of the GenericRequest object with the Thresholds

Phase 1 of the SFX evaluation process filters out services of the SFX-base that do not have BIOSIS as an origin. For each of the resulting services, the information in the GenericRequest object will be matched against the Thresholds -- global and local -- that are connected to these services. The *genome* service connecting BIOSIS and Genome Base will be filtered out if the GenericRequest object does not contain an entry for the genID parameter. The *abstract* service connecting BIOSIS and Medline will be filtered out if an objectLookup for the ISSN value in the GenericRequest object learns that the journal is not abstracted in Medline. Or it could be filtered out if the GenericRequest object does not contain a value for year, volume or issue.

Again, since some Thresholds express the local situation, and since these Thresholds can overrule the global ones, the result of this filtering process will reflect the situation of the local digital library collection. Those services remaining from Phase 1, for which at least one of the Threshold evaluations fails will be rejected as not being relevant. The ones that make it through the complete evaluation process will be presented to the user in the SFX-menu-screen as locally relevant extended services for the current GenericRequest object, hence for the link-source for which the whole process has been initiated by clicking the SFX-button

#### **Resolving locally relevant extended services into URLs with TargetParsers**

Consistent with the just-in-time linking philosophy of SFX, the bundle of relevant services that is obtained as a result of the SFX evaluation process described above, is not resolved into URLs at the moment of their presentation to the user that launched a request for extended services. Rather, for each menu-item in the SFX-menu-screen, the following elements are sent as parameters for a script that will be initiated when a user selects a menu-item:

- The identifier of the GenericRequest object;
- A name referring to the service and its Target, that is represented by the menu-item;
- For some services, overwritable metadata elements from the GenericRequest object (see SFX-menu-screens in Figure 8 , Figure 10, Figure 12 and Figure 14).

When the user clicks a menu-item, the appropriate TargetParser script corresponding to the chosen service and Target is launched. These TargetParsers implement resource-specific link-to syntaxes. They take data from the GenericRequest object as input and compute the URL to which the user will be redirected.

### **Illustrations of project results**

The concrete results of the project are illustrated by Lotus Screencam movies that show how a Ghent and a LANL user navigates in his institutional SFX-aware digital library collection. The Screencams are provided as stand-alone executables that can only be run on WinTel computers. Since the Screencams are large files, their size is mentioned. The Screencams do not contain audio. In addition to these Screencams, some examples are also given by means of screendumps.

#### **Screencam 1: Ghent implementation of BIOSIS (Lotus Screencam executable for WinTel computer; no audio; size 52 Mb)**

The user starts from the Ghent implementation of BIOSIS, and requests the services of the Ghent SFX solution. Services are shown linking from BIOSIS into the Ghent OPAC, into full-text collections, into the LANL implementation of the Science Citation Database, into PubMed and into Ulrich's Serials Directory. At a certain point the user links out to a paper in Cancer, a Wiley InterScience journal that is SFX-aware. Upon request, the user receives similar services from the Ghent SFX solution for citations in that paper. They link him to the LANL OPAC, to full-text at other publishers sites, to PubMed and to the LANL implementation of the Science Citation Database.

**Screencam 2: LANL implementation of BIOSIS** (Lotus Screencam executable for WinTel computer; no audio; size 46 Mb)

The user starts from the LANL implementation of BIOSIS and requests services from the LANL SFX solution. He uses links into the Ghent implementation of Current Contents, into several full-text collections and into the LANL implementation of the Science Citation Database. From the Citation Database, again, he requests SFX-services that lead him into more full-text collections as well as into PubMed. Next, the user returns to his result set in BIOSIS, where he requests services for other records. These lead him into the Journal Citation Reports and to more full-text. For one of the BIOSIS records, the *genome* service appears leading the user to the Genome database.

**Screencam 3: Ghent implementation of Current Contents** (Lotus Screencam executable for WinTel computer; no audio; size 36 Mb)

The user starts from the Ghent implementation of Current Contents, and requests the services of the Ghent SFX solution. The user links out to a paper in JASIS, an SFX-aware Wiley InterScience journal, and -- upon request -- is receiving SFX-services from the Ghent SFX server for citations found in the JASIS paper. Shown are links to PubMed, LISA and to the LANL implementation of the Science Citation Database. For another record from Current Contents, the user links into the Science Citation Database. There, he requests extended services for several records in the result set, which link him to full-text collections etc.

**Screencam 4: Ghent Aleph 500 OPAC** (Lotus Screencam executable for WinTel computer; no audio; size 13 Mb)

The user starts from the Ghent OPAC. He requests SFX-services for a record describing a book and links into Amazon.com. For a journal, he links to the full text collection and to the Journal Citation Reports.

**Screencam 5: LANL Advance OPAC** (Lotus Screencam executable for WinTel computer; no audio; size = 21 Mb)

The user starts from the LANL OPAC and requests services from the LANL SFX solution. For a book record, he links into Amazon.com. For the OPAC record of the journal Cancer he links to the full-text repository and browses towards a paper. In that paper, citations have SFX-buttons since the journal is SFX-aware. The user requests SFX-services for some citations, that bring him to PubMed and to the LANL implementation of the Science Citation Database. For records resulting from the latter service, the user again requests SFX-services that take him to the Journal Citation Reports, the LANL OPAC and to various full-text collections. One of those is -- again -- the Wiley InterScience collection.

**Screencam 6: LANL Inspec** (Lotus Screencam executable for WinTel computer; no audio; size = 29 Mb)

The user starts from the LANL Inspec and requests services from the LANL SFX solution. Services bring him to full-text collections, the LANL OPAC, the Journal Citation Reports. At a certain point the user goes out to the PROLA archive and finds SFX-aware citations for which he requests extended services. Due to the current implementation of the SFX-URL for PROLA little metadata makes it into a GenericRequest object and as such little services result. Back in the Inspec, the user links to the LANL implementation of the Science Citation Database, from which he further links to the LANL OPAC and to full-text collections.

**Screencam 7: the arXiv, consulted by a LANL user** (Lotus Screencam executable for WinTel computer; no audio; size = 19 Mb)

The user searches the Topic interface for the arXiv e-print repository and requests SFX-services for search results. These link him into the Inspec database, searching for references by the authors of the e-prints.

### Screendump example 1:

Figure 8 shows how the link-source record from Figure 3, originating from the Ghent implementation of BIOSIS yields a variety of extended services, amongst other *full\_text*, *holding*, *abstract*, *author*, *cited\_author* and *cited\_reference* that are presented in the SFX-menu-screen. From this menu-screen, the user chooses the full-text link to Wiley, which leads him into an article of the journal Cancer that is SFX-aware (Figure 9). Figure 10 shows how the same user has requested extended services for the third citation in that Wiley article and has received a Ghent SFX-menu-screen. For this citation similar services are available, all linking dynamically from the external Wiley resource back into the Ghent digital library collection. As can be seen, a link to the on-line version of the New England Journal of Medicine has become available. The user chooses to use the *cited\_reference* service that is also available for this citation. This leads him into the SFX-aware LANL implementation of the Science Citation Database (Figure 11) from where he can request extended services, again.

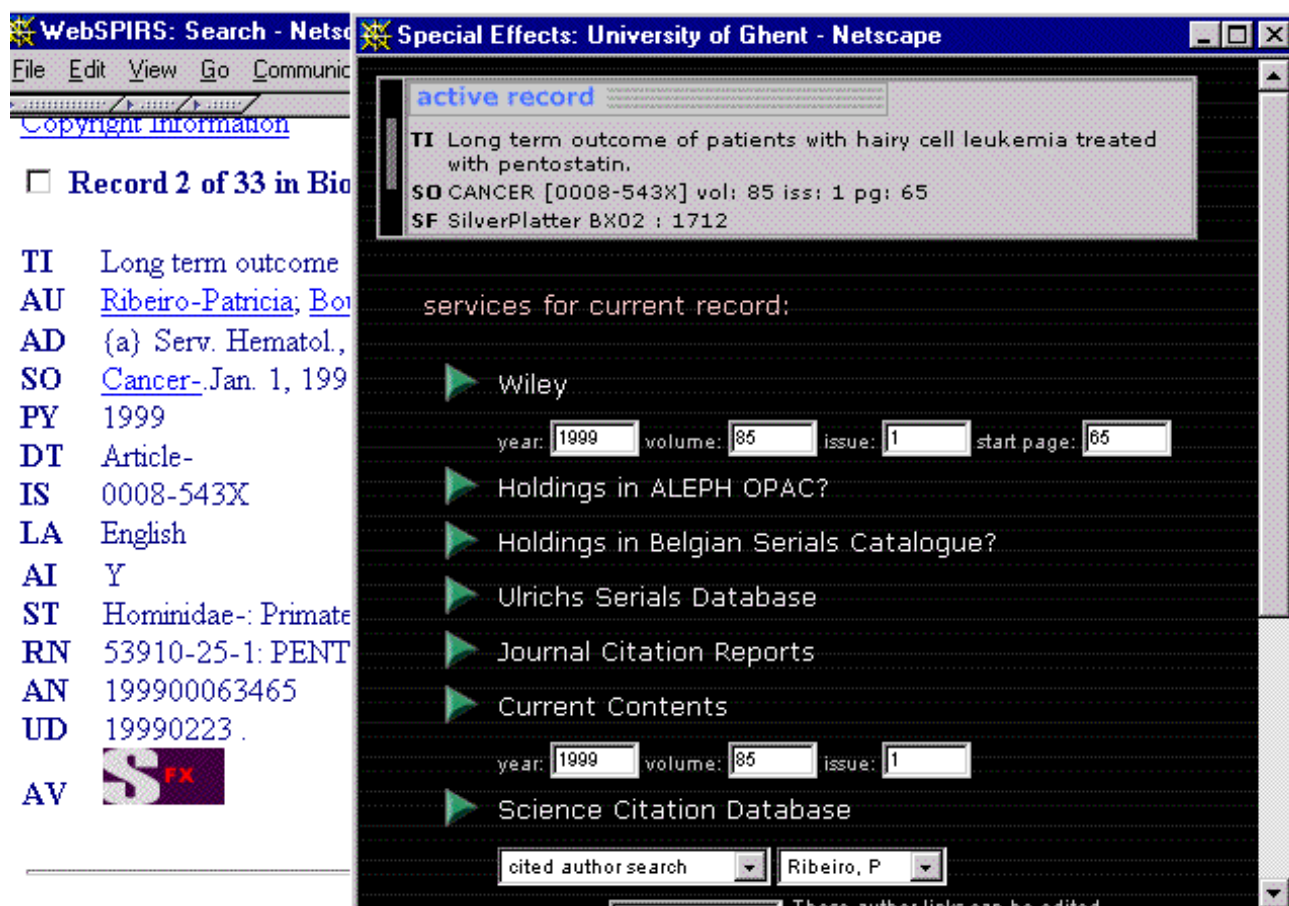


Figure 8: Ghent SFX-menu-screen for link-source from Ghent BIOSIS (record from Figure 3)



WebSPIRS: Search - Netscape

Special Effects: University of Ghent - Netscape

Article Abstract - Netscape

WILEY  
**InterScience®**

PERSONAL HOMEPAGE JOURNAL FINDER SEARCH HELP CONTACT US LOGOUT

ALL JOURNALS PREVIOUS ARTICLE NEXT ARTICLE

**Article Abstract**

**CANCER**

Online ISSN: 1097-0142 Print ISSN: 0008-543X

**Cancer**  
Volume 85, Issue 1, 1999. Pages: 65-71

Original Article

**Long term outcome of patients with hairy cell leukemia treated with pentostatin**

Patricia Ribeiro, M.D.<sup>1</sup>, Fadhela Bouaffia, M.D.<sup>1</sup>, Pierre-Yves Peaud, M.D.<sup>2</sup>, Michel Blanc

**References**

- 1 Saven A, Piro L. Treatment of hairy cell leukemia. *Blood* 1992; **79**: 1111-20. [Medline](#) SFX
- 2 Jaiyesimi I, Kantarjian H, Estey E. Advances in therapy for hairy cell leukemia. A review. *Cancer* 1993; **72**: 5-16. [Medline](#) SFX
- 3 Saven A, Piro L. The newer purine analogues for the treatment of hairy-cell leukemia. *N Engl J Med* 1994; **330**: 691-7. [Medline](#) SFX

ADD HOT ARTICLE

Figure 9: Ghent user follows Wiley *full\_text* service from the SFX-menu of Figure 8

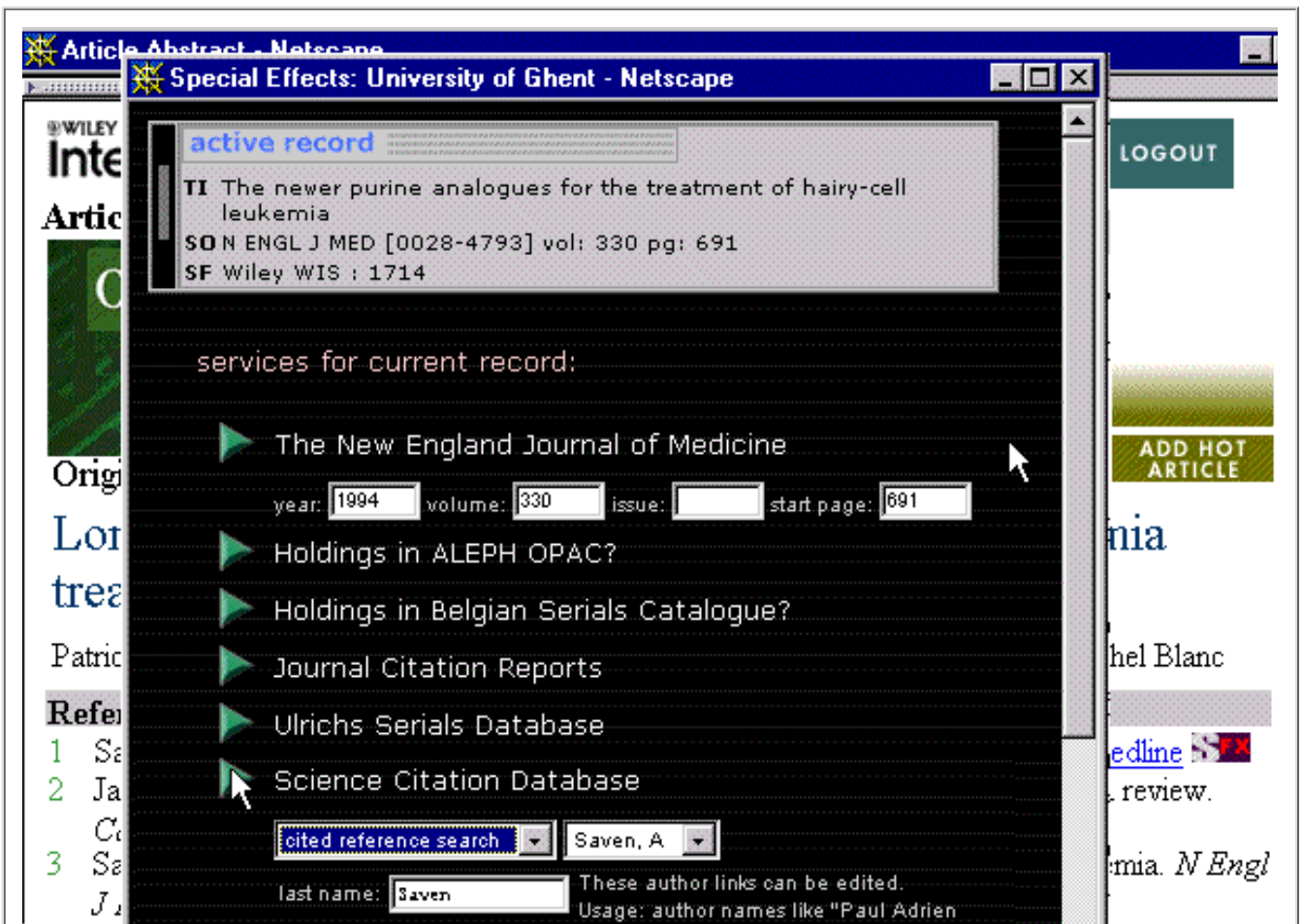


Figure 10: Ghent SFX-menu-screen for link-source from Wiley InterScience (third citation from Figure 9)

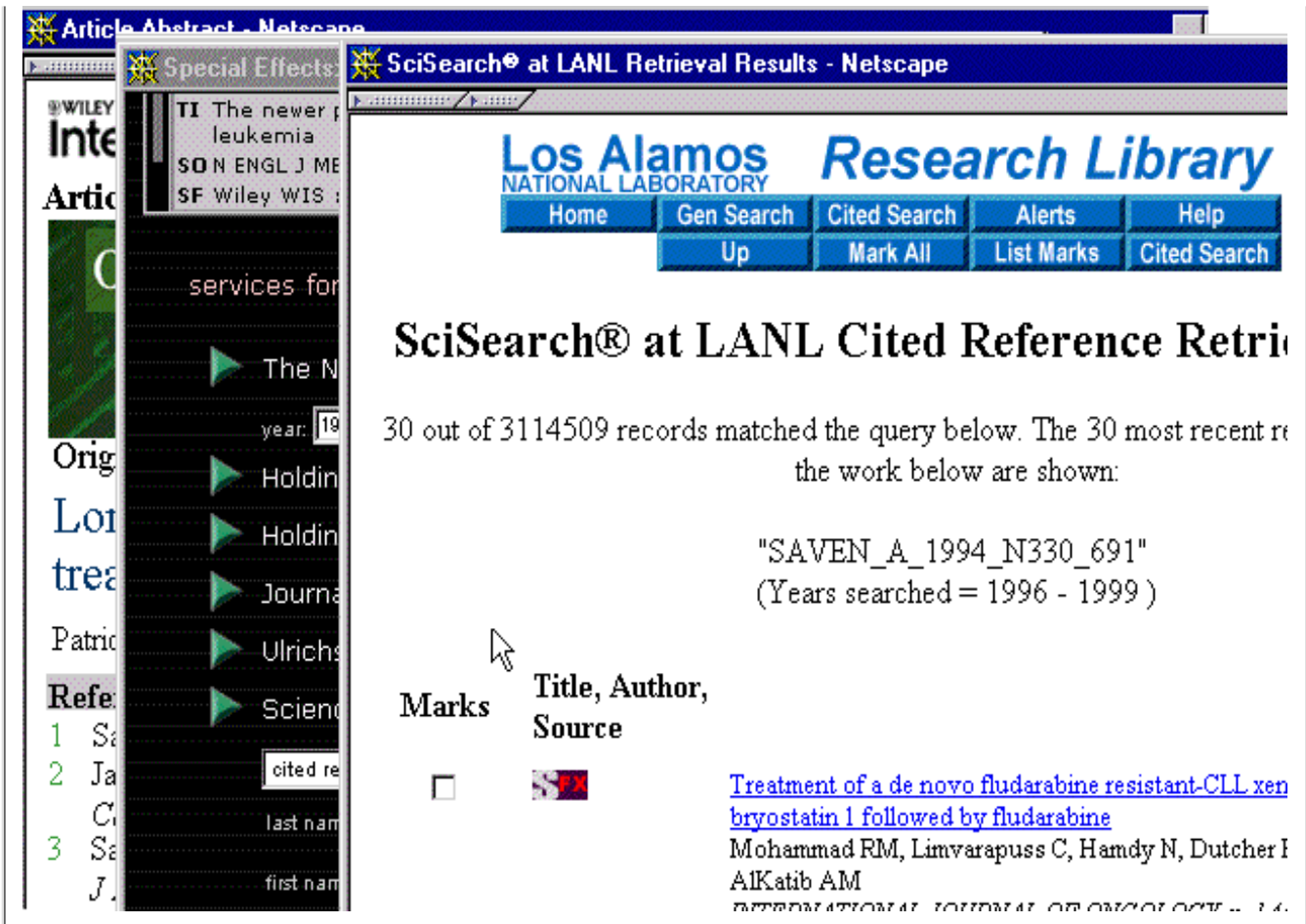


Figure 11: Ghent user follows the *cited\_reference* service from the SFX-menu of Figure 10

### Screen dump example 2:

In Figure 12, the record from Figure 4 originating from the LANL implementation of BIOSIS is used as a link-source. The resulting SFX-menu-screen looks a little different, as an illustration of the fact that another SFX system is being consulted to provide extended services. The LANL localization can also be derived from the OPAC link, that now leads into the Los Alamos Advance catalogue. Another service appears in this screen too: it provides a look-up of sequence information for genome identifiers that were found in the link-source metadata. This service leads the user to the NCBI Genome database using the Entrez link-to syntax (Figure 13).

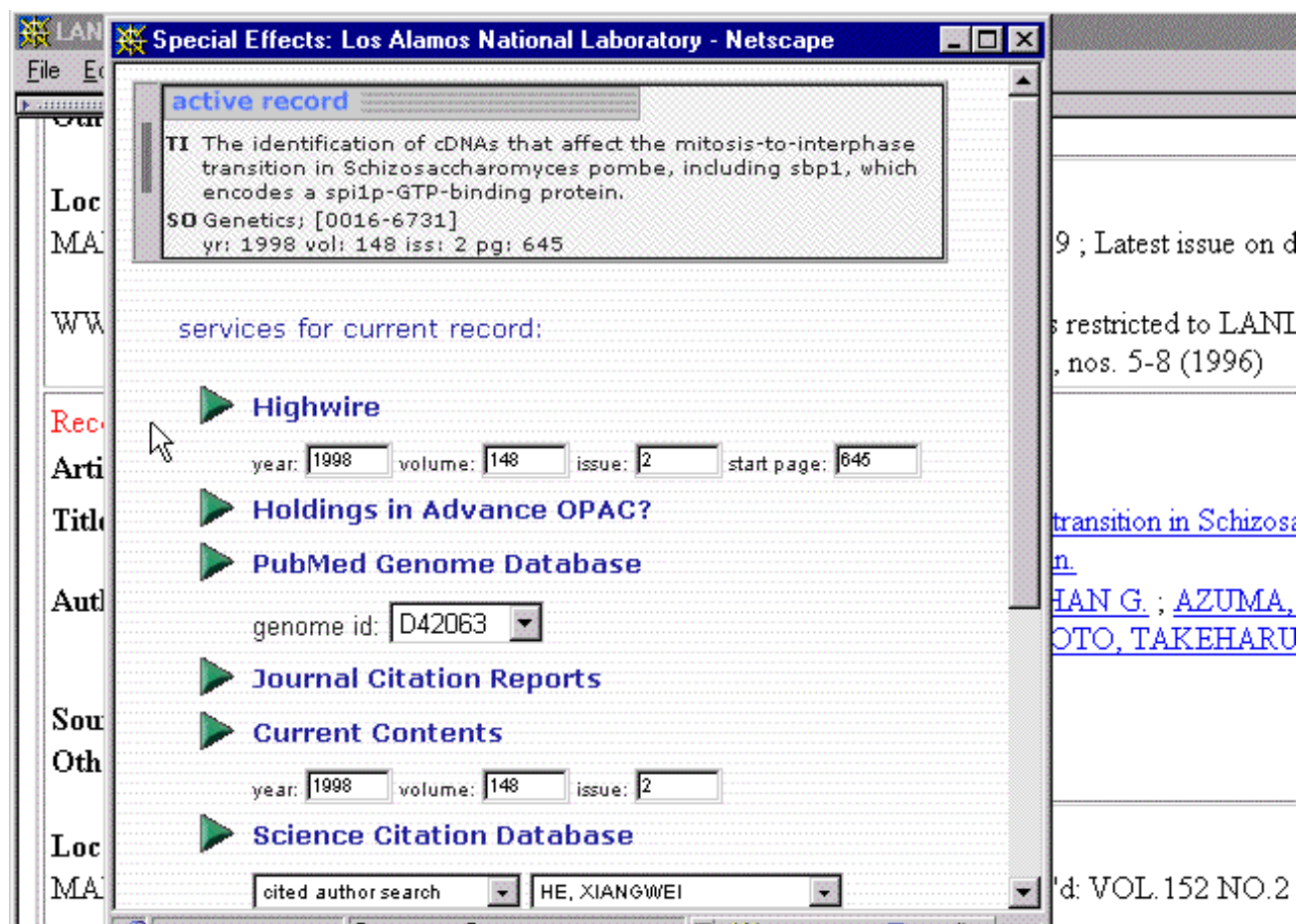


Figure 12: LANL SFX-menu-screen for link-source from LANL BIOSIS (record from Figure 4)

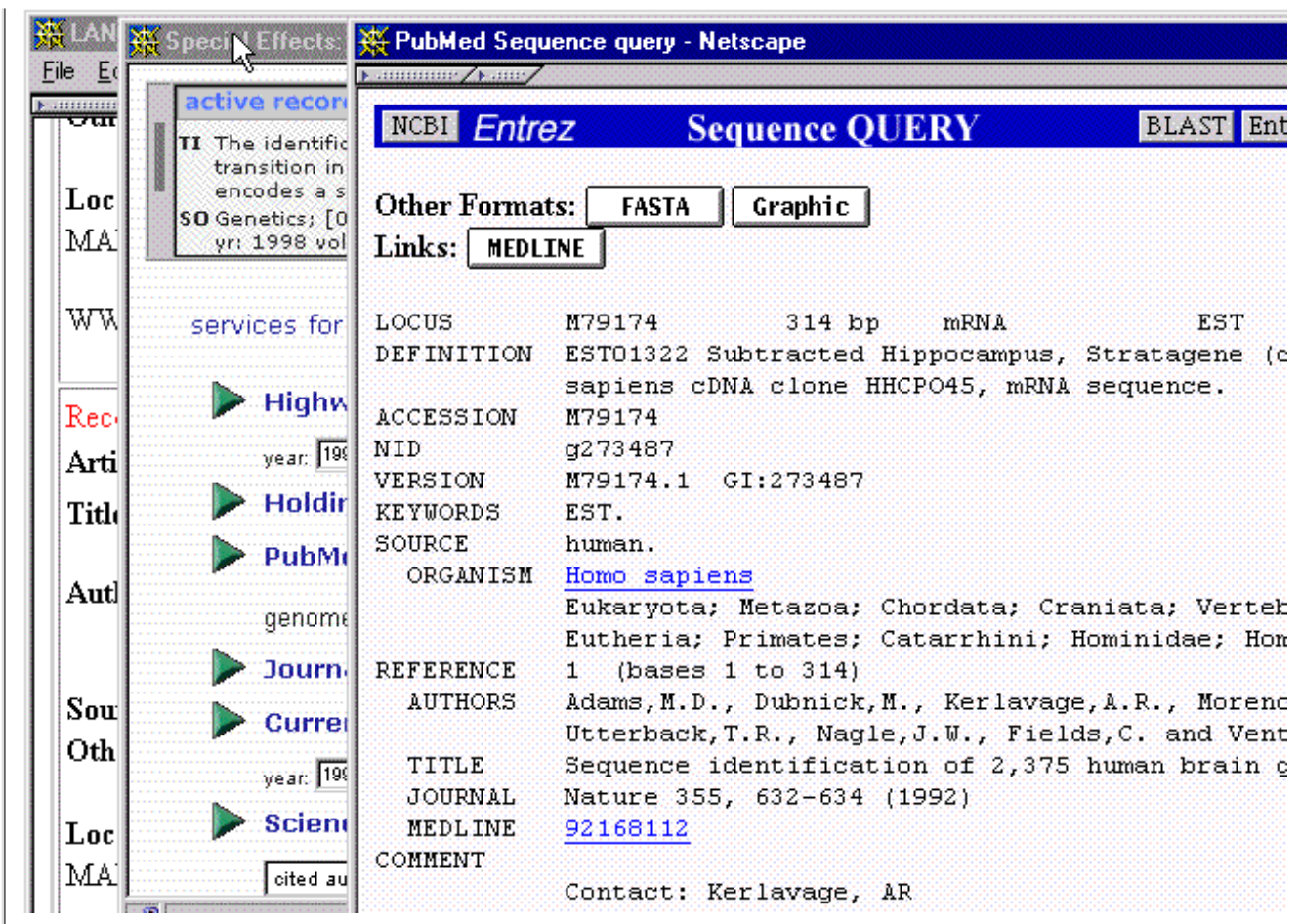


Figure 13: LANL user follows the *genome* service from the SFX-menu of Figure 12

(after selection of identifier M79174 from the pop-down)



### Screendump example 3:

Figure 14 shows the LANL SFX-screen for the first record from the Topic implementation of the arXiv e-print repository shown in Figure 6. Here, an *author* service is available, that provides the option to look-up records in the Inspec database, that abstracts publications authored by the e-print authors as a means to support the evaluation of the reliability of the non-peer-reviewed e-print. The author that is being looked-up has 160 references in the Inspec database (Figure 15).

