

Old Dominion University ODU Digital Commons

Computer Science Faculty Publications

Computer Science

2000

The UPS Prototype: An Experimental End-User Service Across E-Print Archives

Herbert Van de Sompel

Thomas Krichel


Michael L. Nelson
Old Dominion University

Patrick Hochstenbach

Victor Lyapunov

See next page for additional authors

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_fac_pubs

 Part of the [Databases and Information Systems Commons](#), and the [Digital Communications and Networking Commons](#)

Repository Citation

Van de Sompel, Herbert; Krichel, Thomas; Nelson, Michael L.; Hochstenbach, Patrick; Lyapunov, Victor; Maly, Kurt; Zubair, Mohammad; Kholief, Mohamed; Liu, Xiaoming; and O'Connell, Heath, "The UPS Prototype: An Experimental End-User Service Across E-Print Archives" (2000). *Computer Science Faculty Publications*. 9.
https://digitalcommons.odu.edu/computerscience_fac_pubs/9

Original Publication Citation

Van de Sompel, H., Krichel, T., Nelson, M., Hochstenbach, P., Lyapunov, V., Maly, K., . . . O'Connell, H. (2000). The UPS prototype: An experimental end-user service across e-print archives. *D-Lib Magazine*, 6(2). doi: 10.1045/february2000-vandesompel-ups

Authors

Herbert Van de Sompel, Thomas Krichel, Michael L. Nelson, Patrick Hochstenbach, Victor Lyapunov, Kurt Maly, Mohammad Zubair, Mohamed Kholief, Xiaoming Liu, and Heath O'Connell

D-Lib Magazine **February 2000**

Volume 6 Number 2

ISSN 1082-9873

The UPS Prototype

An Experimental End-User Service across E-Print Archives

Herbert Van de Sompel

Los Alamos National Laboratory - Research Library, New Mexico, US, and
Automation Department of the Central Library of the University of Ghent, Belgium
herbert.vandesompel@rug.ac.be.

Thomas Krichel

University of Surrey, UK
T.Krichel@surrey.ac.uk

Michael L. Nelson

NASA Langley Research Center, Hampton VA, USA
m.l.nelson@larc.nasa.gov

Patrick Hochstenbach

Automation Department of the Central Library of the University of Ghent, Belgium
patrick.hochstenbach@rug.ac.be

Victor M. Lyapunov

Institute for Economics and Industrial Engineering
Siberian Branch of the Russian Academy of Sciences, Russia
vic@ieie.nsc.ru

Kurt Maly

Old Dominion University, Norfolk VA, USA
maly@cs.odu.edu

Mohammad Zubair

Old Dominion University, Norfolk VA, USA
zubair@cs.odu.edu

Mohamed Kholief

Old Dominion University, Norfolk VA, USA
kholief@cs.odu.edu

Xiaoming Liu

Old Dominion University, Norfolk VA, USA
liu_x@cs.odu.edu

Heath O'Connell
Stanford Linear Accelerator Center, Stanford, USA
hoc@slac.stanford.edu

Abstract

A meeting was held in Santa Fe, New Mexico, October 21-22, 1999, to generate discussion and consensus about interoperability of publicly available scholarly information archives. The invitees represented several well known e-print and report archive initiatives, as well as organizations with interests in digital libraries and the transformation of scholarly communication. The central goal of the meeting was to agree on recommendations that would make the creation of end-user services -- such as scientific search engines and linking systems -- for data originating from distributed and dissimilar archives easier. The Universal Preprint Service (UPS) Prototype was developed in preparation for this meeting. As a proof-of-concept of a multi-discipline digital library of publicly available scholarly material, the Prototype harvested nearly 200,000 records from several different archives and created an attractive end-user environment. This paper describes the results of the project. This is done in two ways. On the one hand, the experimental end-user service that was created during the project is illustrated. On the other hand, the lessons that the project team drew from the experience of creating the Prototype are presented.

The UPS Prototype project: team, goals, motivation and relation to the Santa Fe Convention

The main aim of the meeting of the Open Archives initiative ([Ginsparg, Luce and Van de Sompel 1999](#)) was to agree on recommendations that would make the creation of end-user services -- such as scientific search engines and linking systems -- for data originating from distributed and dissimilar e-print archives easier. The UPS Prototype project was developed in preparation for this meeting.

The prototype was a feasibility study for the creation of cross-archive end-user services. With the premise that users would very much prefer to have access to a federation of digital libraries, the main 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. This paper describes the most fundamental problems that occurred during the project, and looks at those that are related to the specific nature of e-print collections. The project was also a testbed for the digital library technologies that were selected to create the prototype. These issues are out of the scope of this paper but are described at length in a companion paper ([Van de Sompel, Krichel, Nelson, et al. 2000](#)).

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. The coordinators of the project were Herbert Van de Sompel, Thomas Krichel and Michael Nelson, each of whom brought additional researchers into the project. Most of them never met in person; project communication has mainly been conducted via a list server. Work 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 21, 1999. Experience developed during the prototype was one of the foundations of the concepts brought forward in [the Santa Fe Convention \(Van de Sompel and Lagoze 2000; Open Archives initiative 2000\)](#) that formalizes the recommendations resulting from the Santa Fe discussions.

The aim of the prototype was to demonstrate a multidisciplinary end-user service that covers a variety of e-print and reports services, as a special instance of a cross-archive service. Most e-print and reports services have a focus in a specific discipline. Beyond the communities of scholars that are aware of the existence of discipline-specific points of entry to e-print information, little effort has been made to serve the

communities of libraries, students and interdisciplinary researchers for whom multidisciplinary services are important tools. In addition to increasing the accessibility of e-print data, the existence of such a service helps raise the awareness of alternative communication mechanisms outside a core group that no longer needs convincing.

The amount of cross-archive end-user services is limited (see, for instance, [Plömer and Schwanzl 1996](#); [Plömer and Schwanzl 1997](#); [Canessa and Pastore 1996](#); [Canessa 1996](#); [Powell 1998](#); [Powell and Fox 1998](#)). Most of them are prototypical 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 made the UPS Prototype project a challenging, realistic feasibility study, since it anticipated 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 archives 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)	http://arXiv.org/
CogPrints	(Harnad 2000)	http://cogprints.soton.ac.uk/search
NACA	(Nelson 1999)	http://naca.larc.nasa.gov/
NCSTRL	(Davis and Lagoze 1996)	http://www.ncstrl.org/
NDLTD	(Fox, et al. 1997)	http://www.theses.org/
RePEc	(Krichel 2000a)	http://netec.mcc.ac.uk/WoPEc.html http://ideas.uqam.ca http://netec.wustl.edu/NEP 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 illustrated in [Table 2](#) and [Table 3](#):

- 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 cannot be derived merely from the origin of the record.
- **Scale:** Of the six archives, arXiv is the largest with about 130,000 objects in the collection. It also is the most diverse in terms of contributing authors and institutions ([Table 3](#)).