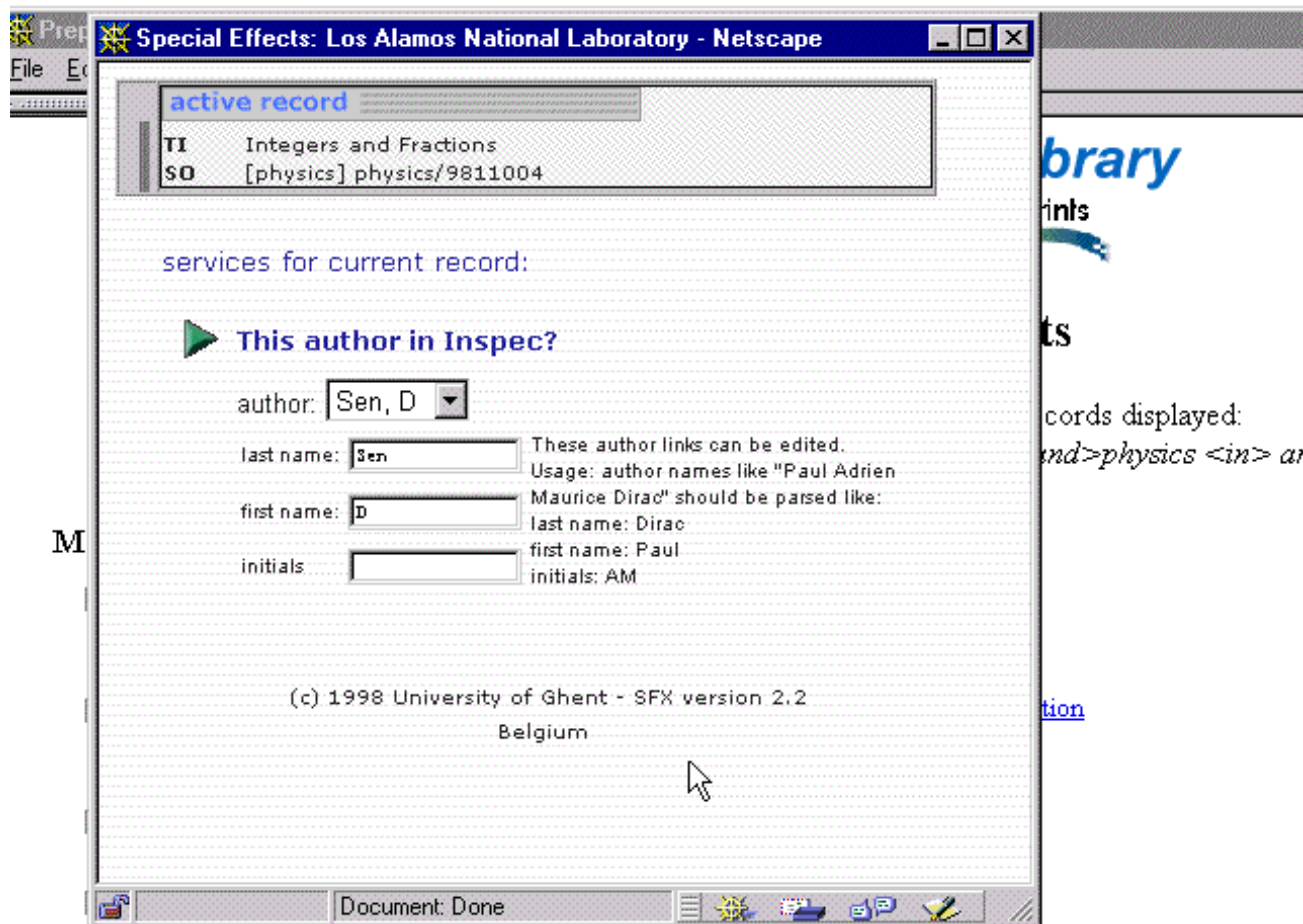


Figure 14: LANL SFX-menu-screen for link-source from the arXiv (first record from Figure 6)

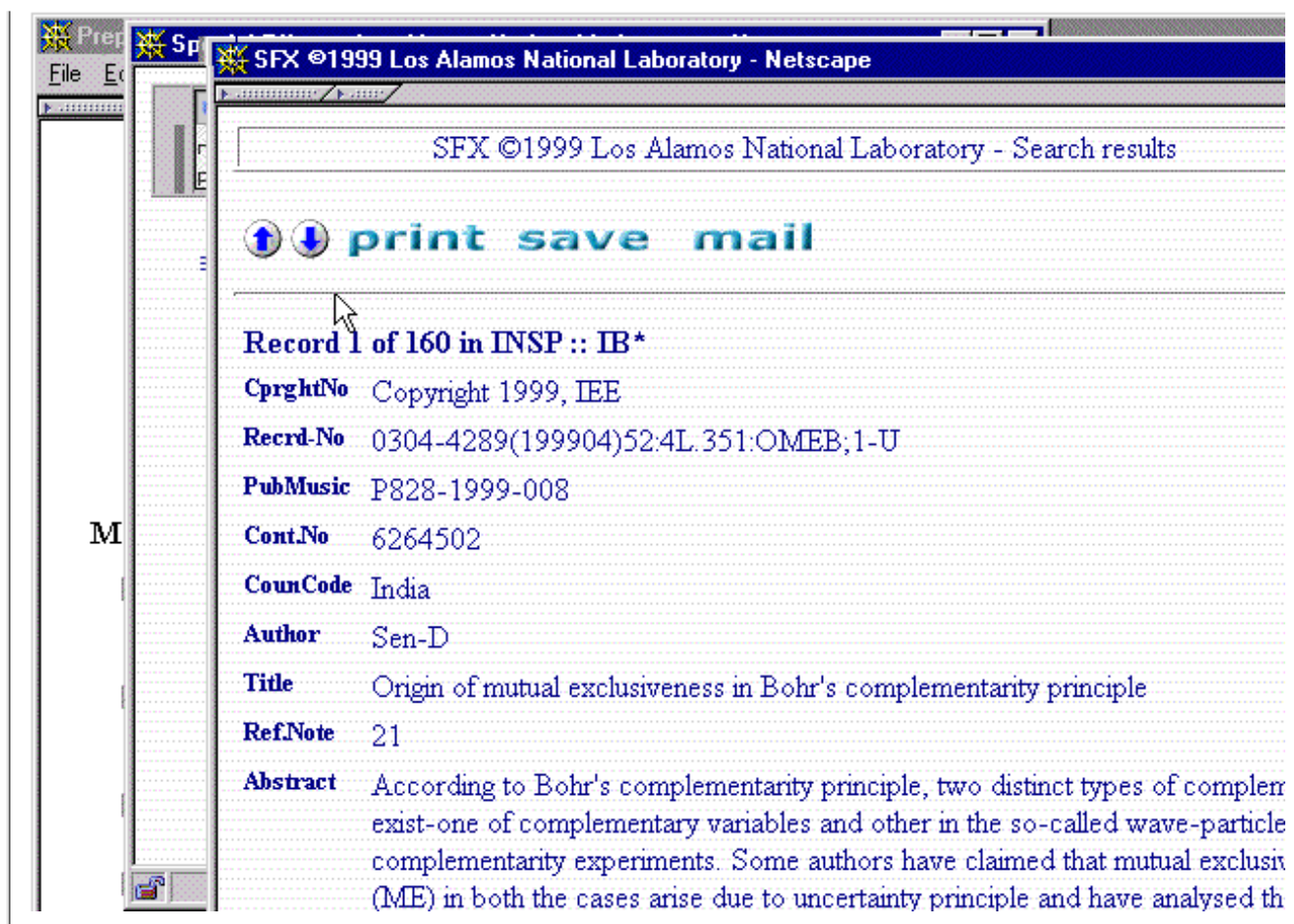


Figure 15: LANL user follows the Inspec *author* service from the SFX-menu of Figure 14

## Comments regarding the SFX redirection component

### *Transferability of SourceParsers*

SourceParsers are easily transferable between different digital library implementations, requiring little or no enhancements to be made. For instance, OPAC systems worldwide support the Z39.50 protocol and respond to requests with MARC formatted records. Making abstraction of the regrettable idiosyncrasies of Z39.50 and MARC implementations, SourceParsers for such OPAC systems can be reused with little editing, apart from the adaptation of Z39.50 parameters such as host, target, port etc... to the local situation. Also, Z39.50 can be used to fetch link-sources from all implementations of databases on a SilverPlatter ERL platform, and -- again -- only the Z39.50 parameters will be different. All implementations of MathSci on such ERL platforms can use the same parsing procedures, making the parsing part of the SourceParser even universal. This is also the case for the SourceParser used for the APS PROLA archive and the Wiley journals, since there is only one implementation of those, worldwide. This approach opens attractive possibilities of sharing SourceParser on a large scale, reducing the overhead in running the solution. Furthermore, it allows information vendors to provide SourceParsers for their resources, keeping full control of the amount of information that they allow to be fetched.

### *The SFX redirection mechanism and namespace specific identifiers*

Important efforts are under way to enable reference linking using DOIs (Paskin 1999a ; Spilka 1999b). Publishers will contribute metadata of their publications along with the corresponding DOIs to the DOI metadatabase. Other publishers can then match references in their own publications to the DOI metadatabase enabling them to insert the DOI of the work represented by the reference next to the reference. Currently, due to lack of support of the handle protocol in mainstream browsers, such a DOI -- say 10.1000/123456789 -- is being hyperlinked as <http://dx.doi.org/10.1000/123456789> and the DOI handle proxy will resolve this link into a single URL, being the

one of the publication at the publisher's site (Paskin 1999b). This hyperlink is a perfect example of a closed link that does not take into account the local context in which the link will be used (see Problem Statement). This closed link causes the Harvard problem to arise, since it does not take into account the possibility of storage of the same work at another -- preferred -- location, for instance the repository of an intermediary or the local institutional full-text warehouse. Also, this mechanism does not allow other, locally relevant extended services to be provided for the reference at hand, since their provision requires the full metadata -- not only the identifier -- of the reference.

The local redirection approach presented by the SFX work does present a pragmatic way to open such a closed linking framework. DOIs can be carried in an SFX-URL pointing at the preferred SFX-server. For instance, for a citation in a Wiley InterScience article that has a static link to DOI 10.1000/123456789 the link can be dynamically rewritten if the existence of a service component is detected. In the case of SFX as a service component, it can become:

<http://isiserv.rug.ac.be/cgi-bin/sfx/bin/menu.cgi?>

[vendorID=Wiley&databaseId=WIS](#) &nameSpace=DOI&

[objectDesc=URLencode\(DOI=10.1000/123456789\)](#)

In essence, such a pragmatic mechanism can redirect the identifier to the redirection component of a selective resolution system, that can decide what to do with it, based on the local context. In the case of SFX, receipt of the above URL causes the launch of a SourceParser. As shown before, typically, this will be the SourceParser corresponding to the serviceDesc part of the URL. Still, upon detection of the "nameSpace" parameter, this default could be overwritten to become the namespace-specific SourceParser, in this example the one for the DOI namespace. This DOI SourceParser would do a so-called reverse look-up in the DOI metadatabase, using the DOI value as a key to fetch the corresponding metadata. Both the fetched metadata and the DOI can then be used in a process to determine the relevant extended services for the citation, including a link to the most appropriate full-text instance. As will be discussed in the next section, for other types of service components, redirection of the DOI without metadata-fetch could be sufficient.

The same mechanism can be used to open the links that are connected to citations carrying identifiers originating from other namespaces such as PubMed and Astrophysics Data System. This can easily be seen by taking the following citation from a Wiley journal that has a static link to PubMed:

Rainer RO, Geisinger KR. Beyond sensitivity and specificity. Am J Clin Pathol 1995; 103: 541-2. Medline

The Medline hyperlink points at:

<http://www4.ncbi.nlm.nih.gov:80/htbin-post/Entrez/query?uid=95259660&form=6&db=m&Dopt=r>

and uses the PubMed ID as a look-up key into the NCBI PubMed. This hyperlink can be pointed at a local resolution solution if it is dynamically rewritten as:

<http://isiserv.rug.ac.be/cgi-bin/sfx/bin/menu.cgi?>

[vendorID=Wiley&databaseId=WIS](#)

[&nameSpace=Medline&objectDesc=URLencode\(Medline=95259660\)](#)

In a similar manner as in the DOI example, this URL can deliver the PubMed ID to a local redirection component. In the case of SFX, this would cause the Medline-namespace SourceParser to be launched, which would fetch the corresponding record from the Medline database. This SourceParser can actually fetch the metadata either from the PubMed implementation of Medline (using the HTTP protocol and the Entrez link-to syntax) or from a local implementation of Medline if one exists. Once that metadata has been turned into a GenericRequest object, the SFX evaluation process can deliver locally relevant extended services. In this way, the initial service of Wiley is augmented with locally relevant services. Also, the quality of the metadata resulting after such a fetch will be higher than that in the original citation, as can be seen from the fact that no issue number is available in the citation of this example, as is common in the medical literature. The fetched record, however, will contain the missing issue number and much more valuable information.

This consideration illustrates that link-source metadata does not necessarily have to be fetched from its origin



resource. In both examples, for a link-source originating in a Wiley journal, the link-source metadata is fetched from the authoritative resource for the namespace of which the link-source carries an identifier.

### ***Modularity of the SFX redirection approach - integration with authentication and authorization efforts***

The current implementation of the SFX local redirection mechanism builds on the CookiePusher mechanism, a consistent SFX-URL and SourceParsers. It can be seen that each of these buildings blocks can be replaced by -- preferably -- more robust alternatives, as long as they perform the same function. The CookiePusher could be replaced by a mechanism in which users register the URL of their local redirection component with information providers. It could also be a parameter that users set in their web-browser preferences and that is recognized by web-servers. The SFX-URL could use a completely different syntax, and its standardization might actually turn the SFX-URL itself into the metadata container interfacing between the redirection component and the service component.

There is a special illustration of the potential of this modularity that is worthwhile to explicitly address, since it offers the promise of combining two important domains of current digital library research: linking and authentication/authorization. Actually, there are some interesting similarities in the approach taken in the SFX linking research and the authentication and authorization effort of the Digital Library Federation (Millman 1999), henceforth referred to as A&A.

The A&A effort takes a certificate based approach to the complex problem of authentication and authorization in a distributed information environment. The user's browser is equipped with a certificate delivered by his institution. When connecting to an information resource, that resource will authenticate the user by asking him for its certificate and by checking its validity. The certificate used in A&A has an extension field, which contains a query URL into an institutional authorization directory server. For each individual of the institution, this authorization server has an entry that contains his authorizations in the various distributed resources of the institutional collection. Such authorizations are described in a triple that consists of the Vendor name, the name of the Service for that Vendor and the access level within that Service. Once an information resource has authenticated the user, it will use the URL contained in the extension of the certificate to query the user's institutional authorization directory server in order to determine his authorization and access level in the resource.

Both efforts - SFX and A&A - have introduced novel mechanisms to solve problems that are commonly addressed via proxying techniques. Both efforts share the conviction that such proxying approaches are not sustainable. Also, both efforts use a local institutional component in their proposed solutions. As such, both deal with the problem of dynamically pointing distributed information resources to that local component. In the Gent&LANL SFX framework, this has pragmatically been addressed by sending the URL of the SFX redirection component to the CookiePusher script. The A&A has taken a more solid but high-entry approach based on certificates. Still, the specific aim of both approaches is the same. The CookiePusher component of the SFX redirection mechanism could easily be replaced via a tie-in with the A&A effort. The URL of the local SFX redirection component could be added to the extension of the A&A certificate, next to the query URL into the local authorization directory server. Or -- even better -- the URL of the SFX redirection component could be part of the response that the authorization directory server sends to the resource that is being connected by the user.

From the perspective of the SFX framework, an important gain of such an approach -- when compared to the CookiePusher -- would be the fact that the certificate mechanism remains valid, even when a user connects to an information resource without going through the aggregated service developed by his institution. Even when connecting directly to an information service rather than going via a menu system that has been enhanced to point at CookiePushers first, will the information on the existence and location of the user's SFX redirection component be accessible for the information resource. Moreover, the certificate approach facilitates the possibility to take into account user preferences regarding the choice of SFX component to be used. Also, SFX ultimately wants to deliver a URL pointing into an information resource. Commonly, authentication is required in order to be able to link into such a resource and theoretically it would be the task of TargetParsers to make this authentication transparent to the user. Current SFX TargetParsers have not dealt with this task, since it is not perceived to be crucial in the demonstration of the feasibility of an open dynamic linking framework: the linking framework does not have to solve the global authentication problem. Still, it can be seen that the SFX framework would benefit from the widespread acceptance of a mechanism for authentication and authorization as proposed in A&A. TargetParsers would be relieved from the cumbersome task of implementing work-arounds to deal with current idiosyncratic authentication solutions.

An additional strong motivation to bundle both efforts is the fact that both need unique persistent identifiers for

scholarly information resources. In A&A these are contained in the Vendor/Service pair; in SFX they are contained in the serviceDesc part of the SFX-URL.

## **Comments regarding the SFX service component**

### ***The impact of a SFX service component building on conceptual services***

The consequences of the introduction -- in the Elektron experiment -- of a linking service built upon a database of conceptual services, have only now become fully evident, due to the complex and distributed nature of the environment in which Ghent&LANL was conducted. Actually, the more resources are added to the environment, the more the elegance and feasibility of the SFX service component become apparent. The introduction of a new SFX-aware resource in an environment requires very limited editing of the SFX-base, in order for all the existing conceptual services that had already been registered in the SFX-base to become immediately available for the new resource too. The dynamic manner in which the SFX system brings up a list of extended services for link-sources of a newly added resource can only be called remarkable. It becomes increasingly difficult to manually predict the outcome of a request for extended services, even when knowing the system and its underlying database in and out, and when studying the content of the link-source record or its GenericRequest object in detail. As an illustration of this, it is noteworthy to mention how SFX delivers a PubMed link for a citation in the Information Science journal JASIS, where Wiley themselves do not have such a link (see Screencam 3 of the examples). This is odd, since Wiley does insert static PubMed links for citations in their journals. Probably they only do so for those with a biomedical subject, because they see insufficient benefit in sending all their citations into the NCBI PubRef process. The dynamic and conceptual SFX approach does not require such precomputational processes and is able to recognize the validity of presenting a PubMed link for a citation in JASIS on the fly. Another remarkable and appealing example is where full-text links presented by SFX lead from a citation in a Wiley journal to an article in another Wiley journal. At the time of the experiment, Wiley did not offer such a linking service within its own collection, even if they fully control the data to implement it. These examples do not only illustrate the strength of the SFX solution, but more importantly, they are a strong indication of the problems of scale of static linking solutions.

### ***The impact of the redesign of the SFX service component***

The new design of the SFX service component, reflecting aspects of global and local relevance, has a considerable impact on the transportability and manageability of the SFX service component. First, the SFX-base needs to be fed with conceptual services of global relevance with appropriate global Thresholds. Usage of resources such as Ulrich's Serials Directory or the public domain JAKE database (Chudnov 1999) significantly reduce the work involved in doing so. Once global services have been configured, the localization of the set-up requires minimal efforts. As an illustration of this, it is interesting to look once again at the BIOSIS-abstract-Medline example. This service was initially localized in Ghent by filling out the names of the local SourceParser for BIOSIS and the local TargetParser for Medline. Both parsers implement the desired connection with the SilverPlatter ERL platform that locally hosts the databases. LANL has a different implementation of BIOSIS, currently running on a Geac Advance system, while no local Medline is available. Therefore, LANL chose to use the PubMed implementation as a Target. When transporting the SFX-base -- that had been localized for Ghent first -- over to LANL, very limited editing of the SFX-base had to be done to activate the BIOSIS-abstract-Medline service in the new environment: the global service and its global Thresholds remained valid. The parser values for the Los Alamos version of BIOSIS and Medline were used to overwrite the Ghent values, as shown in Table 9. The Threshold indicating that Medline is only available from 1985 onwards in the Ghent collection, was overwritten by a 1965 Threshold for LANL. In this case, the local and global Threshold are equal. The elegance of the PubMed Entrez link-to-mechanism and the availability of the complete Medline collection caused Ghent to reconsider the Target -- hence TargetParser -- to be used in favor of the PubMed implementation. Upon this decision, again, very limited editing of the Ghent SFX-base had to be done.

In Ghent&LANL, most of the TargetParsers are implemented as Perl scripts. Towards the end of the experiment, advantage has been taken of the launch of a preliminary version of the S-Link-S Calculator (Openly Inc. 1999). This Calculator is designed to compute URLs based on input metadata and XML templates that describe link-to-syntaxes in an S-Link-S compliant manner (Hellman 1998). As such, SFX TargetParser scripts that perform the computation of the URLs can eventually be replaced by templates. Those templates describe the link-to-syntax and can be used as input for the S-Link-S Calculator. The experiment ended with a hybrid solution, in which the

SFX service component was adapted to dynamically choose between the two mechanism that became available to compute URLs: the TargetParsers and S-Link-S templates. TargetParsers can be shared between different digital library implementations, since many will link into the same resources or family of resources. Again, this reduces the overhead in running the solution. But a tie-in between the SFX and the S-Link-S work, may eventually further diminish the administration of the SFX solution if publishers start and contribute link-to templates and corresponding metadata to the S-Link-S framework. Sharing of TargetParsers would then be replaced by using S-Link-S templates from the framework that has already been set up by Eric Hellman to collect them.