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 UPS Prototype project aimed to create a cross-archive search facility with an overlaid linking service. The different phases leading to the creation of the prototype were:

- [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 was collected from the originating archive initiatives. This data was stored in a new repository that was the subject of the end-user services created. For each of the archives, the data was 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 was collected; the full-content associated with that metadata was left in the originating archives and hence, the end-user service points at the full-content there.

[Table 3](#) presents some figures related to the data harvested for the project. The "Records" column shows the total amount of records resulting from the data-gathering phase for each archive initiative. The "Records in UPS" shows the amount that made it into the Prototype. The "Records linked to full content" shows for how many of the UPS records an associated full-text file exists.

ARCHIVE INITIATIVE	RECORDS	RECORDS IN UPS	RECORDS LINKED TO FULL CONTENT
ArXiv	128,943	85,204	85,204
CogPrints	743	742	659
NACA	3,036	3,036	3,036
NCSTRL	29,690	25,184	9,084
NDLTD (*)	1,590	1,590	951
RePEc	71,359	71,359	13,582
Total	235,361	187,115	112,516

Table 3: Figures regarding the amount of collected and processed records

(*) Only data from NDLTD Virginia Tech is involved in the experiment.

Metadata conversion

The metadata collected during the data gathering phase is expressed in a variety of metadata formats. There are as many metadata formats as there are archives. The [ReDIF version 1](#) format ([Krichel 2000b](#)) -- as used in the RePEc initiative -- was chosen to be the common metadata format for the UPS Prototype project and all data was converted into this format. Consequently, conversion procedures had to perform a mapping of non-ReDIF metadata to the ReDIF format. 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 conversion to ReDIF, important efforts are undertaken to augment the quality of the metadata. The most important interventions made to the metadata relate to:

- Achieving an appropriate level of subtagging of metadata elements: Several input metadata formats do not have adequate subtagging of important data fields. In order to address these problems, the native data was parsed via several routines that use heuristics to try and achieve the desired level of subtagging.
- Creation of a UPS namespace of unique identifiers: It was decided to accord unique UPS identifiers to metadata records originating from all archives. The hierarchical structure of the UPS identifiers is derived from the identifier-conventions used in ReDIF. All record identifiers in UPS start with a value that unambiguously reveals the origin archive of the record.
- Dealing with multiple manifestations of a work: An e-print is 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. In some metadata sets *post e-print* information is available within the e-print metadata record. It was decided to maintain this concept in the converted data: e-print metadata and *post e-print* metadata related to the same work were stored in distinct zones of a single metadata record.
- Subject classification and keywords: Subject classification schemes that are native to various input datasets were preserved and values were stored in separate ReDIF-tags for each scheme. Author supplied keywords that do not follow a controlled vocabulary were mapped into a shared Keywords tag. In addition, an attempt was made to add a broad subject classification to the complete collection. A scheme from NASA -- as proposed in ([Tiffany and Nelson 1998](#)) -- was chosen. The classification values were generated by an automated process, which mapped information on the research domain of archive data into the common scheme.

While the above illustrates that important steps were taken to try and achieve an appropriate level of metadata quality, the short project time frame prevented the creation of an optimal metadata collection.

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, aggregate, intelligent digital objects optimized for use in digital libraries. 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. 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 is a general concept

that is independent of digital library protocols and systems.

Individual SODA archives were created for arXiv, CogPrints, NACA, NCSTRL, NDLTD and RePEc. The ReDIF metadata files are used as seeds for the creation of buckets. [Screencam 1](#) as well as [Figure 4](#) & [Figure 7](#) -- in the Project Results section -- demonstrate some of the methods available for buckets in the UPS environment. [Screencam 3](#) and [Screencam 4](#) show buckets being approached via the NCSTRL+ user interface and how they are integrated with the SFX linking service.

Creation of NCSTRL+ end-user search facility

It was decided to use the NCSTRL+ environment for the creation of the end-user search facility. NCSTRL+ is an extension of the [Dienst](#) architecture ([Lagoze, Shaw, Davis, et al. 1995](#)) that supports buckets and clusters. Clusters provide a way to partition a dataset along predefined metadata axes. Each cluster divides the dataset into virtual sub-collections.

By building on its [Dienst](#) heritage, the [UPS search interface](#) provides a simple and advanced search interface. The interface was redesigned in order to make it more aesthetically pleasing and comprehensible than the original NCSTRL/NCSTRL+ interface. The simple search option 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. Search results can be resorted according to any cluster without the need to redo the actual search. The search results show brief information from metadata records. Selection of a search result leads the user outside the NCSTRL+ environment, because it causes the bucket corresponding to the search result to display itself.

Addition of a SFX linking service

To link references between archives, the UPS prototype end-user service interfaces to the SFX context-sensitive linking service. This provides a concrete illustration of integrating an e-print environment with other information resources, which is an important aspect of e-print interoperability ([Harnad 1999](#)).

The details of the SFX framework are described in ([Van de Sompel and Hochstenbach 1999a](#)); ([Van de Sompel and Hochstenbach 1999b](#)) and ([Van de Sompel and Hochstenbach 1999c](#)). The fundamental aim of the SFX linking framework is to link information within digital library collections in a fully dynamic way. When an information system is interoperable with SFX -- SFX-aware --, it presents each search result with an accompanying SFX-button. When the SFX-button of a search result is clicked, a list of services that are relevant for this result is dynamically generated. The services SFX generates are context-sensitive. The services will reflect the nature of the digital library collection that is accessible to the user: the service links will point back into the collection of the user's institution.

In the UPS Prototype project the digital object buckets -- rather than the NCSTRL+ service -- are SFX-aware. This turns out to be a very interesting feature. For instance, when a bucket contains metadata describing both an e-print and a formal publication (see [Metadata Conversion](#)) it will autonomously decide to display two SFX-buttons -- one for the e-print and one for the publication. As can be seen from [Screencam 3](#), [Screencam 4](#) and [Figure 7](#) to [Figure 13](#) -- in the Project Results section -- the two buttons may generate quite different services when clicked.