## **Reflection on identifiers, metadata and service components**

It is interesting to further reflect on the nature of the service component and the requirements it imposes on the redirection mechanism of a selective resolution solution. As a starting point, service components that only aim at the delivery of the appropriate full-text instance for a given link-source are considered. Such a service component could operate solely on an underlying database of identifiers, meaning it could be a traditional linking service building on static links between documents. For such service components, it would be sufficient if the redirection mechanism would transfer identifiers only, without bothering about the associated metadata. Such a service component might be sufficient to address the Harvard problem. It might contain a repository of identifiers and locations of full-text for which a local or preferred warehouse other than the default one exists. However, such a repository of identifiers could quickly become very large and difficult to maintain. It can even be considered awkward to maintain such an institutional repository if no full-text is stored locally, but is only being accessed from preferred external aggregators. This consideration points at the desirability of a service component of a different, more abstract nature. Such a service component can build on the logic underlying the distribution of the collection rather than on individual identifiers of material in the collection. Under normal conditions, such logic might tell that all journals of a certain publisher are accessed at a certain warehouse, that certain ISSN numbers have to be accessed from another one, and that an ISSN number has to be accessed in one repository before a certain date and at another one after that date. This level of abstraction drastically reduces the amount of information to be maintained in the service component and hence makes it more scaleable. But it requires metadata of link-sources to act as operators, not only identifiers. Adding to this the fact that the identifiers required to make such a scenario work for link-sources originating from scholarly information resources are not available and will most probably not become universally available any time soon, it must be concluded that service components will have to be able to operate on the basis of metadata in general with identifiers being a special instance of metadata. This imposes a requirement on the redirection mechanism to be able to deliver link-source metadata, not only identifiers.

Adding to the task of the service component the delivery of other extended services, it becomes hard to imagine that a scaleable solution in a highly distributed environment could build on an architecture with a static linking database. Several illustrations of this consideration have resulted from the Ghent&LANL experiment. A more abstract and dynamic service component is required, which will perform some rule-based decision making, that tends towards the evaluation of the relevance of conceptual services as introduced in the SFX work. As will be clear from the SFX experiments, this type of service component requires the availability of link-source metadata in order to be able to function. Again, this imposes a requirement on the redirection mechanism to be able to deliver link-source metadata. This does not mean that identifiers are irrelevant to this type of solution: on the contrary, since it has been shown that identifiers -- from whichever namespace -- are a welcome tool to enable the local redirection component to adequately deliver high-quality metadata. Moreover, when an extended service only requires an identifier for its resolution, then the corresponding service link will be as foolproof for a dynamic linking system as for a static one.

The general conclusion of the above is that, realistically, identifiers will not be sufficient to address the problem of delivering extended services in a distributed digital library collection. Metadata is required for scaleable service components to be able to perform their tasks. Moving from left to right on the scale of service components ranging from traditional static linking systems to dynamic linking systems building on conceptual services, the data required for the service components to adequately do their jobs ranges from identifiers to full metadata.

## **Intermediate conclusions**

The "SFX@Ghent & SFX@LANL" experiment has led to important generalizations of the concepts introduced in the Elektron experiment. It has been shown that two components are essential for systems that enable the

delivery of context-sensitive extended services, also called systems supportive of selective resolution: the redirection mechanism and the service component. An in-depth description of the implementation of both components in the SFX framework has been given. Although both have been discussed in relation to one another, it has also been shown that they act as separate components that exchange information in a unique metadata interchange format. For the experiment, this format was internally defined and inspired on the structure of the GenericRequest object, since -- by lack of non-SFX local redirection components -- interoperability at this level was not an issue. If more local redirection solutions and more local service solutions become available, standardization of this interchange format will become important.

In essence, the SFX redirection mechanism can be combined with a service component of a very different architecture, even one that builds on a static linking database of identifiers. The current implementation of the SFX local redirection mechanism builds on the CookiePusher mechanism, a consistent SFX-URL and SourceParsers. Each of these buildings blocks can be replaced by more robust alternatives, as long as they perform the same function. This property has been illustrated by suggesting a possible tie-in with the authentication and authorization effort of the Digital Library Federation. The investment required to make systems SFX-aware using the CookiePusher and the SFX-URL has been minimized. Still, it will be easier to implement SFX-awareness in resources that deliver information in a dynamic rather than in a static manner. It has been shown that the SFX local redirection mechanism can be used to redirect namespace-specific identifiers to a local service component. This proves the capability of the SFX redirection component to open closed linking frameworks, which is seen as a powerful illustration of the feasibility of the approach.

The SFX service component can also operate with a different redirection method, as long as that supports delivery of link-source metadata and its origin to the service component. The Ghent&LANL information environment, with its many resources and different technologies running those resources, has led to a design in which the SFX linking service has become a totally neutral module in the digital library that can potentially interoperate with every other system in the environment. Its redesign, reflecting the notions of global and local relevance of services, has led to an important reduction in the overhead of running the solution. In addition to that, the possibility to share SourceParsers, TargetParsers and S-Link-S templates further diminishes the administrative overhead.

As far as can be verified, Ghent&LANL has been the first experiment in which bi-directional context-sensitive linking between distributed resources under control of different authorities has been demonstrated in the scholarly communication environment. As can be seen from the examples, the net result of making systems SFX-aware and delivering extended services for link-sources originating from those systems via the SFX-menu-screen, is a fully hypertextual scholarly information environment in which jumping between related distributed resources becomes possible. As with the most renowned hypertext system -- the World Wide Web -- the ease with which this navigation occurs, can lead to getting lost in the information space. At this point, this feature is seen as a compliment to the solution, since no comparable navigational capability has been demonstrated before.

Ghent&LANL has only briefly touched on the feasibility to incorporate the promising, new and subversive mechanisms of scholarly communication (Okerson & O'Donnell 1995) in the open and dynamic SFX linking framework. This has been done by making the Topic implementation of the Los Alamos arXiv e-print server SFX-aware, and by providing an *author* service for it, thus enabling users to look up references to e-print authors in the Inspec database. A further exploration is required to conclusively demonstrate that the SFX framework can be used to dynamically interconnect traditional and novel scholarly information resources in the same way as this has been demonstrated for the interconnection of the traditional scholarly resources in the course of Gent&LANL.

# Part 3: Applying SFX to integrate e-prints with the established scholarly communication system in the UPS Prototype experiment

---

## Introduction

Part 2 described the "SFX@Ghent & SFX@LANL" experiment that led to important generalizations of concepts that had been introduced in the Elektron experiment, described in Part 1. To a large extent, the feasibility to use an open dynamic linking framework to interconnect scholarly information resources as well as the characteristics of such a framework have adequately been demonstrated and described in Part 2. Still, the application of the framework in yet another environment could provide a conclusive proof, especially if that environment would be dramatically different than the ones of the Elektron and "SFX@Ghent & SFX@LANL" experiments. The Universal Preprint Service -- UPS -- Prototype project conducted as a preparation for the first meeting of the Open Archives initiative (Ginsparg, Luce and Van de Sompel 1999) has provided such an environment, that was both different with regard to content of the collection as with regard to its technical implementation.

The main aim of that meeting was to agree on recommendations that would make the creation of end-user services -- such as scientific search engines, recommendation systems and linking systems -- for data originating from distributed and dissimilar e-print archives easier. As a preparation for the meeting, the UPS Prototype project was initiated. The central aim of the project was the identification of the key issues in actually creating an experimental end-user service for data originating from important existing, production archives. It was expected that a better understanding of the problems would facilitate the Santa Fe discussions on making recommendations to archives regarding their openness to cross-archive services. But the UPS Prototype project also provided an excellent testbed to experiment with digital library technologies. Given the active involvement of the author in both the Open Archives initiative and the UPS Prototype, it is evident that one of the technologies being applied in the Prototype was the SFX linking system.

This part gives an in-depth overview of the UPS Prototype project. While only one section and the major part of the Illustrations of Project results are explicitly dedicated to the application of the SFX framework in the Prototype, it is essential to include a description of all aspects of the project as a means to illustrate the complexity that it dealt with and in order to be able to view the project results in the correct perspective. Consistent with the central aim of the UPS Prototype project, important project results are out of the scope of this thesis and as such have not been included. Those results have been brought forward at the first meeting of the Open Archives initiative and are described in detail in (Van de Sompel, Krichel, Nelson, et al 2000). The results had a considerable impact on the Santa Fe Convention (Van de Sompel and Lagoze 2000 (see Appendix) ; Open Archives initiative 2000) that resulted from the meeting. The convention provides a set of relatively simple but potentially quite powerful interoperability agreements that facilitate the creation of mediator services for distributed and dissimilar e-print archives. Within the scope of this thesis are the project results concerning the application of the SFX framework in the UPS environment.

The author started the UPS Prototype project with Thomas Krichel and Michael Nelson. This trio became the coordinators of the project. Each of them brought additional researchers into the project. Most of them have never met in person; project communication has mainly been conducted via a list server. The UPS Prototype project was sponsored by the Research Library of the Los Alamos National Laboratory and by the WoPEc project of the JISC funded e-Lib program. It started around the end of June 1999 and was finalized with a report on the project results given by the coordinating trio as the opening presentation for the Santa Fe Meeting of the Open Archives initiative on October 21st 1999.

## **The UPS Prototype project**

The UPS Prototype project aimed to create the following user services:

- A cross-archive search facility;
- A linking service integrating the archive data with other scholarly information resources.

Additionally, the group wanted to take advantage of the availability of an important dataset to explore a specific archive architecture built around intelligent digital objects. It was hoped that a better understanding of its relation to the creation of the desired end-user services could ultimately also support possible future discussion of the Open Archives initiative on recommendations about the architectural design of archives.

At the beginning of the project, the decision was made to create a multidisciplinary end-user service, as a special instance of a cross-archive service. Outside of the communities of scholars who are aware of the existence of discipline-specific points of entry to e-print information, there is an important market consisting of libraries, students and interdisciplinary researchers for which a multidisciplinary service is most probably a welcome tool. In addition to increasing the accessibility of e-print data, existence of such a service helps raise the awareness regarding the feasibility of alternative communication mechanisms outside a core group that no longer needs to be convinced. Along with the advocating that SPARC is undertaking in this area, concrete illustrations of what is achievable can be important promotional tools.

The amount of cross-archive end-user services is limited (see, for instance, (Plümer and Schwänzl 1997; Plümer and Schwänzl 1996; Canessa and Pastore 1996; Canessa 1996; Powell 1998; Powell and Fox 1998). Most of them are prototypal, do not provide a linking service and do not compare in scale to what the UPS Prototype set out to realize. The Astrophysics Data System is a noteworthy exception. Most of the services are discipline-specific and none of them works across as many initiatives as the UPS Prototype. This makes the UPS Prototype project a challenging, realistic feasibility study since it anticipates that future end-user services will have to deal with the complexity caused by an environment in which discipline-oriented as well as institution-based -- hence multidisciplinary -- archives with dissimilar architectures will co-exist.

## **The archive initiatives included in the UPS Prototype project**

The UPS Prototype project set out to create end-user services for data originating from some major archive initiatives: arXiv.org (commonly known as the the Los Alamos E-Print Archives), Cognitive Sciences Eprint Archive (CogPrints), the Digital Library for the National Advisory Committee for Aeronautics (NACA), the Networked Computer Science Technical Reference Library ( NCSTRL), the Networked Digital Library of Theses and Dissertations (NDLTD) and Research Papers in Economics (RePEc). Table 1 provides links to descriptions of these initiatives as well as to their end-user service(s).

ARCHIVE INITIATIVE	DESCRIPTION	USER SERVICE
ArXiv	(Ginsparg 1994)	<a href="http://arXiv.org/">http://arXiv.org/</a>
CogPrints	(Harnad 2000)	<a href="http://cogprints.soton.ac.uk/search">http://cogprints.soton.ac.uk/search</a>
NACA	(Nelson 1999)	<a href="http://naca.larc.nasa.gov/">http://naca.larc.nasa.gov/</a>
NCSTRL	(Davis and Lagoze 1996)	<a href="http://www.ncstrl.org/">http://www.ncstrl.org/</a>
NDLTD	(Fox et al. 1997)	<a href="http://www.theses.org/">http://www.theses.org/</a>
RePEc	(Krichel 2000a)	<a href="http://netec.mcc.ac.uk/WoPEc.html">http://netec.mcc.ac.uk/WoPEc.html</a> <a href="http://ideas.uqam.ca">http://ideas.uqam.ca</a> <a href="http://netec.wustl.edu/NEP">http://netec.wustl.edu/NEP</a> etc.

**Table 1: Links to the e-print archive initiatives for which data is involved in the UPS Prototype project**

These archive initiatives are dissimilar in many senses, as is illustrated in Table 2 and Table 4:

- **Submission model:** Some archives use a procedure in which material is submitted to a central system. Others handle the submission in a decentralized manner, for instance by submission to distributed systems that are part of the archive initiative.
- **Publication model:** In some archives, authors submit papers directly to the e-print archive. In other archives, the submission is handled at the level of the author's affiliation (organization, department, etc.).
- **Storage facility:** The archive initiatives with a centralized submission mechanism also keep the submitted data in a central repository. Archive initiatives with a decentralized submission procedure store the data in the distributed systems of which the archive initiative consists, but some also create a central mirror that keeps all the data.
- **Native end-user service:** All archives initiatives except RePEc offer a native end-user service. This service can be built around a central index that refers to all the data in the archive initiative. This is the case for all systems with a central storage facility. For others it relies on searching of decentral indexes referring to each of the systems that make up the archive initiative. For NDLTD, each of the decentral indexes must be searched separately as long as the federated search functionality (Powell and Fox 1998) is in an experimental stage. Although NCSTRL originally supported distributed searching, its current production version only supports centralized searching.
- **Third-party service:** For some archives, data is also being made searchable via third-party services. RePEc goes to the extreme in this scenario, by fully relying on third-parties for end-user services.
- **Discipline:** Some archives are discipline-oriented in the sense that knowing that a record originates from a certain archive or sub-archive is equal to knowing its research area. Other archives are multidisciplinary in the sense that such knowledge can not be derived merely from the origin of the record.
- **Scale:** Of the six archives, arXiv is by far the largest with about 130,000 objects in the collection. It also is the most diverse in terms of contributing authors and institutions (Table 4).

ARCHIVE INITIATIVE	SUBMISSION MODEL	PUBLICATION MODEL	STORAGE	NATIVE USER SERVICE	3rd PARTY USER SERVICE	DISCIPLINE ORIENTED
	Central	Author	Central	Central	Yes	Yes
	Decentral	Organization	Decentral	Decentral	No	No
			Mirror			
ArXiv	C	A	C	C	Y	Y
CogPrints	C	A	C	C	N	Y
NACA	C	O	C	C	N	Y
NCSTRL	D	O	D	D => C	N	Y
NDLTD	D	O	D	D, (C)	N	N
RePEc	D	O	D, M	-	Y	Y

Table 2: characteristics of archives involved in the UPS Prototype project

## The phases of the UPS Prototype Project

The different phases of the UPS Prototype project are:

- Data gathering;
- Metadata conversion;
- Creation of SODA archives;
- Creation of NCSTRL+ end-user search facility;
- Addition of a SFX linking service.

### *Data gathering*

At a first stage of the project, data is collected from the originating archive initiatives. This data can then be stored in a new repository that becomes the subject of the end-user services to be created. For each of the archives, the data is collected at a fixed moment in time, around July 1999. Updates to the originating archives are not reflected in the new repository: the UPS prototype builds on static dumps of archive-data. Only the metadata is collected; the full-content associated with that metadata is left in the originating archives and hence, the end-user service will have to point at the full-content there.

In order to obtain the archive data, the maintainers of the archives are contacted. In the course of these contacts, it becomes apparent that -- willingly or unwillingly -- most archive initiatives do not have clear indications on the terms and conditions for usage of their data. But all of them agree to make their metadata available for experimental purposes. The issue then becomes how to get the metadata out of the archives. In addressing this problem, an insight is gained in the mechanisms for metadata extraction supported by the archive initiatives. It turns out that some archives do not have protocols to support harvesting and as such a single static dump of data is delivered by the administrator of the archive. Other archives do provide such protocols but differ in the richness of the criteria that are available for metadata extraction as well as in the publication of these features. All archives in this latter group support accession date as a harvesting criterion, making periodic gathering of updates feasible. Other harvesting criteria that occur are subject and author-affiliation. Still, because some archives do not support a harvesting mechanism and because those that do require different protocols, the archive metadata is collected only once; the archives are not polled for updates afterwards. An overview of the above is given in Table 3.



ARCHIVE INITIATIVE	SINGLE DUMP	HARVEST		TERMS & CONDITIONS
		critierion	documented	documented
	Yes	Subject	Yes	Yes
	No	Date Affiliation	No	No
ArXiv	Y	S,D	N	N
CogPrints	Y	-	-	N
NACA	Y	-	-	-
NCSTRL	Y	D	Y	N
NDLTD (*)	Y	-	-	-
RePEc	Y	S,D	Y	Y

**Table 3: Possibilities for data extraction of archives involved in the UPS Prototype project**

(\*) From now onwards, information regarding the NDLTD initiative will refer to data originating from the Virginia Tech since only data from that NDLTD-node is involved in the experiment.

Table 4, presents some figures related to the data harvested for the project. The "Records harvested" column shows the total amount of records resulting from the data-gathering phase for each archive initiative. The meaning of the other columns of Table 4 will be addressed in the remainder of this paper.

ARCHIVE INITIATIVE	RECORDS HARVESTED	ReDIF RECORDS	BUCKETS IN UPS	BUCKETS LINKED TO FULL CONTENT	UNIQUE AUTHOR AFFILIATIONS
ArXiv	128,943	85,204	85,204	85,204	17,983
CogPrints	743	743	742	659	14
NACA	3,036	3,036	3,036	3,036	100
NCSTRL	29,690	29,690	25,184	9,084	93
NDLTD	1,590	1,590	1,590	951	1
RePEc	71,359	71,359	71,359	13,582	2,453
<b>Total</b>	<b>235,361</b>	<b>191,622</b>	<b>187,115</b>	<b>112,516</b>	<b>22,844</b>

**Table 4: Figures regarding the amount of collected and processed records**

### **Metadata conversion**

The overall quality of the metadata available for the creation of the user services undoubtedly has an important impact on the quality and types of services that can be created. Therefore, during the metadata conversion phase, important efforts are undertaken to augment the quality of the metadata.

### The choice for a single metadata format

The metadata collected during the data gathering phase is expressed in a variety of metadata formats, as shown in Table 5. It turns out that there are as many metadata formats as there are archives. The reasons for this can be explained when taking into account the context of the creation of the archives: the initial motivations for setting up an archive initiative, the community to be served, the environment in which the archive emerges, etc. For instance, the arXiv.org metadata format clearly illustrates the intention to avoid overhead in the author self-submission mechanism that could prevent authors from actually submitting. The result is a format lacking some essential subtagging of metadata fields. On the other hand, the university library is actively involved in the NDLTD effort at Virginia Tech, which leads to the usage of the elaborate MARC format.

ARCHIVE INITIATIVE	NATIVE METADATA FORMAT
arXiv	internal_old; internal_new
CogPrints	internal
NACA	refer (Lesk 1978)
NCSTRL	rfc-1807
NDLTD	MARC
RePEc	ReDIF version 1

**Table 5: metadata formats for data collected from archives**

All data is converted into a single metadata format. This is a useful exercise to obtain an in-depth insight in the peculiarities and problems with the delivered metadata formats. This insight enables identifying aspects of the metadata that can lend themselves to data-augmenting procedures. In addition, converting the metadata to a single format removes an unnecessary complication from the creation of the intended end-user service. It would indeed be possible to create an end-user service based on data expressed in heterogeneous metadata formats. But this would introduce an extra complication in the phase of the creation of the end-user service, which would actually draw the attention away from the essential aims of that phase.

The ReDIF version 1 format (Krichel 2000b) -- as used in the RePEc initiative -- is chosen to be the common metadata format for the UPS Prototype project. Consequently, conversion procedures will have to perform a mapping of non-ReDIF metadata to the ReDIF format. The following motivates the choice for ReDIF:

- ReDIF is designed in such a way that it can easily be extended with non-native ReDIF-fields. As such, during the mapping process, fields can be added if ReDIF does not provide appropriate native fields in which the non-ReDIF data can be stored;
- ReDIF is a rich format. Converting it to one of the other formats would result in a downgrade of the quality of the metadata;
- There is an important set of software tools to manipulate data expressed in the ReDIF format;
- ReDIF is under direct control of a researcher of the coordinating trio, allowing for quick decision making regarding possible required enhancements.

Table 6 shows a sample record expressed in the ReDIF format.

Template-Type: ReDIF-Paper: 1.0

Title: Forecasting market shares using VAR and BVAR models: A comparison of their forecasting performance

Author-Name: Francisco F. R. Ramos

Author-Email: fframes@fep.up.pt

Author-Workplace-Name: Faculty of Economic, University of Porto

Note: Type of original Document - Winword 2.0; prepared on IBM PC; to print on HP/Epson; figures: included.  
Word document submitted by ftp

Length: 41 pages

Keywords: Automobile market; BVAR models; Forecast accuracy; Impulse response analysis; Marketing decision variables; Specification of marketing priors; variance decomposition; VAR models

Classification-JEL: C11; C32; M31

Abstract: This paper develops a Bayesian vector autoregressive model (BVAR) for the leader of the Portuguese car market to forecast the market share. The model includes five marketing decision variables. The Bayesian prior is selected on the basis of the accuracy of the out-of-sample forecasts. We find that our BVAR models generally produce more accurate forecasts of market share. The out-of-sample accuracy of the BVAR forecasts is also compared with that of forecasts from an unrestricted VAR model and of benchmark forecasts produced from univariate (e.g., Box- Jenkins ARIMA) models. Additionally, competitive dynamics of the market place are revealed through variance decompositions and impulse response analysis.

Creation-Date: 19960123

File-URL: <http://econwpa.wustl.edu/econ-wp/em/papers/9601/9601003.ps>

File-Format: Application/PostScript

File-URL: <http://econwpa.wustl.edu/econ-wp/em/papers/9601/9601003.ps.Z>

File-Format: application/postscript/unixcompressed

File-URL: <http://econwpa.wustl.edu/econ-wp/em/papers/9601/9601003.pdf.Z>

File-Format: Application/pdf/unixcompressed

File-URL: <http://econwpa.wustl.edu/econ-wp/em/papers/9601/9601003.pdf.zip>

File-Format: Application/pdf/zipped

File-URL: <http://econwpa.wustl.edu/econ-wp/em/papers/9601/9601003.pdf.gz>

File-Format: application/pdf/gnuzipped

File-URL: <http://econwpa.wustl.edu/econ-wp/em/papers/9601/9601003.pdf>

File-Format: application/pdf

Handle: RePEc:bob:wuwpem:960100

**Table 6: a sample record expressed in the ReDIF paper template**

### The conversion to ReDIF

During the conversion to ReDIF, significant efforts are made to increase the quality of the metadata in order to achieve a level that is suitable for the creation of effective end-user services. Hereafter, some major issues arising during this process are discussed.

### *Achieving an appropriate level of subtagging of metadata elements*

Several input metadata formats do not have the same level of subtagging as ReDIF does. This is especially the case for author and author affiliation data. In some formats, multiple authors are provided in a single field, where ReDIF expects repeated Author-Name fields, one per author. Similarly, some formats provide author affiliation as part of the author field, where ReDIF expects a separate Author-Workplace-Name field. In the internal\_old format of arXiv, title, author as well as author affiliation information is combined into a single field.

In order to address these problems, the native data is parsed via several routines that use heuristics to try and achieve the desired level of subtagging. The difference in the amount of input and ReDIF records for arXiv -- as shown in Table 4 -- is the result of the decision to discard the arXiv data expressed in the internal\_old format, because the desired subtagging could not adequately be achieved within the available timeframe.

Table 7 shows how conversion routines restructure author and author-affiliation information for an input record from arXiv into the ReDIF structure. Other records in arXiv or other archives may have other ways of dealing with multiple authors and multiple affiliations, requiring a lot of attention in the conversion phase.

authors: M. J. Drinkwater (1) M. D. Gregg (2) ((1) University of New South Wales, (2) University of California, Davis, and Institute for Geophysics and Planetary Physics, Lawrence Livermore National Laboratory)
<b>author field for a record from arXiv</b>
Author-Name: M. J. Drinkwater
Author-Workplace-Name: University of New South Wales
Author-Name: M. D. Gregg
Author-Workplace-Name: University of California, Davis, and Institute for Geophysics and Planetary Physics, Lawrence Livermore National Laboratory
<b>ReDIF version of the author information after procedural conversion</b>

**Table 7: conversion of an arXiv author field to a ReDIF representation**

### *Creation of a UPS namespace of unique identifiers*

Within ReDIF, items -- such as metadata records, authors and institutions -- receive unique identifiers. These identifiers have a hierarchical structure. For records that describe resources, e.g. e-prints or software, ReDIF has four levels of identification that reflect the distributed nature of the RePEc initiative. They are:

- The *authority*, which refers to the agent that enforces the coherent identification scheme or to the namespace which results from this identification scheme;
- The *archive*, which refers to the place where archive data is physically maintained;
- The *series*, which refers to a group of resources within an archive;
- The *item*, which refers to a metadata record within a series.

As such, each metadata record in ReDIF receives a unique identifier of the form *authority:archive:series:item*. For the RePEc environment, that uses the ReDIF metadata format, the *authority* always is RePEc.