In the course of the UPS Prototype project, the SLAC/SPIRES system has been made SFX-aware. SLAC/SPIRES is a free citation database for high-energy physics. It contains all the references from published papers and e-prints in that research domain. SLAC/SPIRES can be used to dynamically request a list of all references of a publication, using metadata about the publication as parameters. This service is integrated in SFX. When a user requests extended services for buckets describing a high-energy physics e-print, one of the services allows him to get the reference list of the e-print from SLAC/SPIRES. But since SLAC/SPIRES is SFX-aware, all references returned by SLAC/SPIRES are equipped with an SFX-button, allowing the user to request extended services for each 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.

Summary of the problems arising during the creation of the UPS Prototype

Construction of this prototype demonstrated several issues that are likely to recur in any attempt to build a federation among independently managed archives:

- In order to obtain the archive data, the maintainers of the archives need to be contacted individually.
- 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.
- 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.
- Overall, the lack of a shared framework organizing all of the above hinders a smooth operation of the Prototype project.
- Numerous problems that occur are related to metadata-issues:
 - Metadata formats used in individual archives lack an appropriate level of subtagging;
 - Metadata obtained from individual archives is inconsistent and not normalized;
 - Even extensive interventions during the metadata conversion phase could not prevent the negative impact that poor metadata quality has on the search and linking facilities developed in the course of the project;
 - There is a need for mutually agreed namespaces for e-print identifiers, author names, author-affiliations, subject classification schemes, etc.

Project results

As a result of the UPS Prototype project, an experimental cross-archive end-user service was presented at the Santa Fe Meeting of the Open Archives initiative. At 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 in the [Appendix](#) by means of Lotus Screencam movies and screendumps that show how a user navigates the multidisciplinary UPS Prototype environment.

The recommendations made to the Open Archives group

The most important results of the Prototype project are the lessons learned while actually doing the project and the insights gained in the problems related to the creation of cross-archive services. These insights -- briefly summarized above -- enable the identification of some crucial areas where work is required in order to more easily accommodate the creation of third party services. As a result of the UPS Prototype project, the coordinating trio drew some conclusions and presented them to the Open Archives group, as a means to facilitate the discussions at the Santa Fe meeting.

A strong distinction between data provider and service provider

As can be seen from [Table 2](#), all archive initiatives that are part of the project have a storage facility and a submission mechanism. Most of them also have a native user-service, while they lack appropriate features for data-extraction. This typical architecture is represented in [Figure 1](#). In order to enable the easy creation of third-party services, this concept needs to be revised. To enable federation with other archives, the essential functions of an archive must be:

- a submission mechanism;
- a long-term storage facility;
- a machine interface that allows for data-access by third-parties.

A native end-user service remains a possible, but not an essential, function of an archive. The proposed open archive architecture is represented in [Figure 2](#).

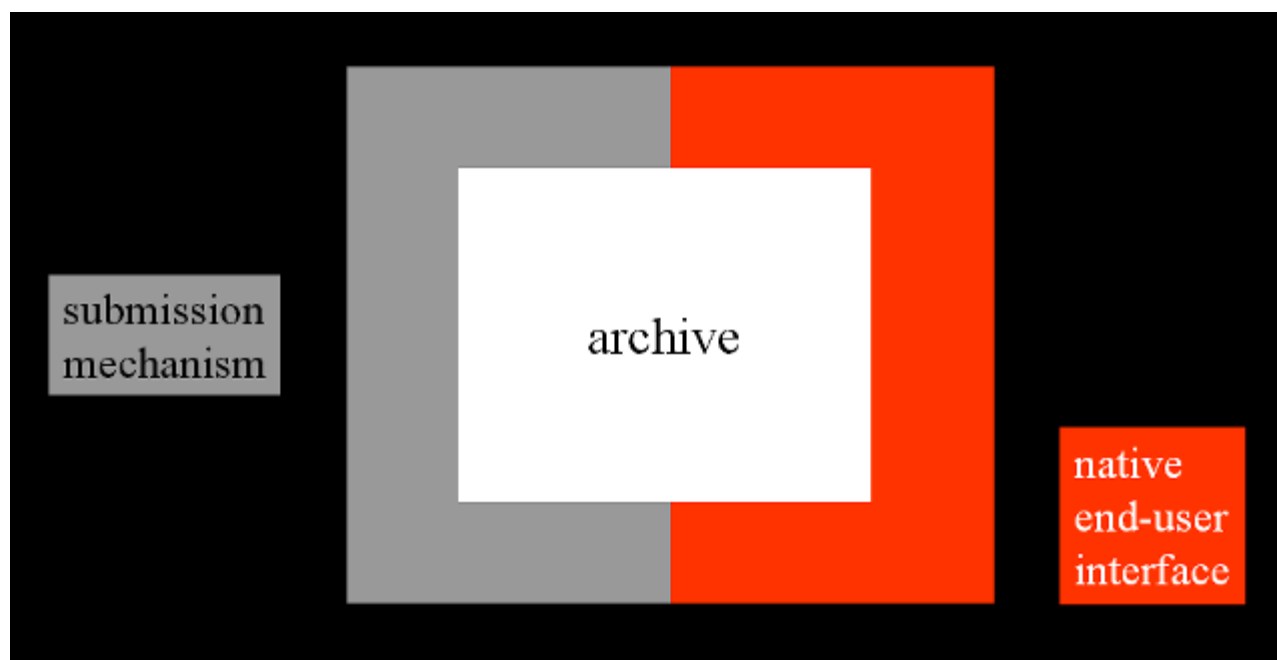


Figure 1: Typical archive architecture

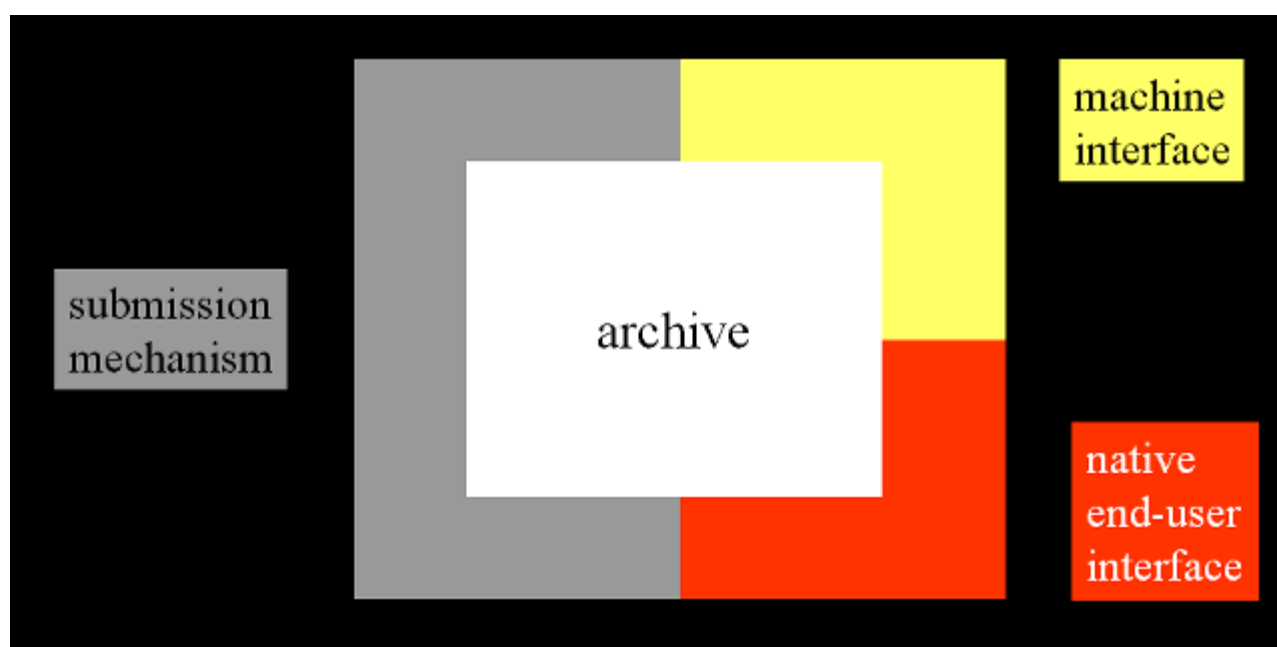


Figure 2: Open archive architecture

As a result of such a reorientation, a scenario emerges in which the maintainers of archives can be seen as data providers that concentrate on the crucial functions of data-submission and long-term storage of submitted data. Minimally, these data providers also enable third-party service providers to create end-user services based on their data. In some cases, they fully rely on them for this. This concept is represented in [Figure 3](#).

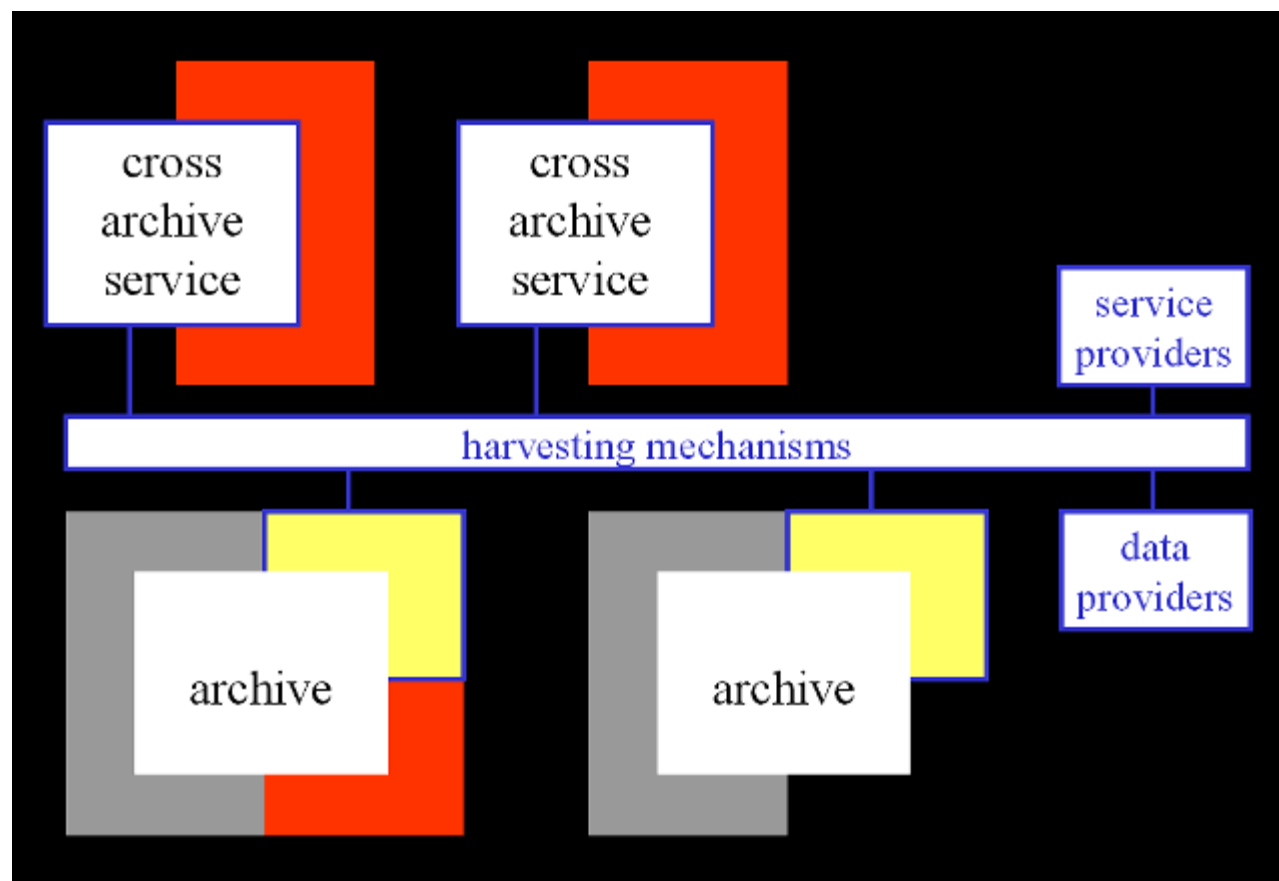


Figure 3: Data providers and service providers

A harvesting approach with rich harvesting criteria

As a mechanism for data-access by service providers, a harvesting approach rather than a distributed search scenario is proposed. The main justification for this approach is the ease with which these are able to be part of a highly distributed environment. Web-harvesting tools have sufficiently demonstrated this property that will become crucial in anticipation of a situation in which thousands of archives, both discipline-specific and institution-based, may become available. Harvesting approaches have the additional attractive property that they allow data enhancing procedures to be run on the collected data.

Enhancements such as normalization, augmentation and restructuring are applied to data originating from different sources in order to create consistent end-user services. In a harvesting scenario, these activities can be dealt with in a batch manner. Distributed searching solutions may ultimately develop the same degree of scalability and flexibility, but given the current status of research, that may not happen on the short to mid-term time range that the Open Archives initiative has set itself to outline and implement a workable set of interoperability recommendations.