It is decided to accord unique UPS identifiers, inspired on this structure, to metadata records originating from the other archives too. This is done because the ReDIF tools that will be used require such identifiers. But -- more importantly -- it is anticipated that unique identifiers for all records in the UPS Prototype collection will be important for the creation of the SODA archives and for the addition of the

SFX linking service. Both will be discussed in other sections of this paper.

As such, existing identifiers are restructured to become compliant to the ReDIF identifier format and to guarantee uniqueness within the UPS collection. Actually, a UPS namespace is created. The condition imposed by the RePEc community that archive and series identifiers must be non-meaningful, fixed length strings is not maintained. That would require renaming certain portions of the existing identifiers of other initiatives, which is considered to be politically incorrect. Consistent with the logic of the ReDIF identifier, the *authority* and the *archive* identifiers are chosen to be the same for centralized archives. Some archives, such as the decentralized NCSTRL and NDLTD do not have the notion of sub-archive (*series*) within an *archive* and therefore, the *series* part is chosen to be void. Table 8 shows some original identifiers and their ReDIF-ed representations.

ARCHIVE INITIATIVE	ORIGINAL IDENTIFIER	UPS IDENTIFIER
ArXiv	astro-ph/990686	xxx:xxx:astro-ph:990686
CogPrints	cogprintscomp/199806013	cogprints:cogprints:cogprintscomp:199806013
NACA	naca-tn-3929	NTRS:NACA::naca-tn-3929
NCSTRL	ncstrl.vatech-cs.TR-92-39	ncstrl:vatech-cs::TR-92-39
NDLTD	etd-32298-134558	NDLTD:VTETD::etd-32298-134558
RePEc	RePEc:wuk:mcpmdp:007	RePEc:wuk:mcpmdp:007

**Table 8: creating unique identifiers for the UPS prototype environment**

### *Dealing with multiple manifestations of a work*

A general problem regarding the handling of metadata referring to multiple instances of the same work (Paskin 1999a) also arises in this conversion phase. An e-print is very commonly only the first manifestation of a work that is followed by other publications such as a peer-reviewed paper, a conference proceeding, a chapter in a book etc. Several input metadata formats provide specific fields to contain information regarding the life after the e-print within the metadata record describing the e-print. ReDIF, however, would typically create a new metadata record for each manifestation and would use the template matching the publication type. This policy illustrates the ReDIF view on metadata, but also the fact that the RePEc initiative sets out to describe the discipline of Economics as a whole rather than only the e-prints that it uses. However, in the context of the experiment, splitting up e-print metadata records into multiple ReDIF records is not appropriate, mainly because the information provided in the specific fields reserved for published manifestations is commonly too limited to enable the creation of a representative separate record. For instance, the process could include guessing whether the peer-reviewed paper has the same title and same authors as the e-print. Taking advantage of the extensibility of the ReDIF format, it is decided to use the paper template (see Table 6) as a basis and to add a Manifestation metadata-cluster -- that can consist of several metadata fields -- to it (see Table 9). Each such cluster will refer to manifestations of the same work published after the e-print. It will become especially important for the SFX linking service, since it will allow for the creation of separate linking features for the e-print and for the published manifestations.

<b>version:</b>	1.1	<b>Template-Type:</b>	ReDIF-Paper 1.0
<b>authors:</b>	Crusio, Wim E.	<b>Author-Name:</b>	Crusio, Wim E.
<b>e-mail:</b>	crusio@citi2.fr	<b>Author-Email:</b>	crusio@citi2.fr
<b>catcode:</b>	bio.bio-socio	<b>Series:</b>	cogprintsbio
		<b>Classification-Ila:</b>	Sociobiology
<b>abstract:</b>	Mealey's evolutionary reasoning is logically flawed. Furthermore, the evidence presented in favor of a genetic contribution to the causation of sociopathy is overinterpreted. Given the potentially large societal impact of sociobiological speculation on the roots of criminality, more-than-usual caution in interpreting data is called for.	<b>Abstract:</b>	Mealey's evolutionary reasoning is logically flawed. Furthermore, the evidence presented in favor of a genetic contribution to the causation of sociopathy is overinterpreted. Given the potentially large societal impact of sociobiological speculation on the roots of criminality, more-than-usual caution in interpreting data is called for.
<b>from:</b>	Wim E Crusio	<b>X-Cogprints-Submitter-Name:</b>	Wim E Crusio
<b>title:</b>	The sociopathy of sociobiology	<b>Title:</b>	The sociopathy of sociobiology
<b>userid:</b>	CrusioW	<b>X-Cogprints-Userid:</b>	CrusioW
<b>file-html-main:</b>	bbsmeal.htm	<b>File-URL:</b>	<a href="http://cogprints.soton.ac.uk/archives/bio/papers/199805/199805001/doc.html/bbsmeal.htm">http://cogprints.soton.ac.uk/archives/bio/papers/199805/199805001/doc.html/bbsmeal.htm</a>
		<b>File-Format:</b>	text/html
		<b>File-Function:</b>	Main file
<b>context:</b>	pjour	<b>Manifestation-Type:</b>	prar
<b>pages:</b>	552	<b>Manifestation-Pages:</b>	552
<b>pubn:</b>	Behavioral and Brain Sciences	<b>Manifestation-Journal-Title:</b>	Behavioral and Brain Sciences
<b>volume:</b>	18	<b>Manifestation-Journal-Volume:</b>	18

<b>year:</b>	1995	<b>Manifestation-Journal-Year:</b>	1995
<b>number:</b>	3	<b>Number:</b>	199805001
<b>idcode:</b>	bio/199805001	<b>Handle:</b>	CogPrints:cogprints:cogprintsbio:199805001
		<b>Order-URL:</b>	<a href="http://cogprints.soton.ac.uk/abs/bio/199805001">http://cogprints.soton.ac.uk/abs/bio/199805001</a>

**Table 9: a CogPrints record describing two manifestations of a work and its ReDIF version for UPS**

#### *Addition of a normalized Manifestation-type value*

A Manifestation type field is added to the ReDIF format, in order to store normalized values referring to the type of publication described by the metadata (see Table 9). Taking advantage of the hierarchical properties of the ReDIF framework, for some archives, Manifestation type values can be added at the level of the *authority* to be inherited by all *items* falling under the authority. For other archives, existing Manifestation type values at the *item* level are mapped into their normalized representations.

#### *Preservation and addition of subject classification and keywords*

Several actions are undertaken to preserve the subject classification and keywords included in the various archives during the ReDIF conversion process. Also, each input record is accorded a subject-classification taken from a broad multi-disciplinary subject-classification scheme.

##### → Preservation of the existing subject classification

Subject classification schemes that are native to various input datasets are preserved. In order to accommodate for those in ReDIF, a *Classification-name-yyyy* tag is introduced, where *name* is an identifier for the classification scheme and *yyyy* is the revision year of that scheme:

- Classification-ACM-1991: The classification scheme used by the Association for Computing Machinery, in its version of 1991 (ACM Publications Dept. 1991);
- Classification-MSM-1991: The Mathematics classification scheme devised by the American Mathematical Society, in its version of 1991 (American Mathematical Society 1999).
- Classification-JEL: The classification system of Journal of Economic Literature (American Economic Association 1999) that is commonly used to classify economics texts. This scheme is used in RePEc and the year qualifier is omitted for historical reasons.

##### → Preservation of existing keywords

Author supplied keywords that do not follow a controlled vocabulary are mapped into the ReDIF Keywords tag at the item level. This is the case for keyword data from arXiv, CogPrints, NCSTRL, NDLTD and RePEc.

##### → Addition of a broad subject-classification

In addition to the preservation of the classification schemes and keywords used in the input archives, an attempt is made to add a broad classification to the complete collection. It is decided to use a scheme from NASA, as proposed in (Tiffany and Nelson 1998), and to create the *Classification-Ila* tag to hold the classification term(s).

The hierarchical structure of the ReDIF metadata simplifies the implementation of such a broad classification. It allows expressing a shared subject-classification value for all records of a *series*, of an *archive* or evening an *authority*. The classification accorded to the higher level is then inherited by all elements that are lower in the hierarchy:

- For RePEc data, a single Ila Classification for "Economics" is added at the level of the RePEc

*authority.*

- For NCSTRL data, a single Ila Classification for "Computer Science" is added at the level of the NCSTRL *authority*.
- For NACA data, a single Ila Classification for "Aeronautics" is added at the level of the NTRS *authority*.
- For CogPrints, a single Ila Classification is added at the level of the *series*. That classification is the representation of the discipline of a CogPrint sub-archive in the Ila Classification (see Table 9).
- For arXiv, the Ila Classification is added at the level of the item. An item can receive multiple Ila classifications reflecting the fact that documents in arXiv can be cross-posted to several discipline-specific sub-archives. The added classifications are actually the representations of the disciplines of the sub-archives in the Ila classification scheme.
- For the multidisciplinary NDLTD data, no Ila Classification is added, because it proves to be impossible to add it in a procedural manner within the timeframe of the project.

### *Removal of duplicate records*

The NCSTRL framework harvests data submitted to the CoRR sub-archive of arXiv. As such, those records are provided twice in the input dataset. It is decided to remove the CoRR records from NCSTRL and to maintain the original ones from arXiv. Removal of the records is facilitated because the NCSTRL records carry native identifiers that reveal their provenance.

### *Dataset remains non-optimized*

While it has been shown above that important steps are undertaken to try and achieve an appropriate level of metadata quality, the short project time frame prevents several issues from being addressed:

- Subtagging of author names into last name, first name and initials is not performed, although it is crucial for user-interfaces that have an index-browsing feature;
- No attempt is made to normalize several representations of an author name or author affiliation to a single one. Table 4 shows that the amount of lexically unique author affiliations would clearly justify an exploration of normalization techniques (French, Powell, and Schulman 1997);
- No attempt is made to try and achieve a level of consistency in the syntax for the information referring to other instances of the work, as included in the Manifestation tag.

As will be shown, these and similar issues not addressed during the metadata conversion phase have an impact on the creation of end-user services.

### **Creation of SODA Archives**

Once the input data is converted to ReDIF, it is moved to archives with a Smart Object Dumb Archive (SODA) architecture (Maly, Nelson, and Zubair 1999). The fundamental concept underlying such archives is the transfer of intelligence away from the digital library towards the data objects in the digital library. In a SODA environment, the data objects are called buckets. Their features are described in detail in (Nelson et al. 1999). Buckets are object-oriented, aggregative, intelligent digital objects optimized for use in digital libraries. The basic units of a bucket are referred to as *elements* that are aggregated into *packages*. *Packages* can be further aggregated to contain other *packages* or *elements*. Typically, *elements* are the actual files (pdf, ps, doc, etc.) representing papers, reports, data, or programs. Buckets are designed to be self-contained and mobile, carrying inside themselves all the code and functionality required for operation. For instance, buckets do not need digital library software to display their content: they carry the software required to self-display their own content. Additionally, buckets are designed to be heterogeneous and can grow and acquire new content and features over time. Current buckets need only an CGI-enabled HTTPD server to function. Communication with buckets occurs through bucket methods, which are invoked using HTTP as a transport protocol. While the bucket concept originates in the NCSTRL+ research project (Nelson et al. 1998), it must be regarded independent of digital library protocols and systems.

Individual SODA archives are created for arXiv, CogPrints, NACA, NCSTRL, NDLTD and RePEc. The ReDIF metadata files are used as seeds for the creation of buckets. To create the important amount of



buckets in a batch manner, a script takes each ReDIF file and creates a bucket around it:

- A bucket template, predefined per archive, is untarred, gunzipped, and renamed to an appropriate bucket name that reflects the unique UPS identifier created in the metadata conversion phase.
- The ReDIF file is placed in the bucket, as an *element* in the metadata *package*.
- The ReDIF file is converted to the rfc-1807 format (Lasher and Cohen D. 1995), extended to support the bucket structure. The rfc-1807 metadata file becomes the second *element* in the metadata *package*.
- References to the actual paper -- that remains in the source archive -- are added as *elements* to the bucket. Typically these are references to either ps or pdf versions of the paper or both.

While the UPS Prototype is not optimal for the demonstration of some crucial advantages of buckets, it does provide some important results:

- As will be explained in the section on the addition of the SFX linking service, the bucket approach turns out to be attractive for the addition of value added services. Individual buckets, rather than a complete digital library service can be tailored to accommodate certain services. Such a concept is extremely flexible, since it makes support of certain services a matter of individual objects rather than of a complete collection. This guarantees that such support remains available when a bucket is relocated to or harvested by another digital library.
- The availability of two concurrent metadata formats in the buckets is seen as a modest but interesting illustration of their aggregative capability. But, more importantly, the flexibility derived from the capability to accommodate for this turns out to be valuable for the creation of the end-user services. While some engines may prefer to index the least elaborate of both formats supplied (rfc-1807), the SFX-linking service (see the section on the addition of the SFX linking service) highly values the availability of the extensive ReDIF format. This aggregative ability extends beyond metadata: it is easy to store multiple file formats and encodings. New formats can be generated as they become accepted and be stored along with the original source formats in the bucket.
- The UPS Prototype project is the first demonstration of a digital library with a significant amount of objects stored in a SODA architecture. While it clearly indicates that buckets can be deployed on a large scale, it also shows that further research is required to optimize the bucket footprint:
  - Each bucket generates about 100 kilobytes of overhead. For regular buckets that store full content, this overhead is negligible when compared with the bucket's data. It compares to the storage of 2 additional scanned pages of text (Nelson 1999). For lightweight buckets, this overhead is more noticeable since they only store metadata that is very limited in size. Fortunately, disk space is cheap and the storage overhead is not a significant problem.
  - The current bucket templates require approximately 60 inodes per bucket. Taking into account the size of the UPS collection, approximately 12 million inodes are required. Therefore, the original UPS disk partition that only had 2 million inodes, was significantly extended.

Buckets will always place additional storage requirements on a system, both in terms of kilobytes and inodes. While both demands can easily be met when dealing with collections the size of UPS, further research should result in buckets optimized for production that produce smaller footprints. Such research is already on its way.

ScreenCam 1 as well as Figure 1 to Figure 4 -- in the Project Results section -- demonstrate some of the methods available for buckets in the UPS environment. ScreenCam 3 and ScreenCam 4 shows buckets being approached via the NCSTRL+ user interface.

### ***Creation of NCSTRL+ end-user search facility***

It is decided to use the NCSTRL+ environment for the creation of the end-user search facility.

### **Indexing and Clustering buckets**

NCSTRL+ is an NCSTRL extension that supports buckets and clusters, both of which are important in the context of this experiment:

- Bucket support was added to Dienst by reducing the functionality of the User Interface service of Dienst. The Dienst Describe verb no longer builds a HTML interface, but rather redirects to the bucket and allows the bucket to build its own interface.
- Clusters provide a way to partition a dataset along predefined metadata axes. Each cluster divides the dataset into virtual sub-collections.

For the UPS dataset, the following clusters are defined:

- Archive: A division of the dataset according to the archive from which the bucket contents originates. This cluster is based on the *authority* part of the UPS identifier;
- Archive's Collection: A division of the dataset according to a sub-archive that might exist in the origin archive. This cluster is based on the *series* part of the UPS identifier;
- Author's affiliation: A division of the dataset according to the author's institution. This cluster is based on the Author-Workplace-Name tag of ReDIF;
- Subject: A division of the dataset according to the research subject. This cluster is based on the Ila Classification added for each record during the metadata conversion phase;
- Material Type: A division of the dataset according to the type of publication. This cluster is based on the normalized Manifestation type values accorded for each record during the metadata conversion phase;
- Terms and Conditions: A division of the dataset according to possible access restrictions, e.g. copyrighted, unrestricted, password-based, etc.

### Creation of the Search Interface

Owing to its Dienst heritage, the UPS interface provides a simple and advanced search tool. These interfaces are redesigned in order to make it more aesthetically pleasing and comprehensible than the original NCSTRL/NCSTRL+ interface. The simple search provides a keyword search across the entire bibliographic metadata set. The advanced search provides fielded searches for title, author and abstract. It also provides the capability to restrict searches to specific clusters. A special user-interface element needs to be introduced for the author-affiliation cluster in order to accommodate for the high amount of values that are available (see Table 4). Both search interfaces present the option to display search results by a chosen cluster, and within that cluster sort hits by author, title, date or relevance rank. NCSTRL+ also implements a "Recluster" Dienst verb that allows a search result list to be reorganized along different clusters without having to perform the search again. The list of brief search results is presented by the NCSTRL+ service. Clicking a result item to see the full record causes the corresponding bucket to self-display: the NCSTRL+ service is not involved in this. The original NCSTRL+ interface does not provide a facility to browse extensive indexes as commonly implemented in library catalogues or abstracting and indexing databases.. As a consequence, the UPS interface doesn't either. Still, it is important to note that the lack of appropriate subtagging of input data would have prevented to implement such functionality in a consistent manner. Screenshot 2 and Figure 5 to Figure 6 -- under the heading Project Results -- illustrate the described features of the user-interface.

Scaling problems with NCSTRL+ are encountered while building the UPS Prototype:

- A conflict between the Dienst architecture and the Solaris operating system occurs. Each Dienst publishing authority expects all of its publications to be in a single directory. Even though the prototype uses buckets, the buckets are internally stored and indexed in the same manner as regular Dienst objects. Thus, each bucket occupies 1 subdirectory in the Dienst publishing authority, and Dienst requires that all documents be at the same level within the publishing authority. The Solaris operating system has a limit of 32,767 subdirectories within a directory (Sun 1999). As such, for the large archives -- arXiv and RePEc -- it is not possible to have all the publications in a single publishing authority. Fortunately, both arXiv and RePEc have native sub-archives (*series* in the UPS namespace), none of which currently have more than 32,767 publications. Therefore, each sub-archive is made into a separate Dienst publishing authority. While this does not solve the Dienst/Solaris conflict, it presents a pragmatic workaround.
- Problems with the search engine occur. The native Dienst search engine does not scale beyond approximately 20,000 records. As a result of this problem, only approximately 20,000 records are

indexed for the UPS Prototype as demonstrated during the Santa Fe Meeting of the Open Archives initiative. The Dienst developers recommend the use of freeWAIS-sf (Pfeifer, Fuhr, and Huynh 1996) for larger collections. Unfortunately, freeWAIS-sf has its own scaling problem, since it automatically considers a word to be a stopword if it has more than 20,000 occurrences. With 193,000 records to be indexed, many non-stopwords occur more than 20,000 times (e.g. "galaxies", "dimensional", "temperature", "spin", etc.). More importantly, the archive names themselves become stopwords, making browsing through cluster selection impossible. In addition, the freeWAIS-sf only returns a partial listing of hits -- the ones that it determines to be most relevant. If 1,000 records match a search, freeWAIS-sf will return less than 200. This effectively negates the ability to do browsing on clusters and advanced searching. It is not straightforward to solve these problems in freeWAIS-sf. At the time of writing, the UPS Prototype contains all 200k objects indexed with the freeWAIS-sf engine. For reasons explained above, the functionality is limited. Work is on its way to replace the Dienst repository service with an Oracle-based storage system. This should address most of these problems.

- Another scaling problem occurs due to the fact that values for the author affiliation field are not normalized during the metadata conversion phase. As a result of this, the pop-down field used to restrict advanced searches to author affiliation initially contains about 23,000 lexically unique values (see Table 4), taking the advanced search screen 2 minutes to load. In order to address this problem, an interactive interface element to specify author affiliation is added and some basic normalization of author affiliation values is conducted.

### ***Addition of a SFX linking service***

The UPS Prototype end-user service is adapted to be interoperable with the context-sensitive and dynamic SFX linking solution. This is done in order to provide a concrete illustration regarding the possibility to integrate an e-print environment with other information resources of the scholarly communication mechanism. Such integration has already briefly been demonstrated in the "SFX@Gent & SFX@LANL" experiment (see Part 2), and is also the aim of the JISC/NSF OpCit linking project (Harnad 1999). The availability of a large e-print dataset stored in the unified UPS Prototype environment creates an interesting opportunity for further explorations of this problem domain. The design of the SFX framework used in the UPS Prototype is identical to the one described at length in Part 2.

### **SFX-awareness and buckets**

In the Elektron and Ghent&LANL SFX experiments, SFX-awareness was a matter of information systems. The SFX-button was being displayed for every record hosted by an SFX-aware information system when being displayed as a search result. However, in the UPS Prototype project it is not the NCSTRL+ digital library service that is SFX-aware, but rather the intelligent buckets in the newly created SODA archives. The essential intelligence that is required to be interoperable with SFX is now available within the buckets, not within the NCSTRL+ digital library service:

- The ability to understand that a user with access to a SFX service component is requiring the display of a search result;
- The ability to determine the location of the user's SFX service component;
- The ability to insert an SFX-button for a search result pointing at the user's SFX local redirection component (the insertion of an SFX-button is mentioned in the access log of a bucket, as shown in the yellow zone of Figure 4);

This can -- for instance -- be achieved by basing the creation of buckets on templates that support these features. SFX-awareness is implemented at the level of the individual objects in an information system, not at the level of the information system itself.

This turns out to be an important advantage in the context of this project:

- As will be explained hereafter, it is not always straightforward to imagine extended services that the SFX system can deliver when the link-sources originate from an e-print environment. For data originating from some of the archives involved, it is even impossible within the timeframe available to this project. In case SFX-awareness would be implemented at the level of the NCSTRL+ system, records originating from all archives would be displayed with an SFX-button. For data from

archives for which no extended services are defined, clicking the SFX-button would result in an empty SFX service screen. In order to avoid this, the NCSTRL+ system could be equipped with intelligence regarding the display or non-display of SFX-buttons for instance depending on the origin archive. The bucket approach pushes such intelligence down to the level of the individual objects in the archive, allowing each bucket to decide for itself whether or not to display an SFX-button. As a result of this, the SFX-button will even be displayed when a bucket is being approached directly instead of via a search engine.

- Some ReDIF-ed metadata has a Manifestation cluster, holding information on other manifestations of the work, such as a peer-reviewed paper, a book chapter etc. published after the e-print (see the section on the Metadata conversion). Buckets display such information under a separate "Published" tag. In this case, the bucket approach allows for the dynamic display of two separate SFX-buttons -- one for the e-print metadata and one for the Published information -- which will generate quite different services when clicked.

Once buckets have been made SFX-aware, the local redirection mechanism works in the same way as explained in detail in Part 2. The SFX-URL contains the unique UPS identifier of the bucket accorded during the Metadata Conversion phase (see Table 8). The protocol used to fetch the metadata directly accesses the bucket using the *metadata* method with the ReDIF qualifier (see Screencam 1 and Figure 3). Once the metadata has been fetched, the appropriate SourceParser -- that depends on the origin archive -- can be launched and a GenericRequest object can be created for delivery to the SFX service component. The service component must then dynamically decide which extended services to present.

#### Extended services in an e-print environment

Imagining and delivering extended services for metadata originating from e-print archives is not as straightforward as doing so for metadata originating from resources in the traditional scholarly communication mechanism. This has to do both with the nature of the e-print data and -- to a certain extent -- with the nature of the SFX service component:

- In the traditional communication mechanism, the existence of abstracts in A&I databases for a given link-source can be derived from data-elements in that link-source, basically the journal title and the publication year. This allows for a rules-based approach to present an abstract from a certain A&I database, for a given link-source. In the same manner, a rule can be defined to express the existence of full-text for a link-source from a given A&I database, using ISSN number, publication year, volume and issue number as parameters. In the e-print environment, there is little straightforward indication of the penetration of an e-print into other resources of scholarly communication that can be derived from the provenance of a link-source or from its metadata. As such it is more difficult to define such service-rules.
- The indication of the provenance of a record is essential for the SFX service to work properly. Knowing the origin of a link-source is often a synonym for knowing its broad research area. Knowing the broad research area is a synonym to being able to identify other resources dealing with that research domain. As such, information on the origin of a link-source is important in a rules-based approach to providing extended services for it. This reasoning does not apply to records originating from a multi-disciplinary resource since knowing their origin is not equal to knowing their research area. But as shown in the above, in the traditional scholarly environment the relationship of a link-source with other resources can also be derived from metadata information in individual records. This is not the case for metadata originating from multidisciplinary e-print archives, unless one would take into account the subject field of the metadata (if that would exist). Doing so would open up an important research problem regarding the interpretation of subject information and its representation in a rules-based approach as taken by the SFX service component.

#### The SFX services in the UPS Prototype project

Even given the above restrictions, the SFX linking system introduced for the UPS prototype presents numerous noteworthy service links that illustrate a possible way to integrate the e-print environment with other parts of the scholarly communication mechanism. Because of the above reasoning, SFX-buttons are only implemented in buckets that contain data originating from the discipline-oriented archives arXiv,

NCSTRL and RePEc and not CogPrints, NACA and NDLTD. For some buckets, two SFX-buttons are implemented, one for the basic e-print metadata and an additional one for citation metadata if the bucket has a Published tag.

#### *The SFX services for the e-print*

Table 8 gives an overview of the extended services that are available for e-print metadata in the UPS Prototype implementation of SFX. Consistent with the terminology introduced in Part 2, Source refers to the origin of the link-source, Target to the information resource into which a service link is provided. Service refers to the nature of the service that connects Sources and Targets:

- *author*: look-up of records in an abstracting & indexing database, with the same author as mentioned in the e-print metadata
- *reference*: look-up of the references made in the e-print

SOURCE	SERVICE	TARGET
arXiv	<i>author</i>	Inspec
arXiv:hep-th	<i>reference</i>	SLAC/SPIRES
arXiv:math	<i>author</i>	MathSci
NCSTRL	<i>author</i>	Inspec
RePEc	<i>author</i>	EconLit
RePEc	<i>author</i>	ABI/Inform

**Table 10: extended services for e-print metadata in the UPS Prototype**

Unfortunately, the problems regarding the -- lack of -- metadata quality, is an obstacle in the provision of the *author* service. Even if serious efforts are undertaken in the metadata conversion process to adequately identify individual authors in author fields containing multiple authors, the overall resulting quality is a hindrance for the *author* service to work seamlessly. In order to address this problem, author names in the SFX-screen are shown in editable fill-out boxes, enabling the user to manually correct names that are parsed inappropriately.

The *reference* service connecting arXiv:hep-th with the SLAC/SPIRES database deserves special attention. SLAC/SPIRES is a free citation database for high-energy physics. It contains almost all the references of both published papers and e-prints in that research domain. SLAC/SPIRES has a Web-based implementation that offers a link-to-syntax that can be used to request a list of all references of a publication, using metadata of the publication as parameters. The arXiv has been using this feature since quite a while as a means to enable users to see the list of references for e-prints in its high-energy physics sub-archive. As such, it is evident that this *reference* service is also part of the SFX services that are available for buckets originating from arXiv:hep-th. Moreover, in the course of the UPS Prototype project, the SLAC/SPIRES system has been made SFX-aware. As a result of this, all references returned by SLAC/SPIRES are equipped with an SFX-button, allowing the user to request extended services for each of them. Since such references are to both published and e-print material, an attractive integration of the e-print environment and the traditional scholarly communication environment is achieved, using the references as an intermediate stage.

#### *The SFX services for the published version*

Since the information in Published tag refers to material originating in the traditional scholarly communication environment (journal articles, conference proceedings, books), the whole spectrum of extended services as listed in Table 6 of Part 2 is potentially available. Unfortunately, there is a remarkable lack of consistency in the syntax of the citations available in this tag. Since no attempt is made to normalize this data during the metadata-conversion phase, this cumbersome task is left to the SFX

SourceParsers. In those cases where parsing is successful, again, an attractive integration of the e-print environment and the traditional scholarly communication environment results.

ScreenCam 3, ScreenCam 4 and Figure 7 to Figure 13 -- in the Project Results section -- illustrate the type of extended services that are made available for the buckets in the UPS environment.

## **Illustration of project results**

As a result of the UPS Prototype project, an experimental cross-archive end-user service is presented at the Santa Fe Meeting of the Open Archives initiative. By the time of writing, it is available at <http://ups.cs.odu.edu> and will remain online for an uncertain period of time. For archival purposes, the concrete results of the project are illustrated here by Lotus ScreenCam movies that show how a user navigates the multidisciplinary UPS Prototype environment.

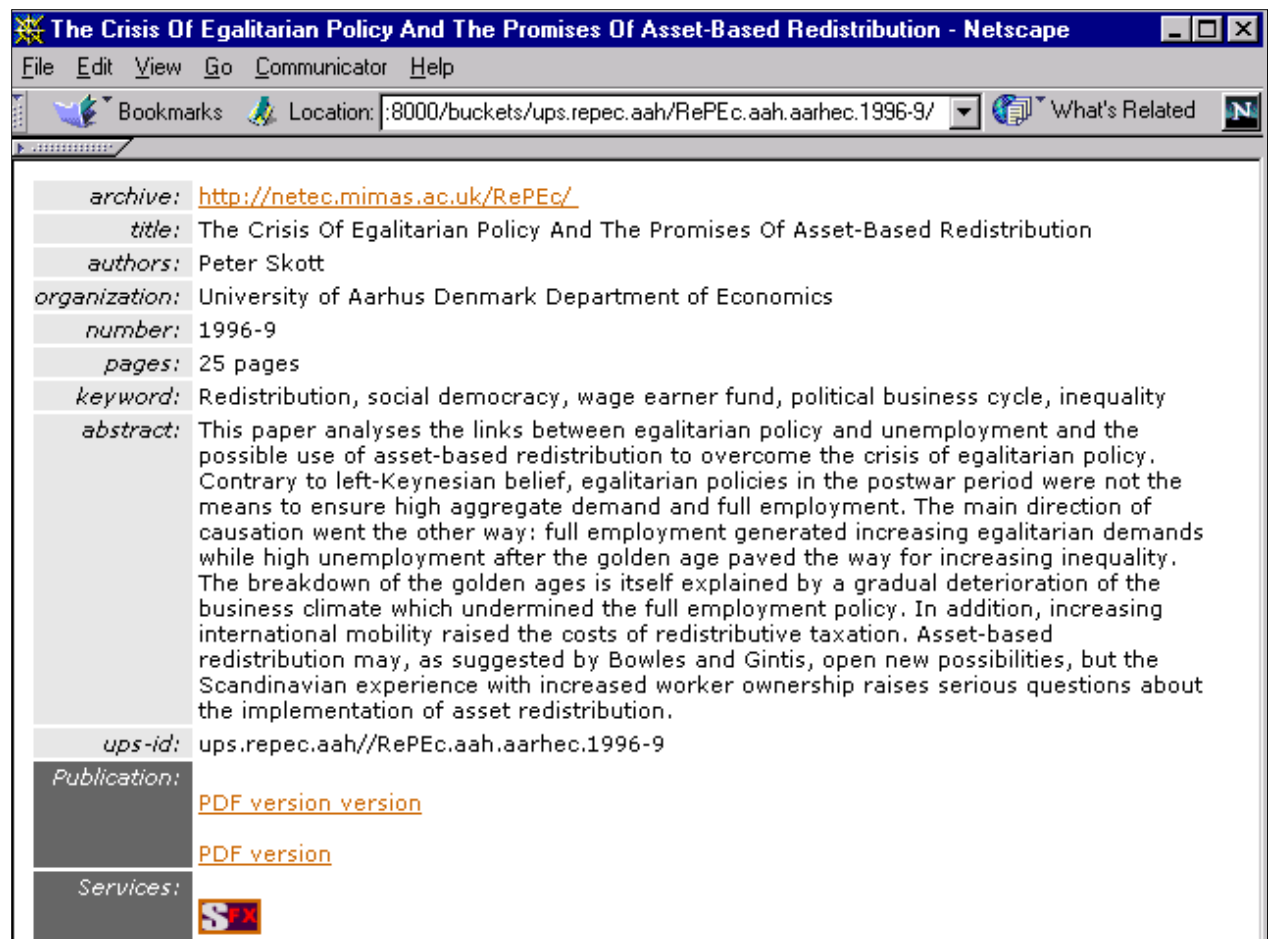
The ScreenCams are provided as stand-alone executables that can only be run on WinTel computers. Since the ScreenCams are large files, their size is mentioned. The ScreenCams do not contain audio. In addition to these ScreenCams, some examples are also given by means of screendumps.

### ***Bucket methods***

**ScreenCam 1: Illustration of bucket methods** (Lotus ScreenCam executable for WinTel computer; no audio; size 6 Mb)

A user is directly accessing a bucket from a web-browser, typing in supported bucket-methods as URLs. The user successively displays the content of the bucket, its rfc-1807 metadata element, its ReDIF metadata element and the access-log of the bucket. The bucket in the example is the one with UPS identifier RePEc:aah:arhec:1996-9. The server on which the SODA archives are stored is [ups.cs.odu.edu](http://ups.cs.odu.edu). Figure 1 to Figure 4 show screendumps illustrating the same bucket-methods. As long as the UPS Prototype archive remains on-line, clicking the URLs in the captions of those figures will yield the results shown in the figures and the screenCam.

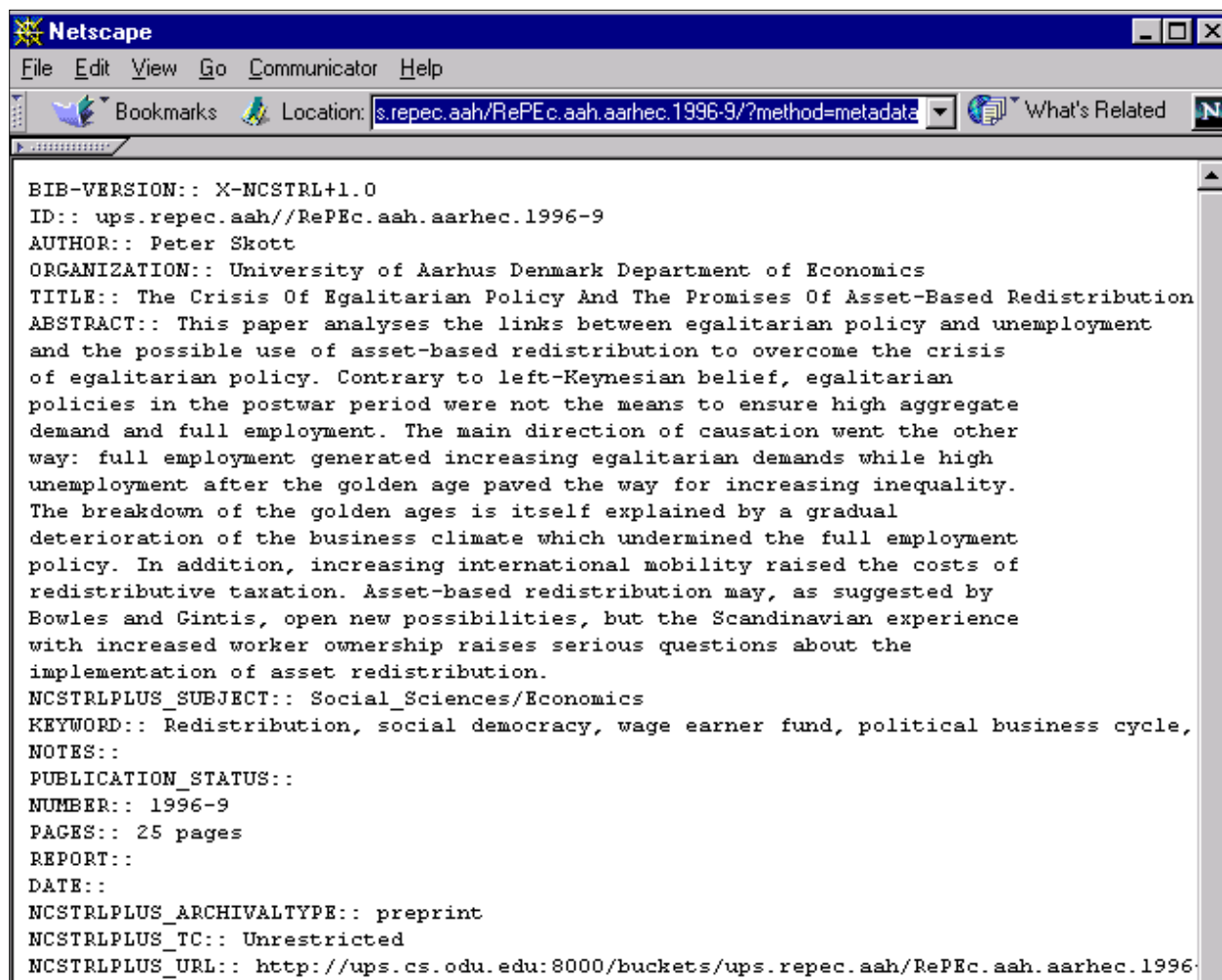
### Screen dump example 1:



**Figure 1: The bucket *display* method**

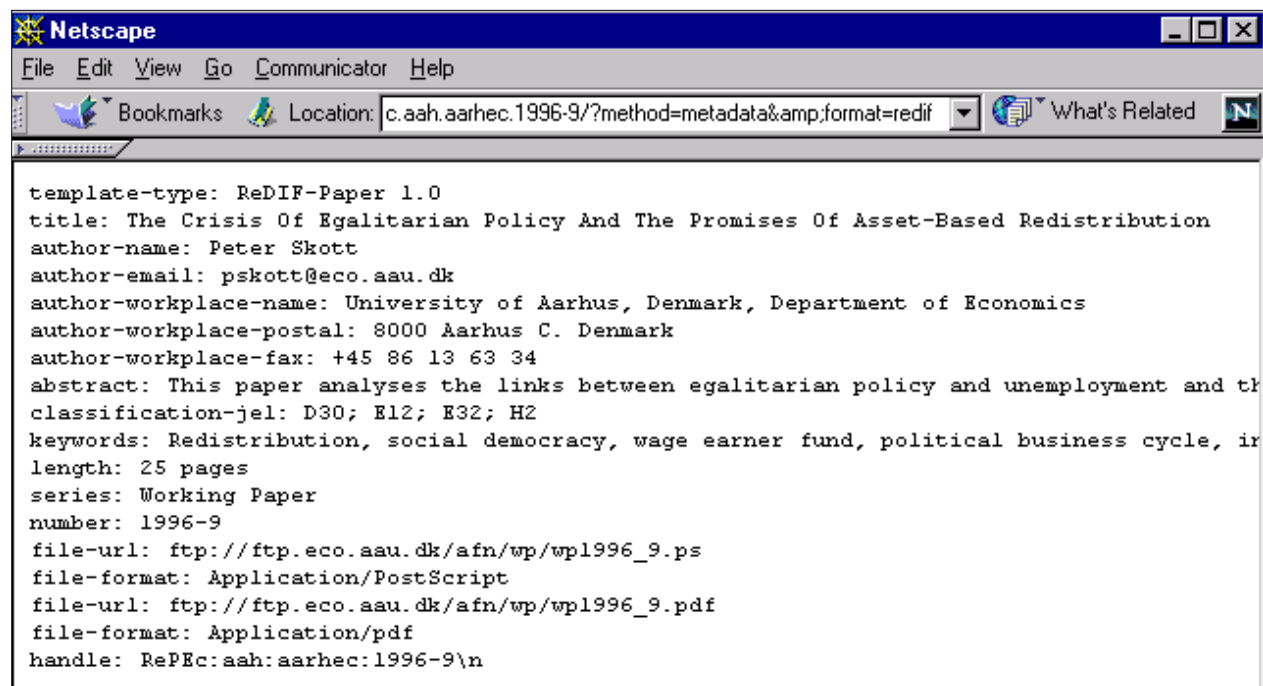
Invocation of the default *display* method via the command  
`http://ups.cs.odu.edu:8000/buckets/ups.repec.aah/RePEc.aah.aarhec.1996-9/`  
results in the auto-display of the bucket.





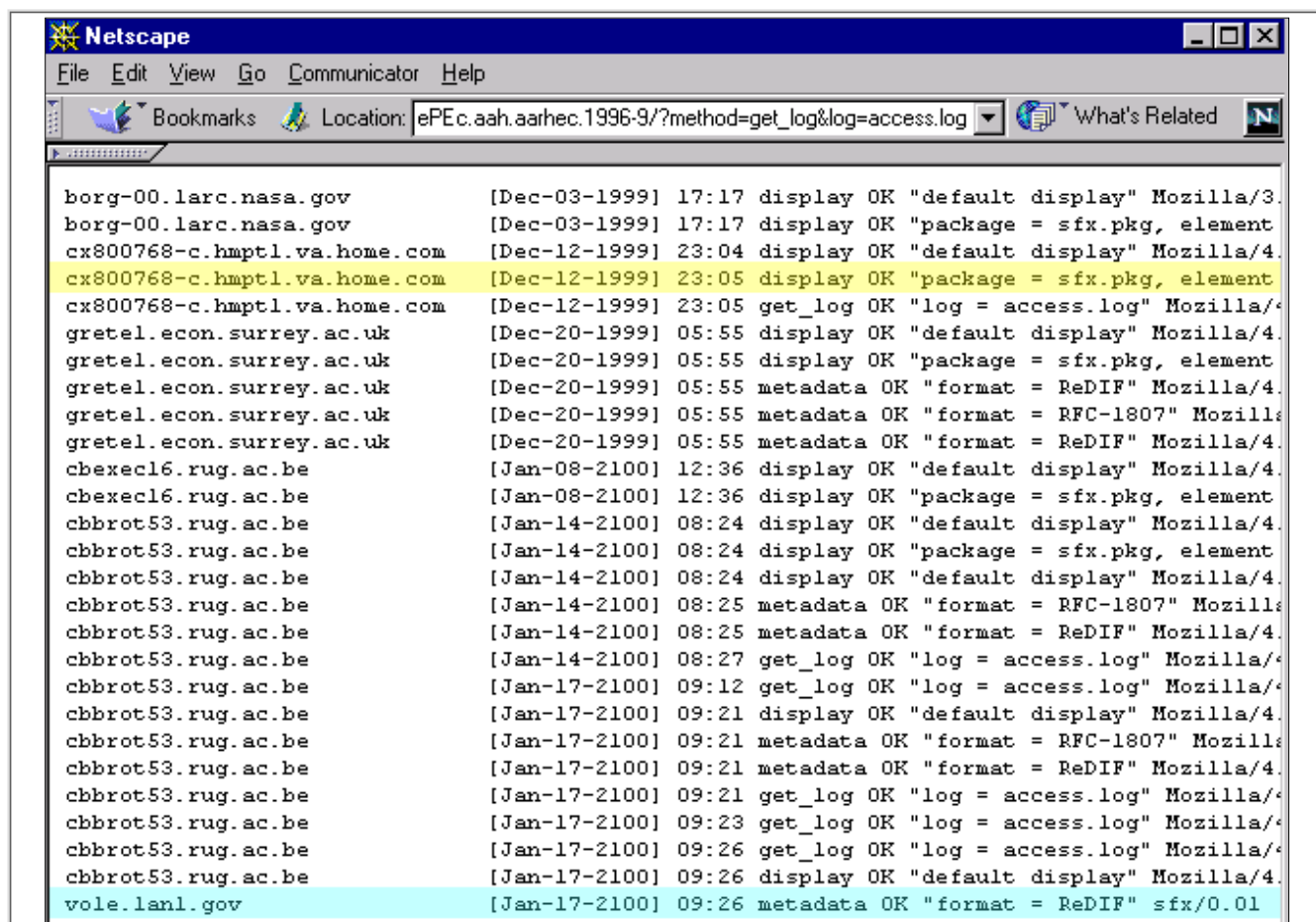
**Figure 2: The bucket *metadata* method shows rfc-1807 data by default**

Invocation of the *metadata* method via the command  
<http://ups.cs.odu.edu:8000/buckets/ups.repec.aah/RePEc.aah.aarhec.1996-9/?method=metadata>  
 results in the display of the rfc-1807 element of the bucket's metadata package.



**Figure 3: The bucket *metadata* method used with the ReDIF qualifier**

Invocation of the *metadata* method with a ReDIF qualifier via the command `http://ups.cs.odu.edu:8000/buckets/ups.repec.aah/RePEc.aah.aarhec.1996-9/?method=metadata&format=redif` results in the display of the ReDIF element of the bucket's metadata package.



**Figure 4: The bucket log method**

Invocation of the log method via the command

`http://ups.cs.odu.edu:8000/buckets/ups.repec.aah/`

`RePEc.aah.aarhec.1996-9/?method=get_log&log=access.log`

results in the display of the logs showing previous accesses to the bucket.

### **The NCCTRL+ search interface**

**ScreenCam 2 : The NCCTRL+ search interface** (Lotus ScreenCam executable for WinTel computer; no audio; size 10 Mb)

The main interface features of the NCCTRL+ search facility, created for the UPS Prototype are illustrated. The screenCam starts by showing the simple search screen that has a single field to search the collection. It also has a Display option, allowing search results to be grouped by a chosen cluster and within the chosen cluster by author, title, date or relevance rank. Next, the features of the advanced search are illustrated. The top -- Search -- part of the advanced interface is the area for fielded searches. Author, Title and Abstract fields can be combined by Boolean and/or. The middle -- Filter -- part of the interface allows the user to limit a search to certain clusters. The Archive Collection and Subject Division fields are not implemented, but included to illustrate functionality that could be created in a more flexible project time span. The special interface element created to deal with the high amount of author affiliations is also illustrated. The bottom part of the screen has the same features as the Display part of the simple search interface. After having performed a search, search results are reorganized along another cluster than the one chosen at the time of performing the search. Figure 5 and Figure 6 illustrate the described features of the advanced interface.

## Screen dump example 2:

UPS - Netscape

File Edit View Go Communicator Help

Bookmarks Location: <http://ups.cs.odu.edu:8000/> What's Related

# UPS PROTO

Simple search Advanced Search Help

### Search specific bibliographic fields

Author

Title

Abstract

Combine fields with ☒ AND ☐ OR **search**

### Filter options

Archive

Archive's Collections

Author's Affiliation:  
*please type part of it:*   
*then select one:*

Subject:  
*Subject Classes*

*Subject Divisions*

Material Type

Terms and Conditions

### Display options

Group results by

Sort results by

**search**

Figure 5: the advanced interface

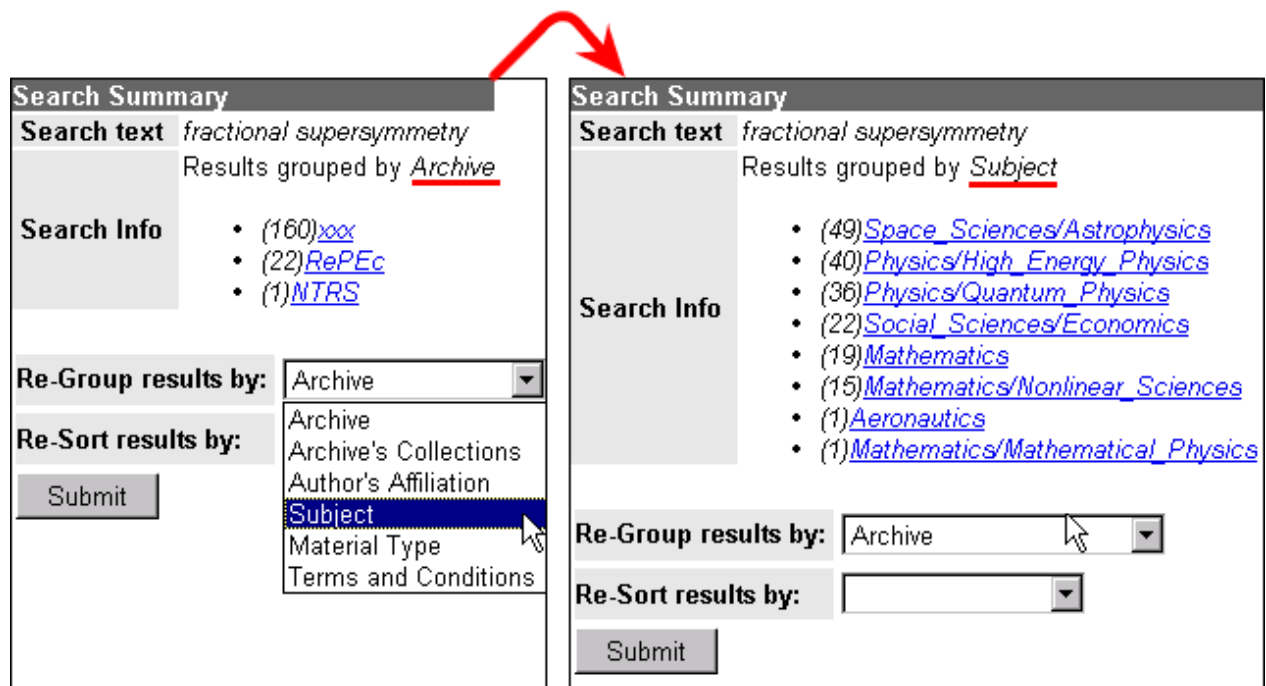


Figure 6: reorganizing search results along another cluster

### Buckets and the SFX linking service

**Screencam 3 : Buckets and the SFX linking service** (Lotus Screencam executable for WinTel computer; no audio; size 50 Mb). The screencam is recorded with the Dienst indexing engine in place. Also, the SFX server matches the collection of the University of Ghent.