For this purpose, data providers need to support a harvesting interface enabling service providers to periodically poll the archives to collect data that is relevant for their end-user services. Such end-user services can create their identity in a variety of ways. Services can be discipline-oriented,

multidisciplinary, regional, only for discovery of certain types of material, etc. Therefore, a harvesting interface should support a variety of harvesting criteria such as Accession Date; Publication Date; Subject Classification; Author Affiliation.

A framework that brings together data providers and service providers

An organizational framework needs to be established to bring data providers and service providers together. Such a framework must establish trust between the parties involved, promote a spirit of collaboration and provide access to information that is required for all parties to operate successfully in the data-provider/service-provider space. The framework needs to address:

- **Terms and conditions of data use:** In order to overcome the ambiguity regarding the terms and conditions of use of archive data that is identified in the course of the project, the framework must accommodate a means by which data providers can express both the openness and limits of the usage of their archives. Data providers need a means to express that service providers are welcome to use archive data for any purpose, provided that they tell the data provider how they are using the data and that they are not using it in a way that the data provider has explicitly prohibited.
- **Technical characteristics:** The framework must accommodate a way in which the data providers can describe the technical characteristics of their archive, such as harvest protocols, harvest criteria, metadata format(s), subject classification scheme(s), material-type scheme(s), identifier scheme(s), etc.
- **Administrative characteristics:** The framework must also provide supporting information such as contact addresses for data providers and service providers, the content of archives and services, etc.

A concrete proposal concentrating on the terms and conditions was prepared and presented to the Open Archives group ([Krichel 1999](#)).

A universal unique identifier namespace for e-prints

There is a need for a shared universal namespace for e-print identifiers just as they exist for journals (ISSN), books (ISBN) and articles, e.g., Astrophysics Data System Bibcode ([Schmitz et al. 1995](#)), DOI ([Paskin 1999](#)) and PubMed ID ([NCBI 1998](#)). Such identifiers become crucial as a means to keep track of the provenance of e-print data when it starts to penetrate into scholarly resources outside the native archive. This is important in de-duplication processes, in linking services, for citations and citation databases as well as to track the lifecycle of a research effort from the e-print, to the publication, to citations.

In addition, namespaces for metadata-elements that are important as a harvesting criterion -- author affiliation, discipline -- are desirable.

Metadata

The lack of quality of the metadata available in the UPS Prototype project has an important, baleful influence on the creation of cross-archive services as well as on the quality of services that can be created. In order to solve this problem, data enhancement procedures need to be run to improve the quality of existing metadata in archives. In parallel with that, an exploration of submission techniques is required, in order to identify ways in which the data quality can be improved at the source, without demotivating authors by requiring them to submit material with lengthy and complex submission mechanisms.

The addition to the e-print metadata of the references made in the e-print is desirable. As illustrated via e-prints of the high energy sub-archive of arXiv, for which references are available in the SLAC/SPIRES database, doing so can create a seamless integration of the e-print environment with other scholarly information resources. Such integration is attractive from a navigational point of view. But even more

important is that this navigational power can help to level the perceived difference in hierarchy between the established scholarly communication mechanisms and communication via e-print archives. Ultimately, the availability of references in the metadata can also lead to the creation of citation databases services such as the Universal Citation Database envisioned by Cameron ([Cameron 1997](#)), in which scholarly work is included without regard to the venue in which it has been published. Again, the leveling power of such services can be an important tool in the transformation of scholarly communication.

Conclusion

In a four-month timeframe, the project team demonstrated the feasibility of creating a cross-archive end-user service by means of its UPS Prototype system. The project 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 were incorporated in recommendations made to the Open Archives group during their first meeting in Santa Fe. An important concern in the formulation of these recommendations was the balance between the efforts required at the end of the data provider to implement the recommendations and the objective of making it easy for the service provider. The Santa Fe Convention, that is the result of the discussions at that meeting, has adopted many of these recommendations.

The group has adopted the concept of the data-provider/data-implementer space. It has also concluded that -- in a short to mid-term timeframe -- a harvesting approach gives more guarantees for success than a distributed searching approach. For some recommendations, the group went further than the coordinating trio. This was the case for the harvesting protocols and the metadata formats, for which the coordination trio left the choice of these to the data providers, but required them to document their choices in the data-provider/service-provider framework. The group agreed on a standardized harvesting protocol and on a joint basic metadata set, leaving the options open for community-specific metadata formats to be used in parallel. For other recommendations, the group did not go as far as the coordinating trio. This was the case for the "Terms and conditions of data use" aspect of the data-provider/service-provider framework, that is only addressed in a minimal way by the Convention, reflecting the "don't ask don't tell" policy of some archive maintainers. The e-print identifier namespace concept has found its way into the Convention, be it in a pragmatic rendering that can ultimately lead to a more elaborate implementation.

Other recommendations were outside the immediate scope of the first meeting, and they might become elements for future meetings. This is definitely the case for the submission mechanism and its relation to the quality of the metadata. Other issues that might be studied later include identifiers for other metadata elements such as author-affiliation and subject, and also the case for the inclusion of citation data in the metadata.

The project also provided interesting new insights regarding the digital library technologies that have been used. As mentioned in the opening section of this paper, these are out of the scope of this paper, but are described in detail in ([Van de Sompel, Krichel, Nelson et al 2000](#)).

References

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. [http://lib-www.lanl.gov/~jbollen/pubs/Bollen_wows_DL99.zip]

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. [http://lib-www.lanl.gov/~jbollen/pubs/JBollen_NRHM99.pdf]

- Cameron, Robert D. 1997. A universal citation database as a catalyst for reform in scholarly communication. First Monday 2, no. 4. [http://www.firstmonday.dk/issues/issue2_4/cameron/index.html]
- 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.
- 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>]
- 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>]
- Ginsparg, Paul. 1994. First steps towards electronic research communication. Computers in Physics 8, no. 4: 390-6. [<http://arXiv.org/blurb/blurb.ps.gz>]
- 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>]
- Harnad, Stevan. Integrating and navigating eprint archives through citation-linking (NSF / JISC - eLib Collaborative Project). June 1999. [<http://www.princeton.edu/~harnad/citation.html>]
- Harnad, Stevan. 2000. CogPrints Project page. [<http://www.ukoln.ac.uk/services/elib/projects/cogprints/>]
- 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].
- Lagoze, Carl, E. Shaw, J. R. Davis, and D. B. Krafft. Dienst Implementation Reference Manual. Cornell Computer Science Technical Report TR95-1514, May 1995, [<http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR95-1514>]
- 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. 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-ieeeedl-mln.pdf>]

Nelson, Michael L. 1999. A Digital Library for the National Advisory Committee for Aeronautics. NASA/TM-1999-209127 April . [<http://techreports.larc.nasa.gov/ltrs/PDF/1999/tm/NASA-99-tm209127.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>]

Open Archives initiative. 2000. The Santa Fe Convention. [<http://www.openarchives.org/sfc/sfc.htm>]

Paskin, Norman. 1999. DOI: Current Status and Outlook. D-Lib Magazine 5, no. 5. [<http://www.dlib.org/dlib/may99/05paskin.html>]

Plömer, Judith and Roland Schwönzl. 1996. Harvesting Mathematics. Euromath Bulletin 2, no. 1. [<http://www.mathematik.uni-osnabrueck.de/projects/harvest/euromath.ps.gz>]

Plömer, Judith and Schwönzl, Roland. MPRESS. 1997. [<http://MathNet.preprints.org/>]

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>]

Schmitz, M. and others. 1995. NED and SIMBAD Conventions for Bibliographic Reference Coding. Information & On-line Data in Astronomy p. 259. [<http://cdsweb.u-strasbg.fr/abstract/simbad/refcode/refcode-paper.html>]

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>]

Van de Sompel, Herbert and Patrick Hochstenbach. 1999a. Reference linking in a hybrid library environment. Part 1: frameworks for linking. D-Lib Magazine 5, no. 4. [http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html]

Van de Sompel, Herbert and Patrick Hochstenbach. 1999b. Reference linking in a hybrid library environment. Part 2:SFX, a generic linking solution. D-Lib Magazine 5, no. 4.

[http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt2.html]

Van de Sompel, Herbert and Patrick Hochstenbach. 1999c. Reference linking in a hybrid library environment. Part 3: Generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment. D-Lib Magazine 5, no. 10. [http://www.dlib.org/dlib/october99/van_de_sompel/10van_de_sompel.html]

Van de Sompel, Herbert, Thomas Krichel, Michael L. Nelson and others. 2000. The UPS Prototype project: exploring the obstacles in creating a cross e-print archive end-user service. Old Dominion University Computer Science TR 2000-01, February 2000. [http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.odu_cs/TR_2000_01]

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>]

Acknowledgments

The authors wish to thank the following parties for the provision of e-print data:

- Paul Ginsparg - [arXiv.org](http://arxiv.org) archive
- Stevan Harnad and Robert Tansley - [CogPrints](http://cogprints.org) archive
- Michael L. Nelson - [NACA](http://naca.com) collection
- Carl Lagoze - [NCSTRL](http://ncstrl.org) initiative
- Ed Fox, Anthony Atkins - [NDLTD](http://ndltd.org) initiative
- Thomas Krichel - [RePEc](http://repec.org) initiative

The authors wish to thank the following parties for their active cooperation in the project:

- Heath O'Connell, Richard Dominiak, Patricia A. Kreitz and Louise Addis at [SLAC/SPIRES](http://slac.spires.org) for making their system SFX-aware
- Lieve Rottiers at the [Ghent Library Automation team](http://ghentlibraryautomation.com) for art work
- Jenny Walker at [SilverPlatter](http://silverplatter.com) for data-access

Many thanks for sponsoring and support:

- Deanna Marcum and Rebecca Graham at the [Council on Library & Information Resources](http://councilonlibraryandinformationresources.org) and the [Digital Library Federation](http://digitallibraryfederation.org)
- Donna Berg and Rick Luce at the [LANL Research Library](http://lanlresearchlibrary.org)
- Thomas Krichel, the WoPEc project of the JISC funded e-Lib program
- Richard Johnson and Alison Buckholtz at [SPARC](http://sparc.org) & [ARL](http://arl.org)

Herbert Van de Sompel wishes to thank the [Belgian Science Foundation](http://belgian-science-foundation.org) for a special PhD grant.

Appendix: The UPS Prototype demo system

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 screen dumps.

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:aarhec:1996-9. ups.cs.odu.edu is the server on which the SODA archives are stored. [Figure 4](#) & [Figure 7](#) also illustrate 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:

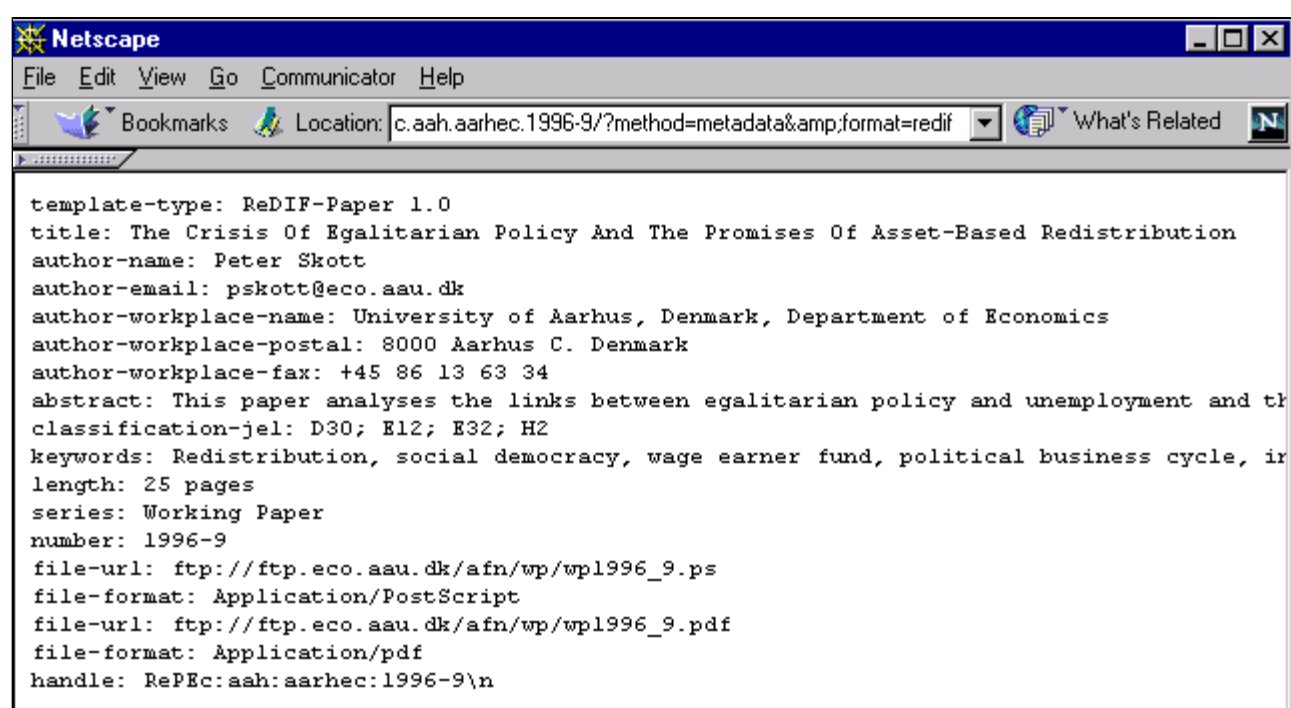


Figure 4: The bucket *metadata* method used with the ReDIF qualifier

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.