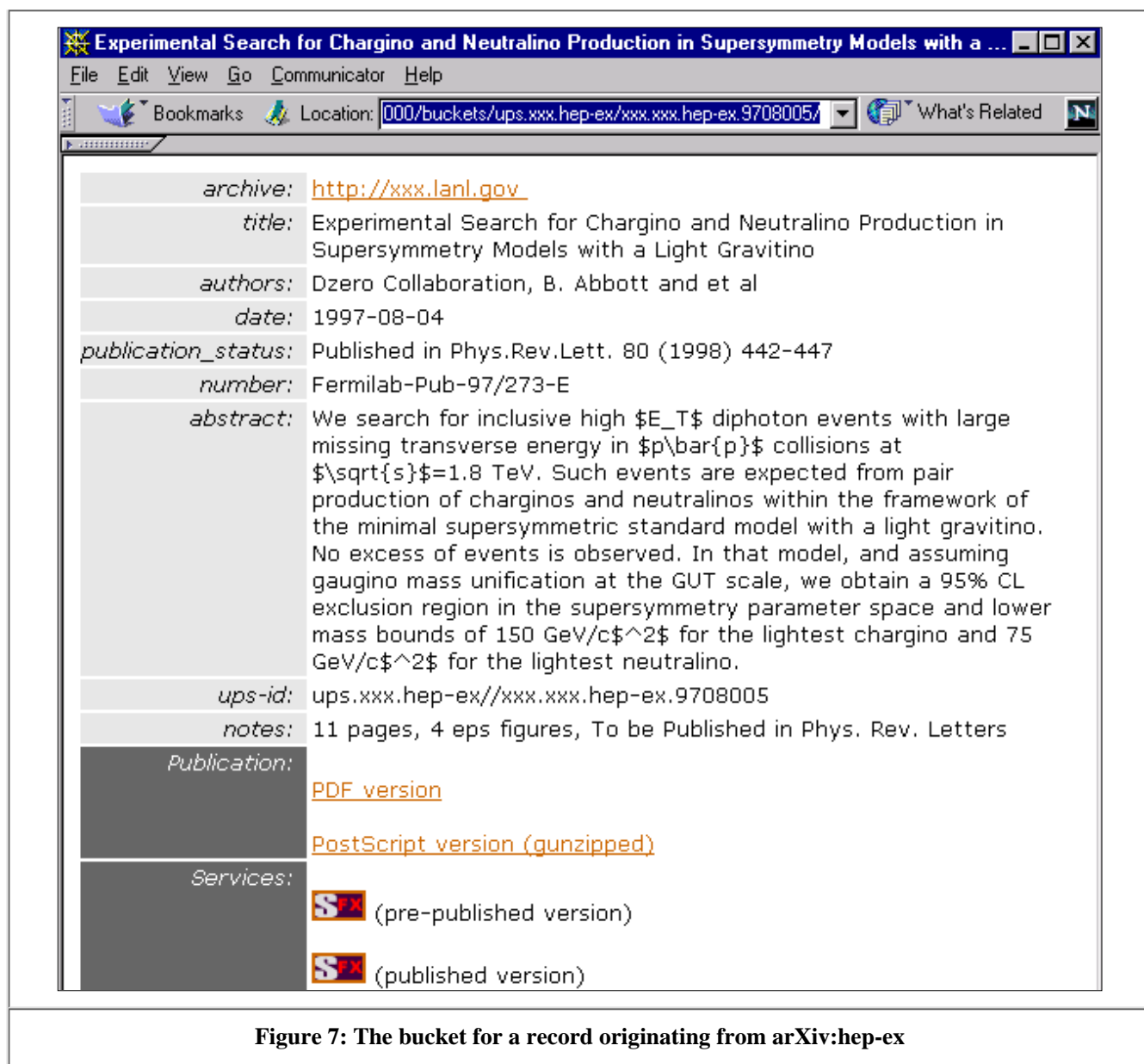
A user performs a search on the term "complexity". Search results are grouped along the Archive cluster, but the user decides to regroup them by Subject. Next, the user scrolls through the brief search results displayed by the NCSTRL+ service. Regularly, he clicks a specific search result. This results in the bucket corresponding with the search result to self-display. Since the buckets are lightweight, the full-content links followed by the user point into the original archives, not the UPS environment. Special attention is accorded to the features of the SFX-linking service. The Screencam shows how buckets containing data originating from RePEc, ArXiv and NCSTRL self-display SFX buttons. The user clicks some of these buttons in order to request extended services. For RePEc buckets, service links are provided that enable the user to look-up authors of the e-print in the EconLit and ABI/Inform databases. For ArXiv buckets this *author* service points at the Inspec database. When the user requires SFX-services for data originating from the ArXiv:hep-th sub-archive, the *reference* service appears. This service points into SLAC/SPIRES that contains the references of publications in high energy physics, and as such also of e-prints submitted to ArXiv:hep-th. Clicking the *reference* link brings the user to the SLAC/SPIRES environment where all references of the e-print described by the current bucket are displayed. The SLAC/SPIRES environment is SFX-aware and therefore SFX-buttons are available for every reference. Some references are to e-prints, others to formal publications. The user clicks several of these SFX-buttons. For formal publications, service links similar to the ones demonstrated in the "SFX@Gent & SFX@LANL" experiment show up. For e-prints, a link to the full-content in the UPS environment shows up. Clicking it causes the corresponding bucket to self-display. The Screencam also shows how buckets can display two SFX-buttons: one for the e-print and one for the published version. When a user clicks the published SFX-button, the *full\_text* service brings him to the article that was formally published after the e-print. At the end of the Screencam, some bucket methods are illustrated again. Amongst others, the access log of a bucket from which SFX services have been requested is being displayed. The log shows that the bucket has been approached several times by a web browser (Mozilla) but also that the SFX service has fetched the bucket metadata (see also blue zone of Figure 4).

**Screencam 4 : Buckets and the SFX linking service** (Lotus Screencam executable for WinTel computer; no audio; size 56 Mb) . The screencam is recorded with the freeWAIS-sf indexing engine in place. The SFX server does not particularly match the collection of any institution. Rather, it illustrates the type of services that can be provided as a means to integrate the e-print environment with other parts of the scholarly communication mechanism.

The user performs a search on "fractional supersymmetry". The search results are sorted by archive, and the user reorganizes them according to subject. He clicks a search result, causing the corresponding bucket to self-display. The bucket displays two SFX-buttons, since it describes an e-print that has been published formally. The e-print SFX-button provides an *author* link into the Inspec database. The SFX-button for the published version provides a wide range of services. From those, the user chooses to follow a *holdings* look-up leading him into the catalogue of the Library of Congress. He also chooses a recommendation service that will advise him on journals related to the one in which the formal publication has occurred. This service is based on the SpreadIt system (Bollen, Van de Sompel and Rocha in preparation) that generates recommendations by applying spreading activation to a previously established set of associative relations among electronic journals. These journal relationships have been generated from user interaction patterns by the @ApWeb methodology (Bollen and Heyligen 1998) applied to logs of usage of electronic journals. Next, the user moves to another search result for which the bucket -- again -- displays two SFX-buttons. The user clicks the SFX-button for the published version. From the SFX-menu, he selects a link to PubList, that displays journal information. He also selects the recommendation link again. At this time, the service presents a different list of related journals, reflecting the fact that -- by now -- he has "traversed" two journals in his actual session. The user moves on to a search result originating from arXiv:hep-th. From the bucket, he requests SFX-services for the e-print version. Since the actual record is about high-energy physics, the *reference* service pointing at SLAC/SPIRES shows up. The user clicks this link and the citations made in the e-print are being displayed by the SLAC/SPIRES system. Each citation has an SFX-button. Clicking the buttons generates the typical extended services. For the second citation selected by the user, an e-print version exists. Therefore, the SFX-menu screen now also shows a link that leads into the UPS Prototype system. Clicking this link causes the UPS bucket corresponding to the cited e-print to be displayed. Next, the user moves on to a search result originating from the RePEc initiative. Clicking the SFX-button generates *author* services pointing into EconLit and ABI/Inform.

### **Screendump example 3:**

The user accesses a bucket originating from the ArXiv:hep-ex sub-archive (Figure 7). The bucket contains SFX buttons for both the e-print version and for a published version. Clicking the latter, results in a wide range of services to become available (Figure 8), variations of which have been demonstrated in earlier SFX experiments. The SFX-button for the e-print version generates two services: an *author* link pointing into the Inspec database and a *reference* link pointing to the SLAC/SPIRES citation database for high energy physics (Figure 9). The user follows the *reference* service causing citations made in the e-print to be displayed by the SLAC/SPIRES system. Each citation carries an SFX-button (Figure 10) and the user decides to request extended services for one of those. Again, a range of services become available (Figure 11). The user follows the recommendation service that provides him with information on journals related to the ones he already has "traversed" during his current session. Their titles are shown at the top of the recommendation screen; the recommendations themselves are listed in the remainder of the screen (Figure 12). Each of the recommended journals also carries an SFX-button. The user moves back to the SFX-menu of Figure 11 and from there follows a service linking him to the e-print version of the cited paper. Since that e-print is also part of the UPS collection, this service points back into the UPS environment, causing the appropriate bucket to self-display (Figure 13).





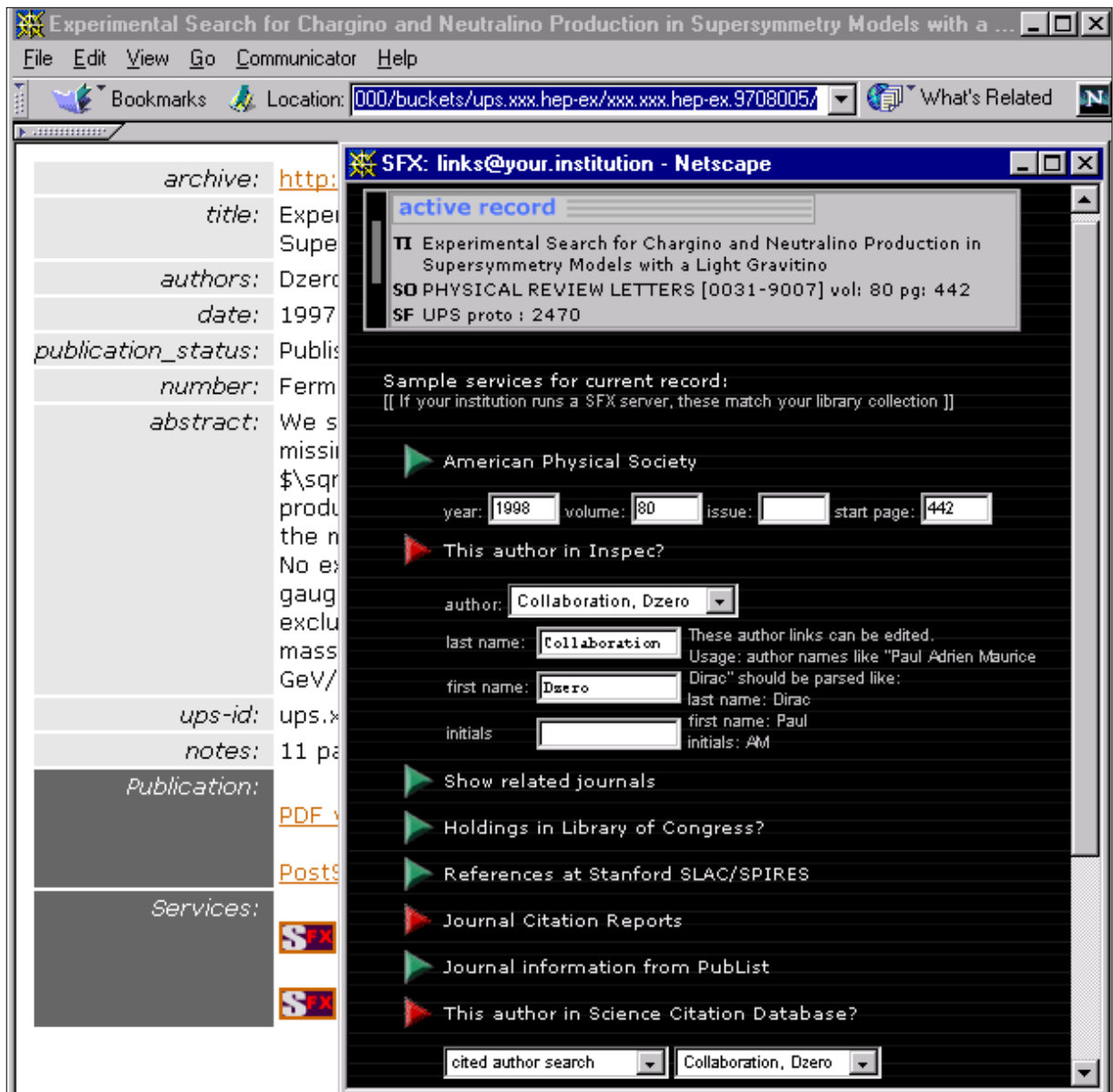


Figure 8: SFX-services for the published version of the paper described by the bucket in Figure 7

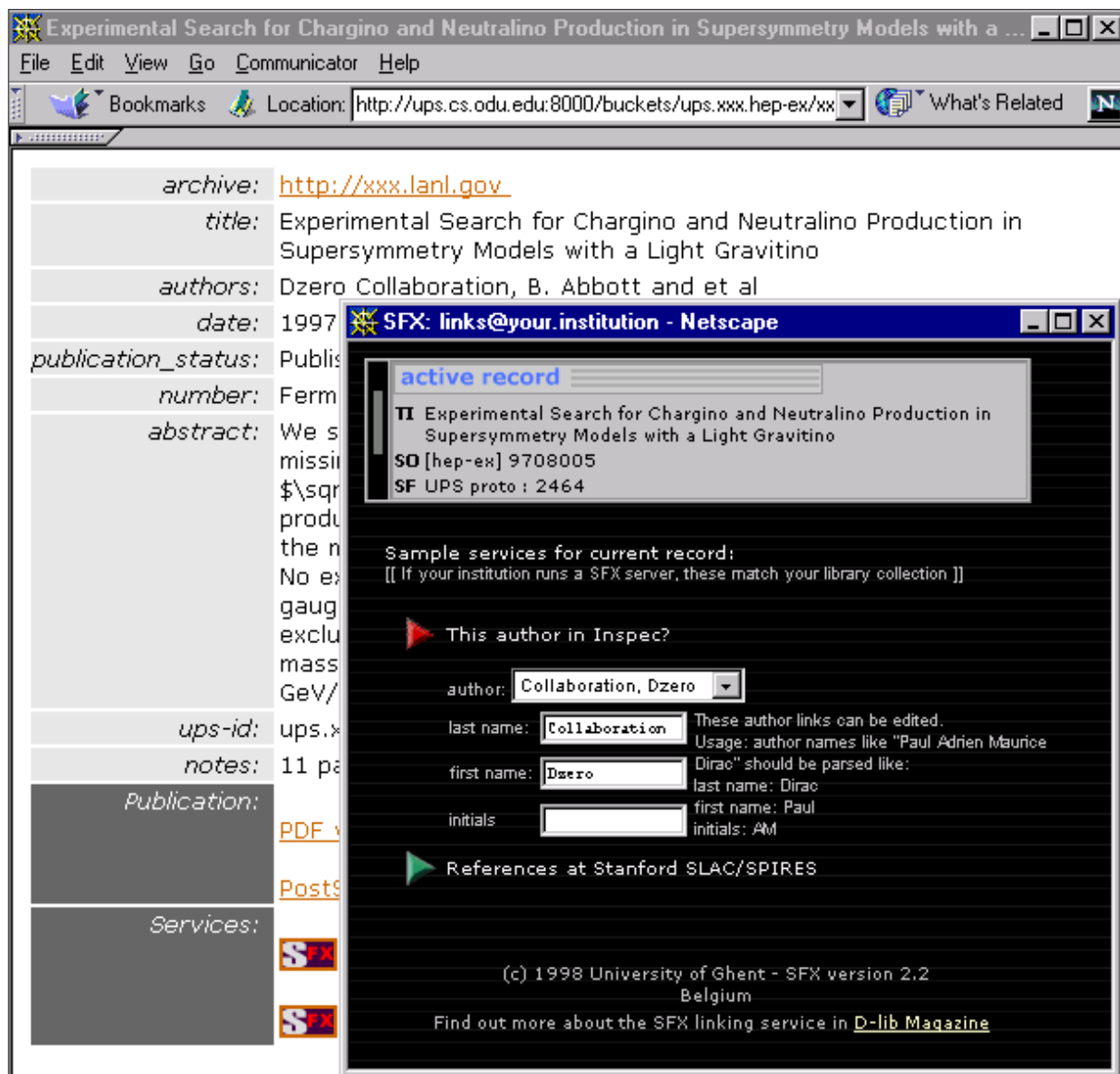
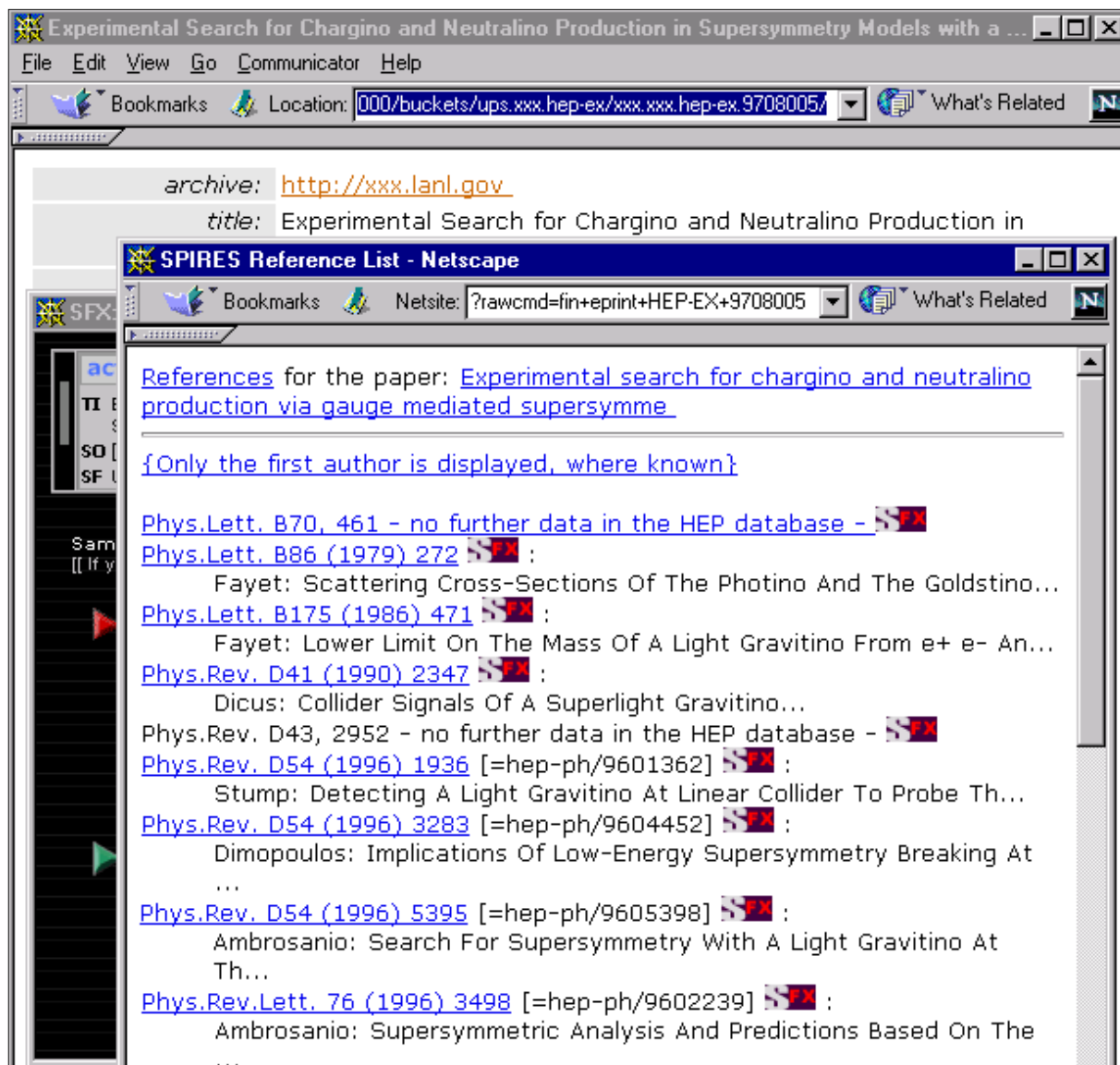


Figure 9: SFX-services for the e-print version of the paper described by the bucket in Figure 7



**Figure 10: The user follows the *reference* link from the SFX-menu screen of Figure 9**

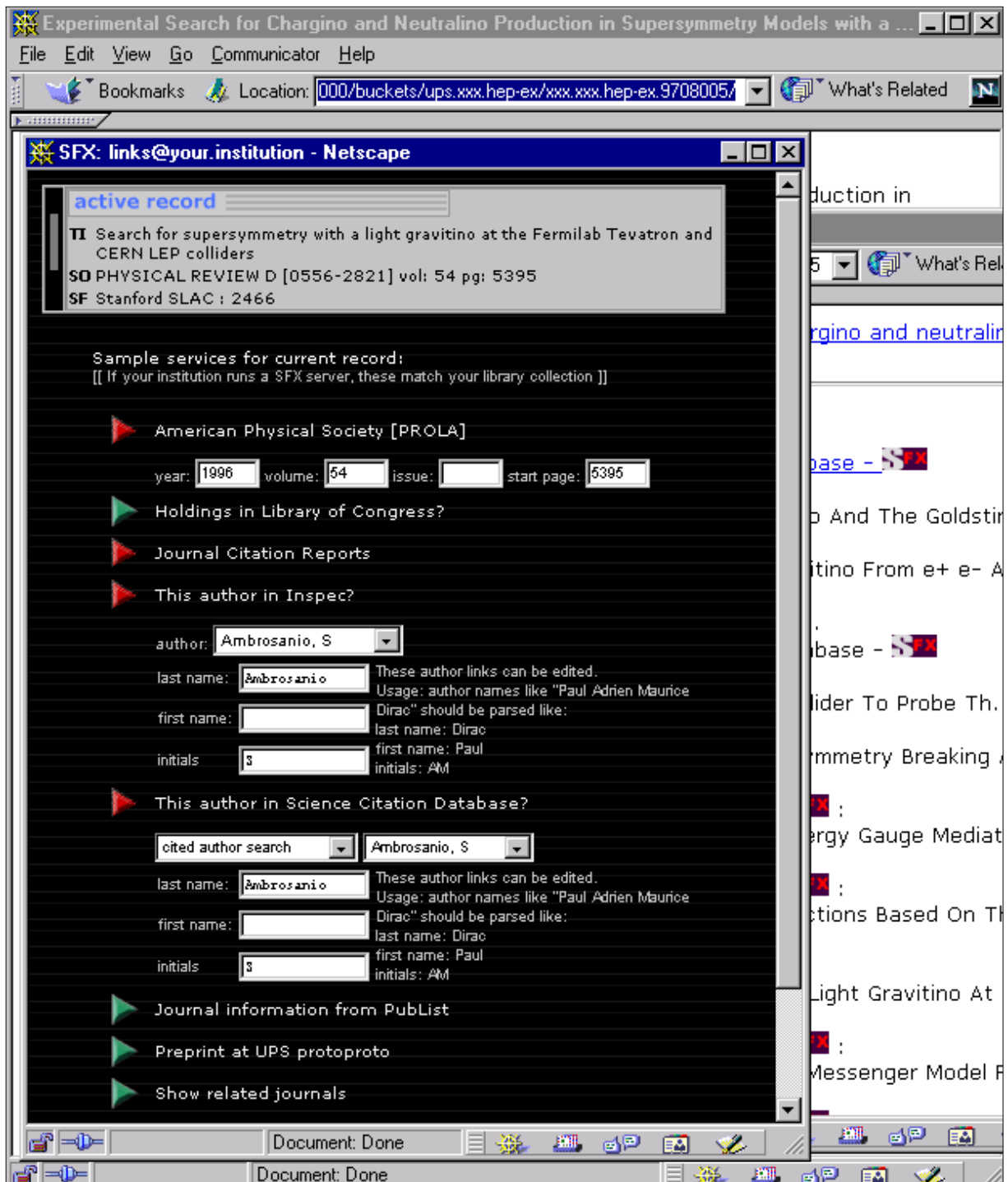
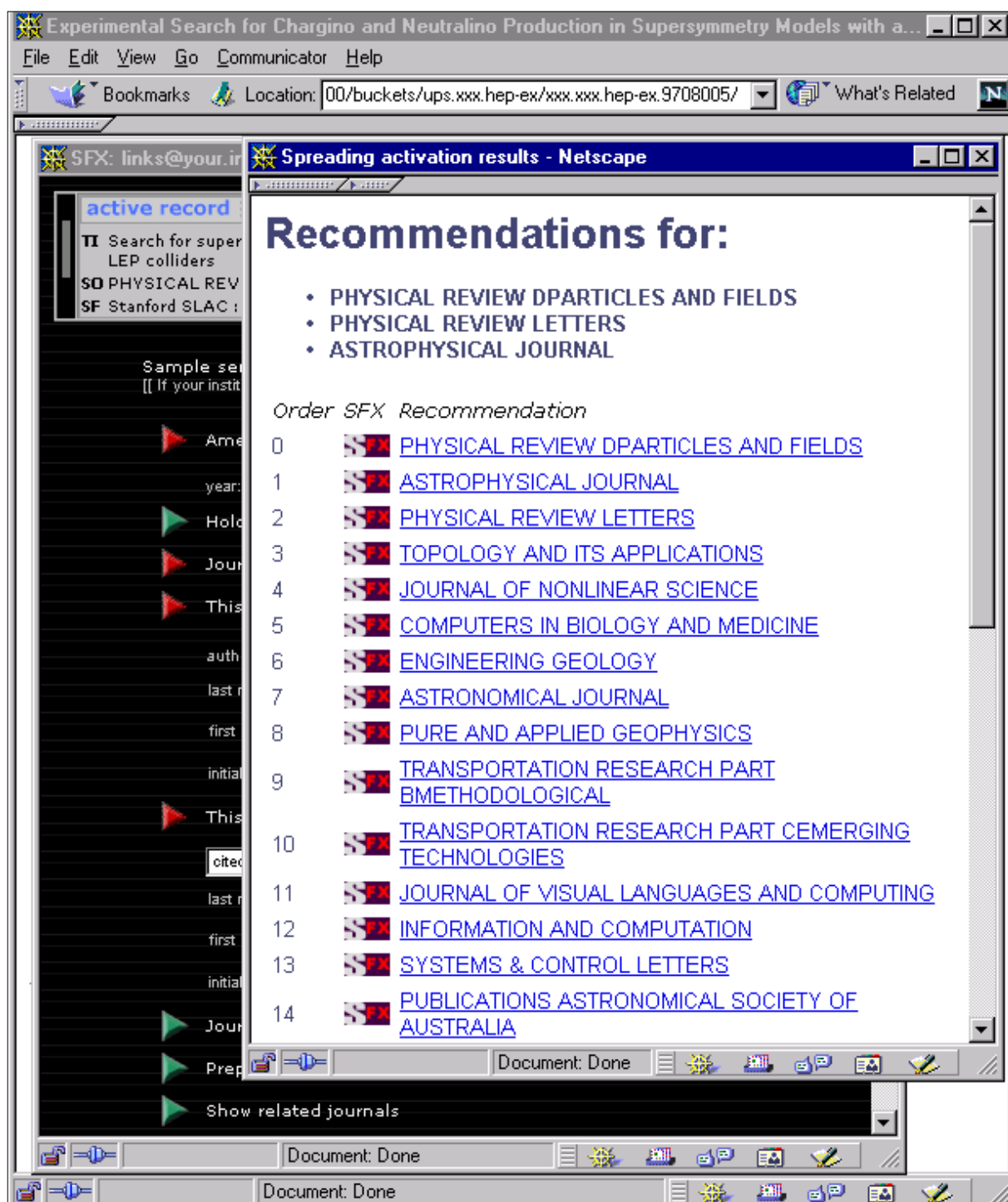