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:

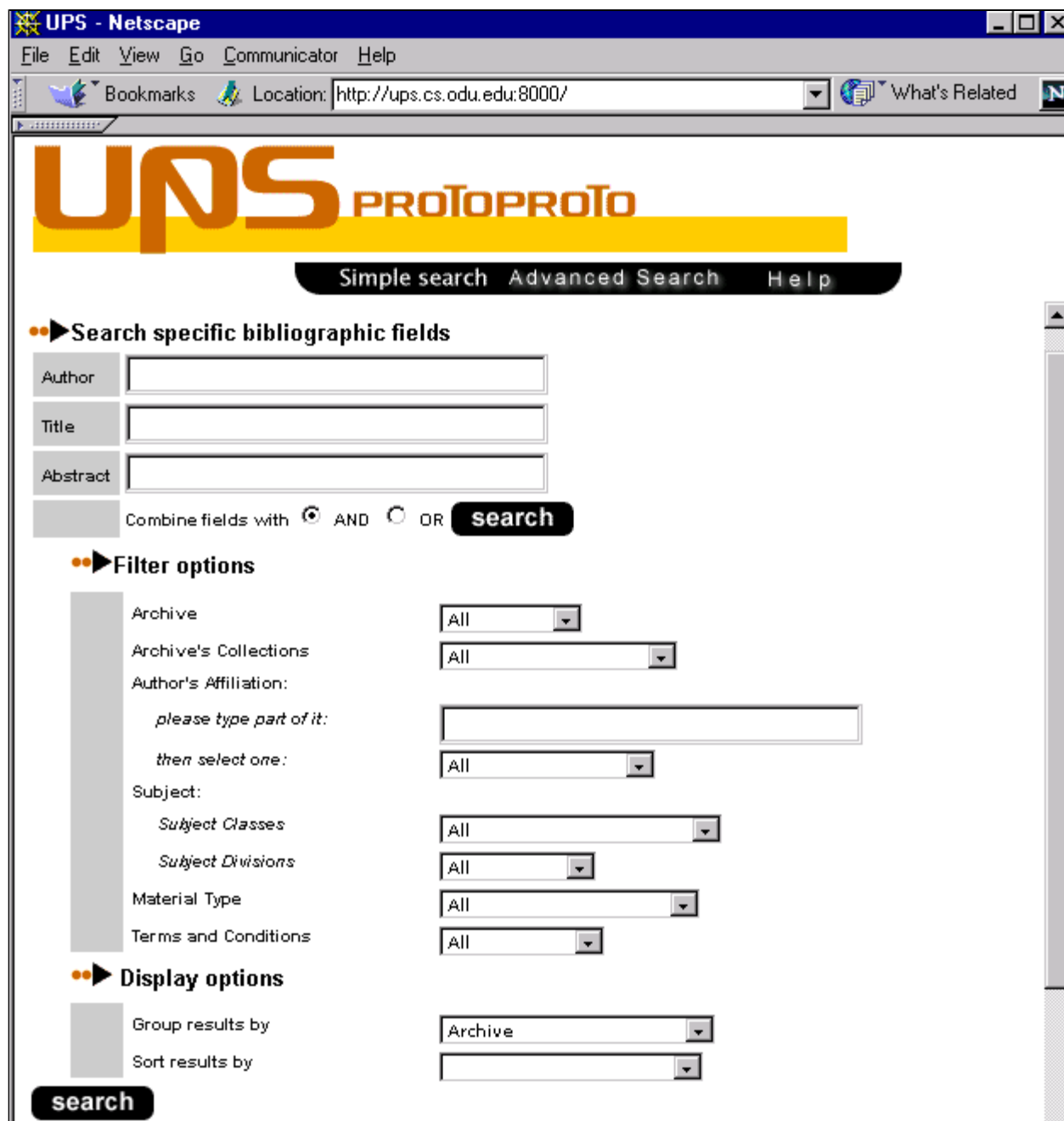


Figure 5: The advanced interface

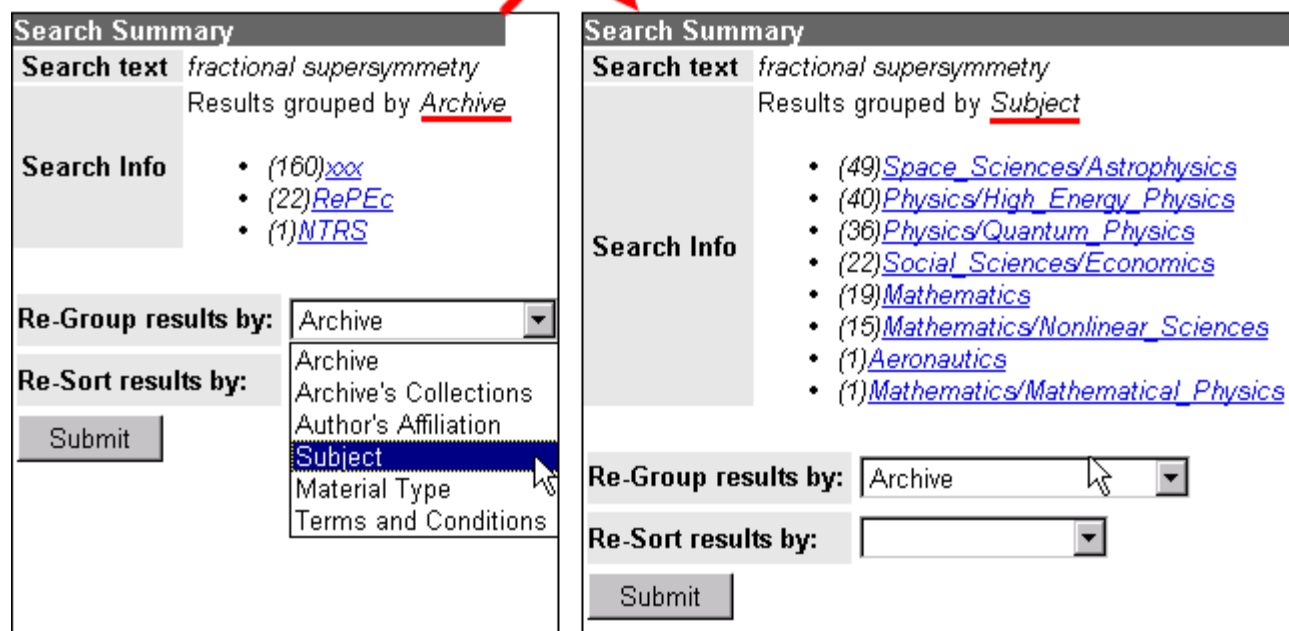


Figure 6: Reorganizing search results along another cluster

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 point 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.