Figure 11: The user requests extended services for one of the citations from Figure 10



**Figure 12: The user checks out the journal recommendations made by the SpreadIt! system by clicking the "Show related journals" menu item**

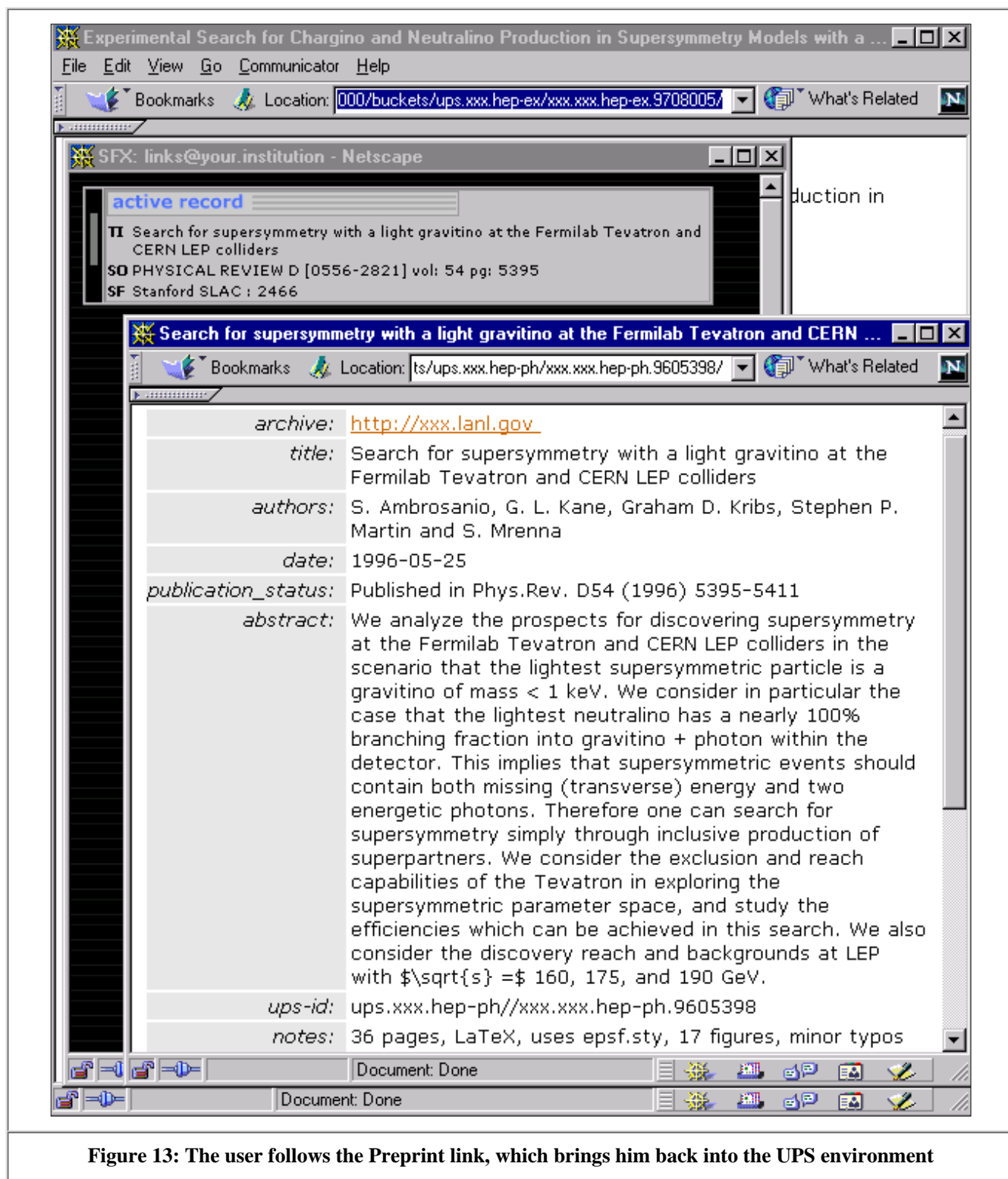


Figure 13: The user follows the Preprint link, which brings him back into the UPS environment

## Intermediate conclusions

In a four-month timeframe, the project team has demonstrated the feasibility to create a cross-archive end-user service by means of its UPS prototype system. The team has identified a number of issues that are crucial in making the creation of such services more straightforward and that aim at the creation of rich, diverse and high quality services. These issues are out of the scope of this thesis. They have been translated into recommendations made to the Open Archives group, at the occasion of their first meeting in Santa Fe (Van de Sompel, Krichel, Nelson, et al 2000).

The UPS Prototype Project also resulted in interesting new insights regarding the digital library

technologies that have been used in the course of the project. The results regarding the application of the SFX framework are within the scope of this thesis. The framework was introduced in an environment that was fundamentally different from the ones of the Elektron and "SFX@Ghent & SFX@LANL" experiments.

The technical environment was different in that the core collection was made up of digital objects handled in a Smart Objects Dumb Archive architecture. The individual digital objects were made SFX-aware, rather than -- as in previous experiments -- the information system used to gain access to records in resources. The fact that the SFX local redirection mechanism could easily and successfully be implemented in such a different architectural set-up is considered to be a strong indication of its feasibility.

The environment was also different because of the nature of the core collection, that described e-prints originating from various archives rather than material that has formally been published, as was the case in the previous experiments. This has consequences that are significant for the SFX service component.

E-print metadata rarely has a quality level that is comparable with that of records in other scholarly information resources such as abstracting and indexing databases. This is related to the author-self-submission procedure used by many e-print archives. Since -- for the delivery of many extended services -- the SFX service component depends on other metadata than an identifier, its decreased performance could have been foretold. Still, for those services that did not require additional metadata and in those cases where the metadata quality was decent, the SFX service component confirmed the flexibility and attractiveness that has been demonstrated in the Elektron and "SFX@Ghent & SFX@LANL" experiments.

There does not exist a global namespace for e-prints. Since conclusive knowledge about the origin of the e-print link-source is essential for the SFX framework, this could prevent the framework from functioning in the e-print environment. In the UPS Prototype this problem has been approached by according a unique UPS-identifier to each bucket, revealing the origin archive of its metadata. In order to function in relation to individual e-print archives, these should have unique archive identifiers and unique record identifiers within those archives. Actually, the usage of such identifiers is one of the important recommendations of the Santa Fe Convention (Van de Sompel and Lagoze 2000 ; see Appendix). As such, the SFX framework will potentially be able to interoperate with e-print archives that comply with the convention.

The project revealed an intriguing future research area regarding the SFX service component. That service component is based on conceptual service relationships between Sources and Targets, whereby Thresholds help to decide on the relevance of the services for a given link-source. To a certain extent, the notion of Sources and Targets is correlated with the notion of Subject, and as such it may be interesting to add the latter to the design.

It can be concluded that the UPS Prototype has demonstrated the feasibility to apply the open and dynamic SFX linking framework in an environment that is in many senses different than the ones of earlier experiments. The SFX redirection component has successfully been applied in relation to UPS buckets and citations in the SLAC/SPIRES database. The SFX service component confirmed its potential, even if it was hindered by the lack of quality of the e-print metadata. Overall, the project has demonstrated the feasibility to integrate the e-print environment with the established scholarly information resources by means of the SFX framework. The most appealing integration resulted from the SFX-awareness of SLAC/SPIRES that holds the references of e-prints in the high energy physics subarchive of arXiv. This made it possible to navigate from the e-print metadata to the citations of the e-print and from those citations either into the traditional scholarly resources or back into the e-print environment. The exploration of such capabilities is the topic of the OpCit linking project (Harnad 1999).



# Conclusions

In the Problem Statement, a categorization of linking frameworks has been introduced by means of the description of static and dynamic linking as well as of open -- or context-sensitive -- and closed linking. The aim of the thesis was to demonstrate the feasibility of interlinking distributed electronic scholarly information resources in a dynamic and context-sensitive manner. This has been achieved by:

- The incremental development of concepts that crystalized to become the SFX linking framework;
- The illustration of the successful application of the SFX framework in various dissimilar digital library environments;
- The demonstration of the generic nature and the modularity of the SFX framework.

The above has been accomplished over the course of three experiments:

- The Elektron experiment (August 1998 - November 1998) in which some initial conceptual foundations of the SFX linking framework were laid. The experiment has been conducted within the boundaries of a subset of the digital library collection of the University of Ghent (Belgium). There was active cooperation from two parties in the information industry.
- The "SFX@Ghent & SFX@LANL" experiment (February 1999 - June 1999) in which the initial concepts of the SFX linking framework were generalized. "SFX@Ghent & SFX@LANL" has been conducted in the complex and dissimilar digital library environments of the Research Library of the Los Alamos National Laboratory (New Mexico, US) and of the University of Ghent (Belgium). There was extensive cooperation from several parties in the information industry.
- The UPS Prototype project (July 1999 - October 1999) in which the generalized concepts of the SFX linking framework were applied in an environment that was in many ways different from the ones of the earlier experiments. There was extensive collaboration with the e-print community and with digital library researchers.

SOURCE					
A & I databases	E G U	E G U	- G U	E G U	not applicable
OPAC	E G U	E G U	- G U	E G U	not applicable
citation databases	- G U	- G U	- - U	- G U	- - U
full-text journals	- G -	- G -	- G -	- G -	- G -
e-print archives	- G U	- - U	- - U	- - U	- - U
	A & I databases	OPAC	citation databases	full-text journals	e-print archives
TARGET					

**Table 1: Resources interlinked over the course of three experiments (E stands for Elektron; G for Ghent&LANL; U for UPS Prototype)**

Table 1 shows the type of electronic information resources that have been dynamically interlinked in each of the experiments. These information resources:

- Covered the whole spectrum of resources commonly used for dissemination and discovery of

scholarly information;

- Ran on a wide variety of technologies;
- Ran on systems adhering to different architectural models;
- Were both local and remote with respect to the institutional environment of the user.

The remainder of this concluding section will describe the most important characteristics of the framework that has enabled this interlinking.

## Systems supportive of selective resolution or context-sensitive linking systems

The most general result derived from the experiments is the identification of the basic components of context-sensitive linking systems -- systems supportive of selective resolution -- as well as a characterization of possible architectural models of such systems. This result has built upon the categorization brought forward in (Caplan & Arms 1999) in relation to the Harvard problem. This problem, also referred to as the 'appropriate copy problem', deals with the delivery of the *appropriate full-text* instance that corresponds with a given citation, whereby *appropriate* relates to context-sensitiveness. Since the delivery of the appropriate full-text instance is a special case of the broader problem of the delivery of *relevant extended services* -- the core problem addressed in the SFX research -- an attempt to generalize the characterization of (Caplan & Arms 1999) was indeed tempting.

Figure 1 shows the basic components of a context-sensitive linking solution:

- The **redirection component** transfers metadata of the link-source -- for which extended services are requested -- from the information resource to which the link-source belongs, to the service component. The redirection mechanism addresses the 'grabbing the link-source' problem that has been referred to in the Problem Statement.
- The **service component** takes metadata from whichever information resource in the digital library collection as an input, delivering extended services as an output.

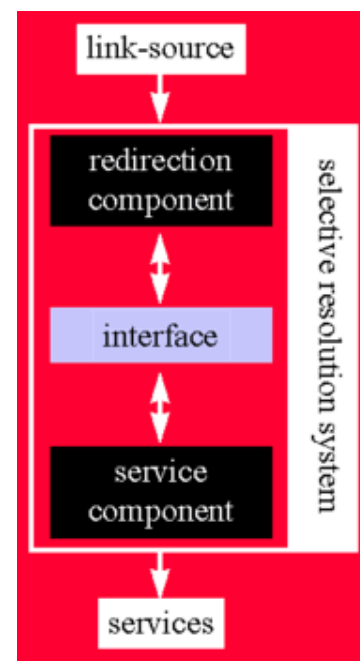


Figure 1: Systems supportive of selective resolution

Moreover, when taking into account the fact that:

- both the redirection component and the service component can be local or remote;
- the redirection order is subject to variation:
  - Redirection of the link-source metadata to the local service component first, using a central service component as a means to complete the set of services that can be presented;
  - Redirection of the link-source metadata to the central service component, whose default services can be overwritten and/or completed after communication with the local service component;

a categorization of systems supportive of selective resolution based on the nature of the service component and the redirection order follows (Table 2).

CATEGORY			
Category 1		central	central
Category 2	a	central & local	local => central
	b	central & local	central => local
Category 3		local	local
		<b>SERVICE COMPONENT</b>	<b>REDIRECTION ORDER</b>

**Table 2: categorization of systems supportive of selective resolution**

In Table 2:

- Category 1 only has a central service component and hence a central redirection mechanism.
- Category 2 has both a central and a local service component that contribute to the presentation of the services. Also, there is some form of communication between both. For this Category, it is possible to imagine both approaches regarding the redirection order mentioned above.
- Category 3 builds purely on a local service component and hence requires a local redirection mechanism.

## The SFX linking framework

The SFX framework that has been applied in all three experiments falls under Category 3 of the categorization shown in Table 2: the SFX framework builds on a local redirection component (Figure 2) and a local service component (Figure 3). Both components interoperate in order to achieve a functional context-sensitive linking system. Nevertheless, the modularity of the design guarantees that the SFX redirection component can potentially operate in an environment with non-SFX service components, and that the SFX service component can function with another redirection mechanism, as long as that mechanism supports transfer of link-source metadata as well as information regarding the origin of the link-source, to the SFX service component.

The SFX linking framework builds on a dynamic service component, that does not rely on precomputed relationships between documents. Rather, it will generate those relationships on-the-fly. Moreover, the SFX linking framework takes a just-in-time approach. The redirection mechanism will only be activated upon explicit request of the user and as such service links will not be generated unless the user has requested them. This is opposed to just-in-case linking approach, in which service links are provided regardless of the fact whether the user will use them or not. It has been shown that the just-in-time approach can dramatically reduce the data-processing delays inherent to dynamic linking.

### *The local redirection component of the SFX framework*

Information resources that can interoperate with SFX -- SFX-aware systems -- insert an SFX-button for each link-source in the result set of a query (left-hand part of Figure 2). This SFX-button is hyperlinked to point at the local redirection component and to contain link-source metadata. The metadata transported on this link can range from a unique identifier of the link-source to all link-source metadata. Consistent with the just-in-time approach of SFX, requesting extended services for a specific link-source is done by clicking this SFX-button. In response to this click, the local SFX redirection component can fetch link-source metadata --

usually -- from the origin resource in order to complete the metadata that was sent over the link (middle part of Figure 2). The SFX redirection component can perform this fetch using whichever protocol it takes. Next, link-source metadata as well as information on its origin will be converted into an interfacing format (right-hand part of Figure 2). At this point, the local redirection mechanism has fulfilled its task and is able to deliver this information to the local SFX service component.

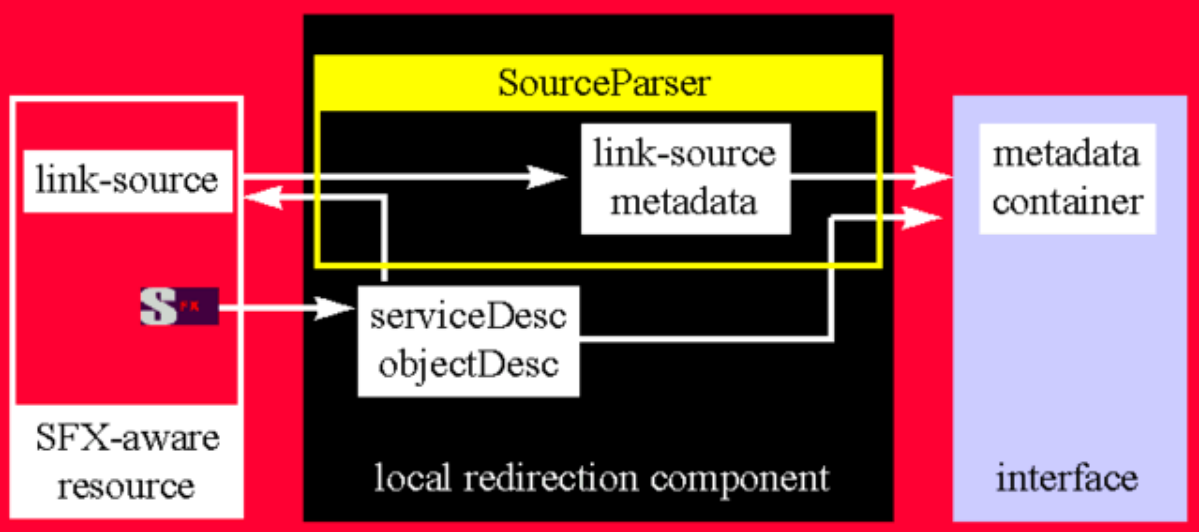


Figure 2: the local redirection component in the SFX framework

The experiments have also resulted in a modular approach towards the implementation of the functions crucial for the operation of the flow described above:

- The Cookiepusher is a pragmatic solution used to dynamically notify an information resource about the existence and the location of a local redirection component accessible to a user connecting to the resource.
- The consistent SFX-URL structure hyperlinking the SFX-button (Table 3) ensures that all local SFX redirection components can interoperate with potentially every information resource in the same way, meanwhile leaving a high degree of flexibility in the implementation of the URL to the information resource.
- The serviceDesc-specific SourceParsers stored at the end of the local redirection component are launched upon receipt of a request in the form of an SFX-URL, which is the result of a user clicking the SFX-button. The SourceParsers uniquely implement the following distinct functions:
  - The interpretation of the information contained in the objectDesc parameter based upon the syntax defined by the vendor;
  - The mechanism to fetch the link-source from its origin resource based on its origin and on the content of its objectDesc.
  - The conversion of the fetched link-source metadata, expressed in the metadata scheme supported by the authority running the origin resource, into a metadata container compliant with the scheme of the metadata interchange format.

GENERAL	target? <span style="color: red;">serviceDesc</span> & <span style="color: blue;">objectDesc</span>
DETAILED	local_SFX? <span style="color: red;">vendorId=&lt;theVendor&gt;&amp;databaseId=&lt;theBase&gt;&amp;objectDesc=&lt;theIdentifier&gt;</span>

Table 3: the consistent syntax of the SFX-URL

### ***About the feasibility of the local redirection component of the SFX framework***

It has been shown that:

- The SFX mechanism for local redirection was successfully applied in various complex environments, and for a wide variety of link-sources originating from distributed information scholarly information resources (see Table 1).
- The implementation of the requirements at the end of the information resources, to be interoperable with the SFX mechanism for local redirection (the CookiePusher and the consistent SFX-URL), causes a minor overhead. Nevertheless it has been pointed out that it is more straightforward to do so for resources that deliver information in a dynamic rather than in a static manner. This was the case for all but one resource dealt with through the course of the experiments.
- The design of the mechanism provides guarantees regarding the minimization of the overall implementation overhead because of the transferability of SourceParsers.
- The SFX mechanism for local redirection can be used to redirect namespace-specific identifiers -- such as the DOI -- to a local service component. This demonstrates the capability of the SFX redirection mechanism to facilitate the opening of closed linking frameworks and hence adapt them for context-sensitiveness.
- Several components of the implementation of the SFX mechanism for local redirection can be replaced by -- hopefully -- more robust alternatives. For instance, it has been noted that the Cookiepusher could be replaced by alternatives, that vary from the registration of the URL of the redirection component with each information resource, to the integration with an ongoing effort on authentication and authorization.

### ***The service component of the SFX framework***

The local service component (middle part of Figure 3) takes the information handed over by the local redirection component (left-hand part of Figure 3) as input, and begins by parsing it into a normalized internal representation object (GenRequest in Figure 3). During this process, the original content can be enhanced and/or augmented using a supporting database. The resulting information object is then fed into the SFX evaluation process in which it will be compared to the SFX-database. This is a special kind of linking database. Unlike traditional linking services, it does not contain any static links between "documents" (records/citations/full-text/etc.) of a collection. Rather, it contains a collection of conceptual services that express potential inter-relationships between documents at the level of the resource to which they belong. Source databases -- the ones that can interoperate with the SFX mechanism for local redirection -- are connected to Target databases -- the ones into which linking is possible -- by means of conceptual services. The validity of the connections is subject to boundary conditions named Thresholds (Figure 4). This approach turns the SFX service component into a fully dynamic linking system. The SFX evaluation process determines the relevance of each of these conceptual services using the content -- or the lack thereof -- in the information object. Next, the resulting bundle of relevant services is sent back to the user in the SFX-menu-screen (lower right of Figure 3). Tailoring the SFX-database -- Sources, Targets, conceptual services and Thresholds -- to reflect the local digital library collection makes those services context-sensitive. Only when the user decides to use a service from the bundle, will the service be resolved -- by TargetParsers -- into a URL to which the user is redirected (top right of Figure 3).

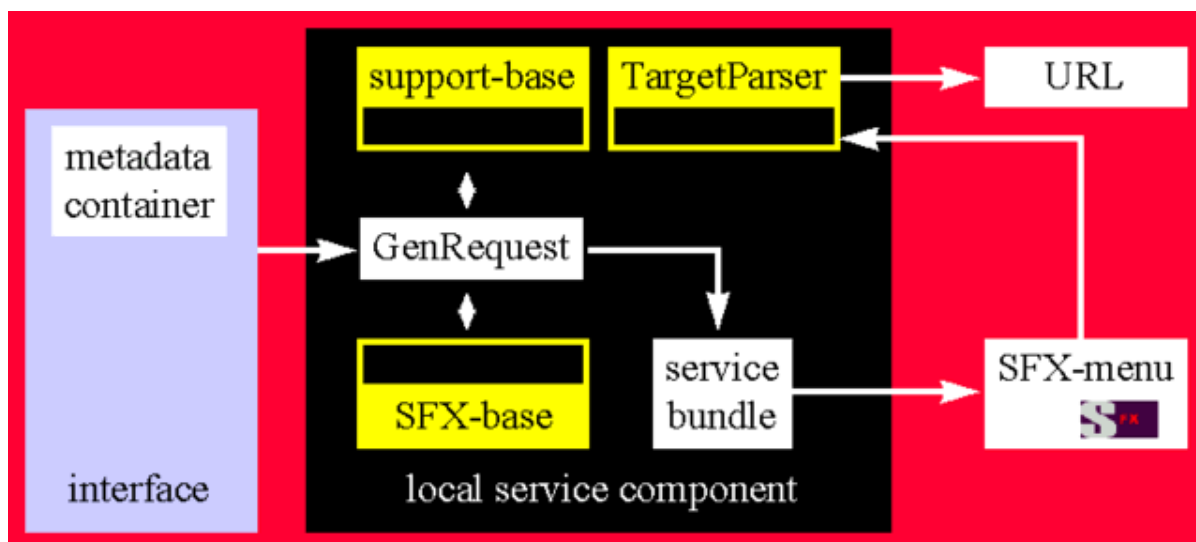


Figure 3: the local service component of the SFX framework

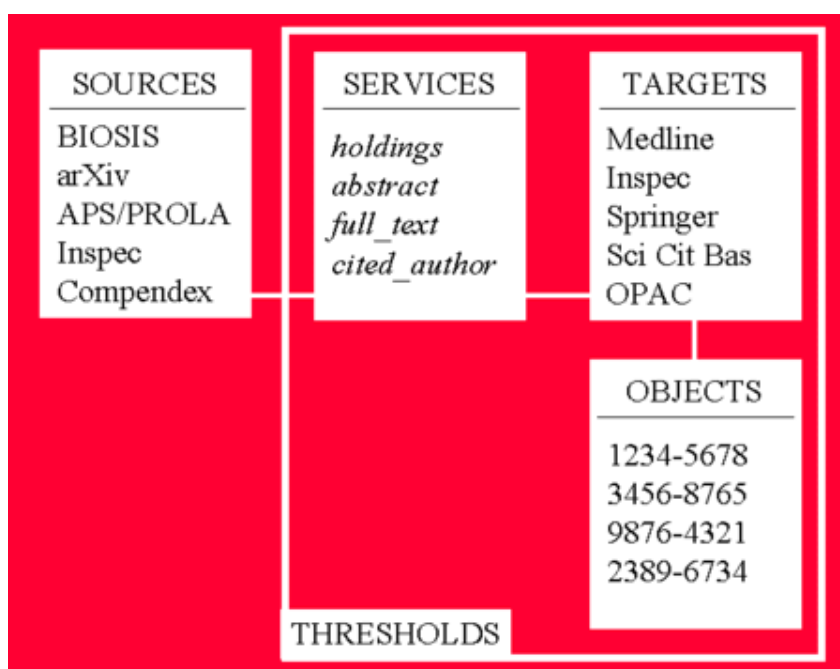


Figure 4: potential inter-relationships between information resources subject to Thresholds in the SFX linking service

#### About the feasibility of the local service component of the SFX framework

It has been shown that:

- The SFX service component is able to deliver a wide variety of extended services in a dynamic way. As shown in Table 1, both Sources and Targets cover the spectrum of commonly used scholarly information resources.
- The SFX service component is a neutral linking service, that can potentially be shared by all information resources of a digital library.
- The design of the SFX service component provides guarantees regarding the minimization of the overall implementation overhead in running the system and in making its services context-sensitive. This is due to the design of the SFX-database, that builds on the concepts of global and local relevance of services as well as global and local Thresholds, making it transferable and manageable. This is also due to the transferability of TargetParsers -- or alternatively -- the feasibility to use S-Link-S templates for the resolution of service links into URLs.



- The dynamic delivery of services, made possible by the notion of conceptual services on which the SFX local redirection component is built, has remarkable consequences. The addition of a new resource in a digital library environment causes a minor overhead in editing the SFX-database. Immediately, all existing conceptual services that had already been registered in the SFX-base become available for the new resource too. Also, at several points, the SFX service component was able to deliver service links for resources, where the static linking systems native to the resources failed to do so.
- By building on conceptual services that are subject to Thresholds, the SFX service component achieves a level of link-verification that can be situated between no verification at all and full verification of service links. It has been argued that this is acceptable:
  - In most cases, full verification of dynamic links is not possible since information resources rarely support a protocol for link-verification.
  - For several extended services, verification is irrelevant since they are shortcuts to searches rather than links in the traditional sense.
  - When an extended service only requires an identifier -- and no other metadata -- for its resolution, then this extended service is as foolproof as a static service.
  - Full verification would require the resolution of all service links into URLs, before the user has chosen to use a service from the bundle. In addition to this resolution, each service link would have to be verified (if possible). The combination of both actions would create significant delays in the delivery of service-links.

## Conclusion

It has been shown that a context-sensitive linking solution -- also named a system supportive of selective resolution -- consists of a redirection component and a service component that share an interface. By means of the local redirection approach taken in the SFX framework, it has been shown that it is feasible to transport link-source metadata originating from scholarly information resources to the doorstep of a service component. By means of the SFX service component, it has been shown that this transported link-source metadata can be used to dynamically generate context-sensitive extended services.

Therefore, the general conclusion is that -- by means of the SFX framework -- it has been demonstrated that it is feasible to interlink distributed electronic information resources in a dynamic and context-sensitive manner.

Moreover, the experiments conducted in the course of the thesis have emphasized the need for an open linking framework. Significant technical and conceptual progress has been made on the path that hopefully will lead towards the establishment of such a global context-sensitive framework. In addition, the experiments achieved more than a mere demonstration of the feasibility of dynamic linking. They also generated significant conceptual and technical progress in the realization of a linking solution that is able to deliver a wide range of extended services in a dynamic manner. They provided numerous indications that dynamic linking offers significant benefits when compared to static linking. Actually, it becomes hard to imagine how such extended services could at all be created in a static manner. Overall, the experiments demonstrated that the net result of the combination of context-sensitive and dynamic linking using the SFX framework, is a fully hypertextual scholarly information environment in which jumping between related distributed resources becomes possible. As far as can be verified, no other linking projects have shown comparable navigational capabilities.

## A closing comment about open linking

The experience with the SFX experiments taught us that the establishment of the means to support open, context-sensitive linking is highly dependent on the cooperation of the information industry. Many established players might be apprehensive about such an idea (Hitchcock et al.

1998b) as it requires far-reaching openness of their services. Proprietary solutions are part of a traditional strategy aiming at the minimization of competition (Porter 1979). A revival of that marketing concept can be found in many parts of the information industry, where the battle for the one stop shop market has exploded. Linking is considered to be a very important matter by major players in the information industry. The following is a citation of Elsevier's Karen Hunter (Hunter 1998):

*"In 1996 I said: 'One of the key roles a publisher should play in the future is creating links -- adding value by integrating information letting people maneuver through the space and get a full range of information.' Amen. My current motto is 'the publisher with the best links wins'. I don't lose sleep over this, but it's a mantra that I keep repeating to all who will listen. No publisher is an island, no information cannot be improved by enriching its context. (Pardon the double negative)."*

In due course, services of such importance will be subject to differential pricing. Wittingly or unwittingly, outsourcing such new information services to commercial parties will lead to a dependency on their integrated solutions. Outsourcing of scholarly publishing to commercial publishers has led to a pricing spiral (Bennett 1998). Although the literature abounds about the serials crisis, the problem should not be seen as restricted to the area of journal literature. At the core of the problem lies the notion of total dependency. It comes as no surprise, finding recent evidence of a sudden price increase with a factor of 3.5 for a commercial database service, after said service had been acquired by a main commercial player in the information industry (Case 1998). A similar situation may lay ahead for linking services, since closed linking frameworks in the hands of commercial parties will make the academic community completely dependent on those solutions, leaving no room for libraries to act in this domain. Hunter's quote not only stresses the importance of linking, it also calls for bridges between publishers -- currently being established -- without mentioning libraries.

Libraries should strive for an alteration of the existing, closed linking frameworks in a direction that enables them to fully exploit the collection they access, acquire or build. The pursuit of the means that enable the creation of extended services, like the ones illustrated in this thesis, should be high on the agenda of libraries worldwide. In the same manner as libraries are uniting in order to formulate guidelines for consortia deals (Turner and Yale University Library 1998), they should bring forward requirements for information systems that enable them to build upon and control extended services for the information they license or acquire. Therefore, the author is pleasantly surprised to be invited to present the open SFX linking framework at the occasion of the April 2000 meeting of the International Coalition of Library Consortia, in a session where publishers will present the closed CrossRef linking initiative (Spilka 1999b). Hopefully, this can become an opportunity to start and seriously discuss the enabling of context-sensitive linking on a broad scale.

This linking domain also opens an opportunity for the subversive (Okerson & O'Donnell 1995) initiatives in the area of scholarly communication to become more widely accepted via an integration into library services. Kling and Covi have already brought to our attention that the marginal situation of new and mostly free electronic-only journals (Harter and Kim 1996 ; Harter 1996) might be overcome by integrating those into the scholarly document system of libraries, indices and abstracting services (Kling and Covi 1995). As such, the adherence to an open framework for interlinking, that would enable libraries to deliver extended services for the alternative e-journals, might be part of the path leading to more general acceptance. A similar remark applies to the e-print servers, that turn out to be very successful in the intended user-community (Ginsparg 1994 ; Luzi 1998). Still, their integration into library services worldwide might be an impulse for a move from a successful subversive communication initiative to a wide-spread accepted publishing model that is more equitable and efficient (Van de Sompel



and Lagoze 2000, see Appendix) than the established ones. The UPS Prototype Project has given an early indication of the feasibility of such integration. It was felt that the navigational power it introduced could help to level the perceived difference in hierarchy between the established scholarly communication mechanisms and communication between scholars via e-print archives.

# References

---

American Economic Association. Journal of Economic Literature Classification System Menu. 1999. [<http://www.aeaweb.org/journal/elclasjn.html>].

American Mathematical Society. 2000 Mathematics Subject Classification. 1999. [<http://www.ams.org/msc/>].

Arms, William Y. 1993. Keynote address: the virtual library. Networking and the future of libraries. *Proceedings of the UK Office for library networking conference*. London: Meckler.

Atkins, Helen. 1999. The ISI® Web of Science® - Links and Electronic Journals: How links work today in the Web of Science, and the challenges posted by electronic journals. *D-Lib Magazine*, 5 no. 9. [<http://www.dlib.org/dlib/september99/atkins/09atkins.html>]

Bates, Marcia J. 1998. Indexing and access for digital libraries and the Internet: Human, database and domain factors. *Journal of the American Society for Information Science* 49, no. 13.

Bollen, Johan, Herbert Van de Sompel, and Luis Rocha. (in preparation). Mining associative relations from website logs and their application to context-dependent retrieval using spreading activation. *Proceedings of the Workshop on Organizing Webspaces (ACM-DL99)*, Berkeley, California.

Bollen, Johan and Frans Heylighen. 1998. A system to restructure hypertext networks into valid user models. *The new review of Hypermedia and Multimedia*. no. 4, pp. 189-213.

Boss, R. W. 1993. Online catalog functionality in the 90s: vendor responses to a Model RFP. *Library Technology Reports*, 29 no. 5

Bennett, Douglas C., et al. 1998. To publish and perish. *Policy Perspectives* 7, no. 4.

Bide, Mark. 1997. *In search of the Unicorn*. London: Book Industry Communication, BNBRF 89. [<http://www.bic.org.uk/bic/>].

Bush, Vannevar. 1945. As we may think. *Atlantic Monthly* 176, no. 1 (July). [<http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm>].

Canessa, Enrique. ICTP: One-Shot World-Wide Preprints Search. 1996. [<http://www.ictp.trieste.it/indexes/preprints.html>].

Canessa, Enrique and Giorgio Pastore. 1996. One-Shot Service Searches Preprint Repositories at a Mouseclick. *Computers in Physics* 10, no. 6: 520.

Caplan, Priscilla. 1999a. A model for reference linking. Report of the working group of the reference linking workshop; May 1999. [<http://www.lib.uchicago.edu/Annex/pcaplan/reflink.html>].

Caplan, Priscilla. 1999b. Report of the second workshop on linkage from citations to journal literature; June 9th 1999, Boston. [<http://www.niso.org/linkrept.html>].

Caplan, Priscilla and William Y. Arms. 1999. Reference linking for journal articles. *D-Lib Magazine* 5, no. 7/8. [<http://www.dlib.org/dlib/july99/caplan/07caplan.html>].

- Carr, Leslie and others. 1995. The distributed link service: a tool for publishers, authors and readers. *Proceedings of the fourth World Wide Web conference*.  
[<http://www.w3c.org/pub/Conferences/WWW4/Papers/178/>].
- Case, Mary M. 1998. ARL Promotes Competition through SPARC: The Scholarly Publishing & Academic Resources Coalition . *ARL Newsletter*, no. 196.  
[<http://www.arl.org/newsltr/196/sparc.html>].
- Caswell, Jerry V. and others. 1995. Importance and use of holdings links between citation databases and online catalogs. *The Journal of Academic Librarianship* 21, no. 2.
- Chudnov, Daniel. 1999. The jake project. [<http://jake.med.yale.edu>]
- Davis, James R. and Carl Lagoze. 1996. The Networked Computer Science Technical Report Library. *Cornell CS TR96-1595* .  
[<http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR96-1595>].
- de Haas, Boy. 1994. MECANO Project: Mechanism of automatic comparison of CD-ROM answers with OPACs. [<http://www.uba.uva.nl/nl/projecten/mecano/>]
- Dempsey, Lorcan. 1993. The future of library systems: integrated or insulated? Networking and the future of libraries. *Proceedings of the UK Office for library networking conference*. London: Meckler.
- Dempsey, Lorcan. 1995. The scandal of serials holding data. *Catalogue & Index*, no. 118.
- Evans, Nancy H. and others. 1989. The vision of the electronic library. *Mercury technical report series* 1. Carnegie Mellon University.
- Fox, Edward A. and others. 1997. Networked Digital Library of Theses and Dissertations An International Effort Unlocking University Resources. *D-Lib Magazine*.  
[<http://www.dlib.org/dlib/september97/theses/09fox.html>].
- French, James C., Allison L. Powell, and Eric Schulman. 1997. Automating the construction of authority files in digital libraries: a case study. *Technical Report No. CS-97-02*.  
[[http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.uva\\_cs/CS-97-02](http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.uva_cs/CS-97-02)]
- Gardner, William. 1990. The electronic archive: scientific publishing for the 1990s. *Psychological Science* 1, no. 6.
- Ginsparg, Paul. 1994. First steps towards electronic research communication. *Computers in Physics* 8, no. 4. [<http://xxx.lanl.gov/ftp/hep-th/papers/macros/blur>].
- Ginsparg, Paul, Rick Luce, and Herbert Van de Sompel. the Open Archives initiative. July 1999. [<http://www.openarchives.org/>].
- Ginsparg, Paul, Rick Luce, and Herbert Van de Sompel. First meeting of the Open Archives initiative. October 1999. [<http://www.openarchives.org/ups1-press.htm>].
- Halstead, Amy. 1999. PROLA: More Than Just a Pretty Acronym. *APS News* 8, no. 8.  
[<http://www.aps.org/apsnews/0899/089914.html>].
- Hamilton, Feona J. 1998. Multi-level linking technology by Swets. *Information World Review*, no. 142 (December).
- Harnad, Stevan. 1999. Integrating and navigating eprint archives through citation-linking (NSF /

JISC-eLib Collaborative Project). [<http://www.princeton.edu/~harnad/citation.html>].

Harnad, Stevan. 2000. CogPrints Project page.  
[<http://www.ukoln.ac.uk/services/elib/projects/cogprints/>]

Harter, Stephen P. 1996. The impact of electronic journals on scholarly communication: a citation analysis. *Public-Access Computer Systems Review* 7, no. 5.  
[<http://info.lib.uh.edu/pr/v7/n5/hart7n5.html>].

Harter, Stephen P. and Hak Joon Kim. 1996. Electronic journals and scholarly communication: A citation and reference study. *Proceedings of the midyear meeting of the American Society for Information Science*, San Diego, CA.  
[<http://php.indiana.edu/~harter/harter-asis96midyear.html>].

Hellman, Eric. 1998. Scholarly Link Specification Framework (SLinkS).  
[<http://www.openly.com/SLinkS/>].

Hitchcock, Steve and others. 1997a. Citation linking: improving access to online journals. *Proceedings of the 2nd ACM International Conference on Digital Libraries*, New York, USA: Association for computing machinery. [<http://journals.ecs.soton.ac.uk/acmdl97.htm>].

Hitchcock, Steve and others. 1997b. Linking everything to everything: Journal publishing myth or reality? ICCC/IFIP conference on electronic publishing '97: New models and opportunities. [<http://journals.ecs.soton.ac.uk/IFIP-ICCC97.html>].

Hitchcock, Steve and others. 1998a. Webs of research: putting the user in control. IRISS '98: Institute for learning and research technology, University of Bristol.  
[<http://sosig.ac.uk/iriss/papers/paper42.htm>].

Hitchcock, Steve and others. 1998b. Linking electronic journals: lessons from the Open Journal project. *D-Lib Magazine*, no. December.  
[<http://www.dlib.org/dlib/december98/12hitchcock.html>].

Hunter, Karen. 1998. Sleepless nights redux. *Against the Grain*, no. February.

International DOI Foundation. DOI Foundation Member List. January 1999.  
[<http://www.doi.org/idf-member-list.html>].

Kierman, Robert. 1998. The next five years: a publisher's ambition. *Serials* 11, no. 2.

King, Donald W. and Nancy K. Roderer. 1978. The electronic alternative to communication through paper-based journals. The information age in perspective: *Proceedings of the ASIS annual meeting*, 1978 White Plains, NY: Knowledge Industry Publications for American Society for Information Science.

Kling, Rob and L. Covi. 1995. Electronic journals and legitimate media in the systems of scholarly communication. *The Information Society* 11, no. 4.  
[<http://www.ics.uci.edu/~kling/klungej2.html>].

Knudson, Frances L. and others. 1997. Creating electronic journal web pages from OPAC records. *Issues in Science & Technology Librarianship* 15, no. Summer.  
[<http://www.library.ucsb.edu/istl/97-summer/article2.html>].

Krichel, Thomas. 1999. The Santa Fe Agreement: a discussion document presented at the Santa Fe Meeting of the Open Archives Initiative. [T.Krichel@surrey.ac.uk]

- Krichel, Thomas. RePEc Documentation. 2000a. [<http://netec.wustl.edu/RePEc/>].
- Krichel, Thomas. ReDIF version 1. 2000b. [[http://openlib.org/acmes/root/docu/redif\\_1.html](http://openlib.org/acmes/root/docu/redif_1.html)].
- Lasher, R. and Cohen D. A Format for Bibliographic Records. *Internet RFC-1807*. June 1995. [<http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1807.txt>].
- Lesk, M. E. 1978. Some applications of inverted indexes on the UNIX System. *Computing Science technical report* 69. Bell Laboratories, Murray Hill NJ.
- Luce, Rick. 1998. Integrating the Digital Library Puzzle: The Library Without Walls at Los Alamos . International Summer School on the digital library 1997 Tilburg: Ticer B.V. [<http://lib-www.lanl.gov/lww/tilberg.htm> ].
- Luzi, Daniela. 1998. E-print archives: a new communication pattern for grey literature. *Interlending and Document Supply* 26, no. 3.
- Lynch, Clifford A. 1997. Building the infrastructure of resource sharing: union catalogs, distributed search, and cross-database linkage. *Library Trends* 45, no. 3.
- Maly, Kurt, Michael Nelson, and Mohammad Zubair. 1999. Smart Objects, Dumb Archives: A User-Centric, Layered Digital Library Framework. *D-Lib Magazine* 5, no. 3. [<http://www.dlib.org/dlib/march99/maly/03maly.html>].
- NCBI. The NLM PubMed Project. 1998. [<http://www4.ncbi.nlm.nih.gov/pubmed/overview.html>].
- Nelson, Michael L. 1999. A Digital Library for the National Advisory Committee for Aeronautics. *NASA/TM-1999-209127* . [<http://techreports.larc.nasa.gov/ltrs/PDF/1999/tm/NASA-99-tm209127.pdf>].
- Nelson, Michael L. and others. 1998. NCSTRL+: Adding Multi-Discipline and Multi-Genre Support to the Dienst Protocol Using Clusters and Buckets. *Proceedings of Advances in Digital Libraries* 98. [<http://techreports.larc.nasa.gov/ltrs/PDF/1998/mtg/NASA-98-ieee-dl-mln.pdf>].
- Nelson, Michael L. and others. 1999. Buckets: Aggregative, Intelligent Agents for Publishing. *WebNet Journal* 1, no. 1: 58-66. [<http://techreports.larc.nasa.gov/ltrs/PDF/1998/tm/NASA-98-tm208419.pdf>].
- Millman, David. 1999. Cross-Organizational Access Management. A Digital Library Authentication and Authorization Architecture. *D-Lib Magazine* 5, no. 11. [<http://www.dlib.org/dlib/november99/11millman.html>].
- Needleman, Mark. 1999. Meeting report of the NISO linking workshop; February 11th 1999, Washington DC. [<http://www.niso.org/linkrpt.html>].
- Okerson, Ann and James O'Donnell. 1995. Scholarly Journals at the Crossroads; A Subversive Proposal for Electronic Publishing. Washington, DC: Association of Research Libraries. [<http://www.arl.org/scomm/subversive/toc.html>].
- Openly Inc. 1999. S-Link-S Calculator. June 1999. [<http://www.openly.com/SLinkS/Calculator/>].
- Open Archives initiative. 2000. The Santa Fe Convention. [<http://www.openarchives.org/sfc/sfc.htm>]

- Otlet, Paul. 1898. Le répertoire bibliographique universel. Compte rendu des travaux du congrès bibliographique international. Paris. 13 au 16 avril 1898.
- Paskin, Norman. 1999a. DOIs used for reference linking. Washington & Geneva. [<http://dx.doi.org/10.1000/143>].
- Paskin, Norman. 1999b. DOI: Current Status and Outlook. *D-Lib Magazine* 5, no. 5. [<http://www.dlib.org/dlib/may99/05paskin.html>].
- Pearl, A. 1989. Sun's link service: a protocol for open linking. *Hypertext '89 Proceedings*. New York: ACM.
- Pfeifer, Ulrich, Norbert Fuhr, and Tung Huynh. 1996. Searching Structured Documents with the Enhanced Retrieval Functionality of freeWAIS-sf and SFGate. *Proceedings of the Third International World Wide Web Conference*, pp 1027-36. [[http://www.igd.fhg.de/archive/1995\\_www95/papers/47/fwsf/fwsf.html](http://www.igd.fhg.de/archive/1995_www95/papers/47/fwsf/fwsf.html)].
- Plümer, Judith and Roland Schwänzl. 1996. Harvesting Mathematics. *Euromath Bulletin* 2, no. 1. [<http://www.mathematik.uni-osnabrueck.de/projects/harvest/euromath.ps.gz>].
- Plümer, Judith and Schwänzl, Roland. MPRESS. 1997. [<http://MathNet.preprints.org/>].
- Porter, Michael E. 1979. How competitive forces shape strategy. *Harvard Business Review*, no. March-April.
- Powell, James. Virginia Tech Federated Searcher. 1998. [<http://jin.dis.vt.edu/fedsearch/ndltd/support/search-catalog.html>].
- Powell, James and Ed Fox. 1998. Multilingual federated searching across heterogeneous collections. *D-Lib Magazine* 9, no. 4. [<http://www.dlib.org/dlib/september98/powell/09powell.html>].
- Publications Dept., ACM Inc. Computing Classification System. 1991. [<http://www.acm.org/class/1991>].
- Rayward, Boyd W. 1994. Visions of Xanadu: Paul Otlet (1868-1944) and Hypertext. *Journal of the American Society for Information Science*, 45 no. 4
- Schmitz, M. and others. 1995. A Uniform Bibliographic Code. *Vistas in Astronomy* 39: 272. [<http://cdsweb.u-strasbg.fr/abstract/simbad/refcode/refcode-paper.html>].
- Scott, Marianne. 1998. Library-Publisher relations in the next millennium: the library perspective. *IFLA Journal* 22, no. 5/6.
- Shishir, Gunavaram. 1996. CGI Programming on the World Wide Web. Sebastopol, CA.: O'Reilly and Associates, Inc.
- Spilka, Susan. 1999a. Wiley InterScience Update. June 1999. [<http://www.wiley.com/about/corpnews/wisupdate.html>].
- Spilka, Susan. 1999b. Reference Linking Service to Aid Scientists Conducting Online Research. December 1999. [<http://www.doi.org/ref-link-press-release-11-99.html>].
- Sun. September 29, 1999. InfoDoc #19895.
- Tiffany, Melissa E. and Michael L. Nelson. 1998. Creating a Canonical Scientific and Technical

Information Classification System for NCSTRL+. *NASA/TM-1998-208955* .  
[<http://techreports.larc.nasa.gov/ltrs/PDF/1998/tm/NASA-98-tm208955.pdf>].

Turner, Bonnie and Yale University Library. International Coalition of Library Consortia. March 1998. [<http://www.library.yale.edu/consortia/>].

Van de Sompel, Herbert. 1991. Heading towards an electronic library: location independent integration of electronic reference sources in library workstations. 10th Annual meeting of the Dobis/Libis User Group. Leuven: Dobis/Libis User Group Secretary.

Van de Sompel, Herbert. 1993. Optimalisatie van de konsultatieketen aan de Universiteit Gent. *Bibliotheekkunde* 51. Kris Clara and Julien Van Borm. Antwerpen: VVBAD.

Van de Sompel, Herbert and Guido Van Hooydonk. 1994. Technology and collaboration: creating an effective information environment in an academic context. Online Information 94. *Proceedings of the 18th International Online Information Meeting*. Oxford and New Jersey: Learned Information (Europe) Ltd.

Van de Sompel, Herbert. 1997a. Integrating CD-ROMs in the digital library. International Summer School on the digital library 1997. Tilburg: TICER B.V.

Van de Sompel, Herbert. 1997b. Tools for the digital library. From database networking to the digital library Padua.

Van de Sompel, Herbert, Thomas Krichel, Michael L. Nelson and others. 2000. The UPS Prototype: an experimental end-user service across e-print archives. *D-Lib Magazine* 6, no. 2. [<http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html>].

Van de Sompel, Herbert and Carl Lagoze. 2000. The Santa Fe Convention of the Open Archives initiative. *D-Lib Magazine* 6, no. 2.  
[<http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html>].

Wang, Peiling and White, Marilyn Domas. 1999. A cognitive model of document use during a research project. Study II. Decisions at the reading and citing stages. *Journal of the American Society for Information Science* 50. no. 2.

Weislogel, Judy. 1998. Elsevier Science Digital Libraries Symposium. *Serials Review* 24, no. 2.