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, which 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 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.

Screen dump example 3:

The user accesses a bucket originating from the ArXiv:he-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).



<i>archive:</i>	http://xxx.lanl.gov
<i>title:</i>	Experimental Search for Chargino and Neutralino Production in Supersymmetry Models with a Light Gravitino
<i>authors:</i>	Dzero Collaboration, B. Abbott and et al
<i>date:</i>	1997-08-04
<i>publication_status:</i>	Published in Phys.Rev.Lett. 80 (1998) 442-447
<i>number:</i>	Fermilab-Pub-97/273-E
<i>abstract:</i>	We search for inclusive high E_T diphoton events with large missing transverse energy in $p\bar{p}$ collisions at $\sqrt{s}=1.8$ TeV. Such events are expected from pair production of charginos and neutralinos within the framework of the minimal supersymmetric standard model with a light gravitino. No excess of events is observed. In that model, and assuming gaugino mass unification at the GUT scale, we obtain a 95% CL exclusion region in the supersymmetry parameter space and lower mass bounds of 150 GeV/c ² for the lightest chargino and 75 GeV/c ² for the lightest neutralino.
<i>ups-id:</i>	ups.xxx.hep-ex//xxx.xxx.hep-ex.9708005
<i>notes:</i>	11 pages, 4 eps figures, To be Published in Phys. Rev. Letters
<i>Publication:</i>	PDF version PostScript version (gunzipped)
<i>Services:</i>	 (pre-published version)  (published version)

Figure 7: The bucket for a record originating from arXiv:hep-ex

Bucket display is the result of the invocation of the default *display* method via the command

<http://ups.cs.odu.edu:8000/buckets/ups.xxx.hep-ex/xxx.xxx.hep-ex.9708005/>

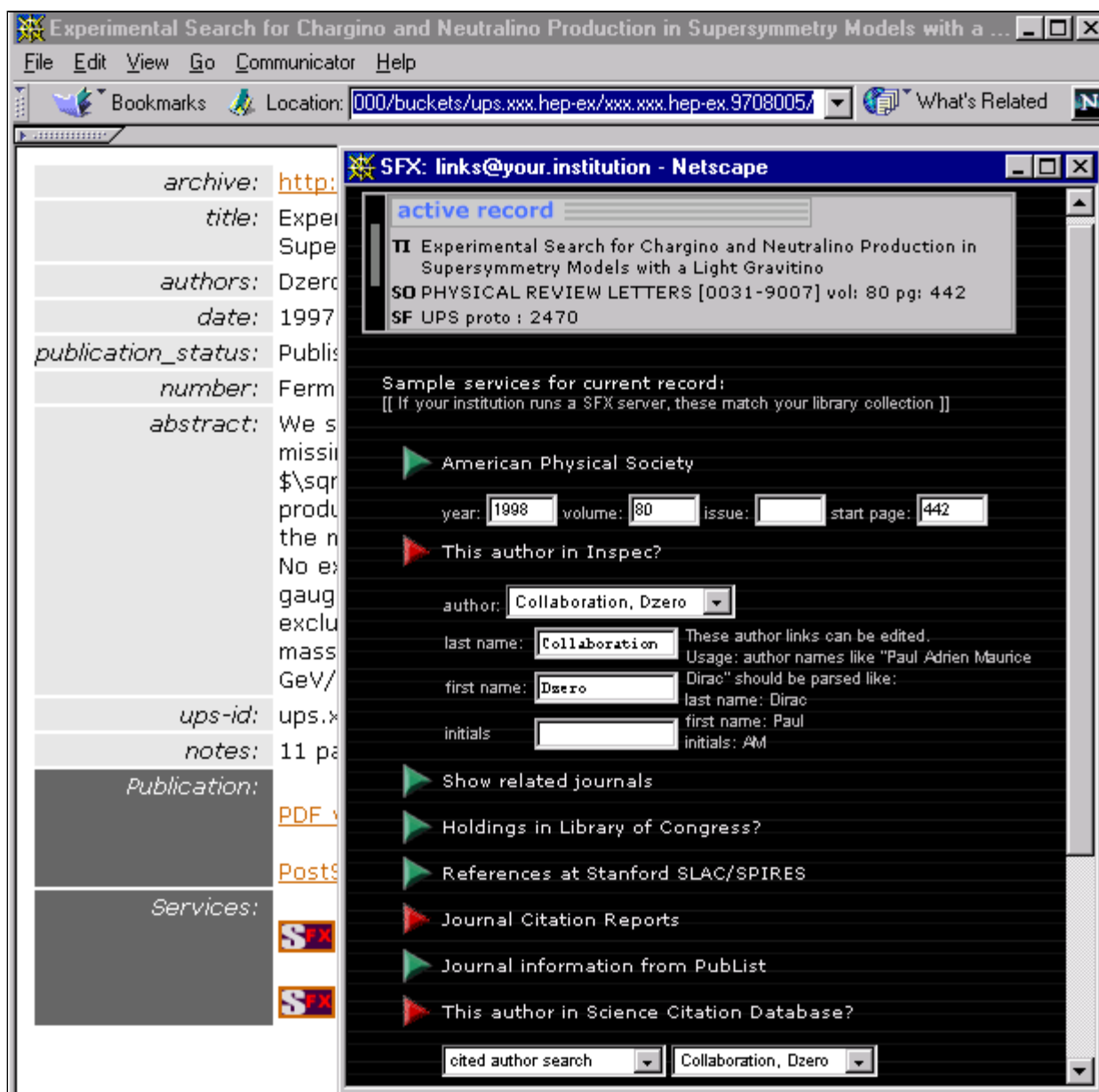


Figure 8: SFX-services for the published version of the paper described by the bucket in [Figure 7](#)

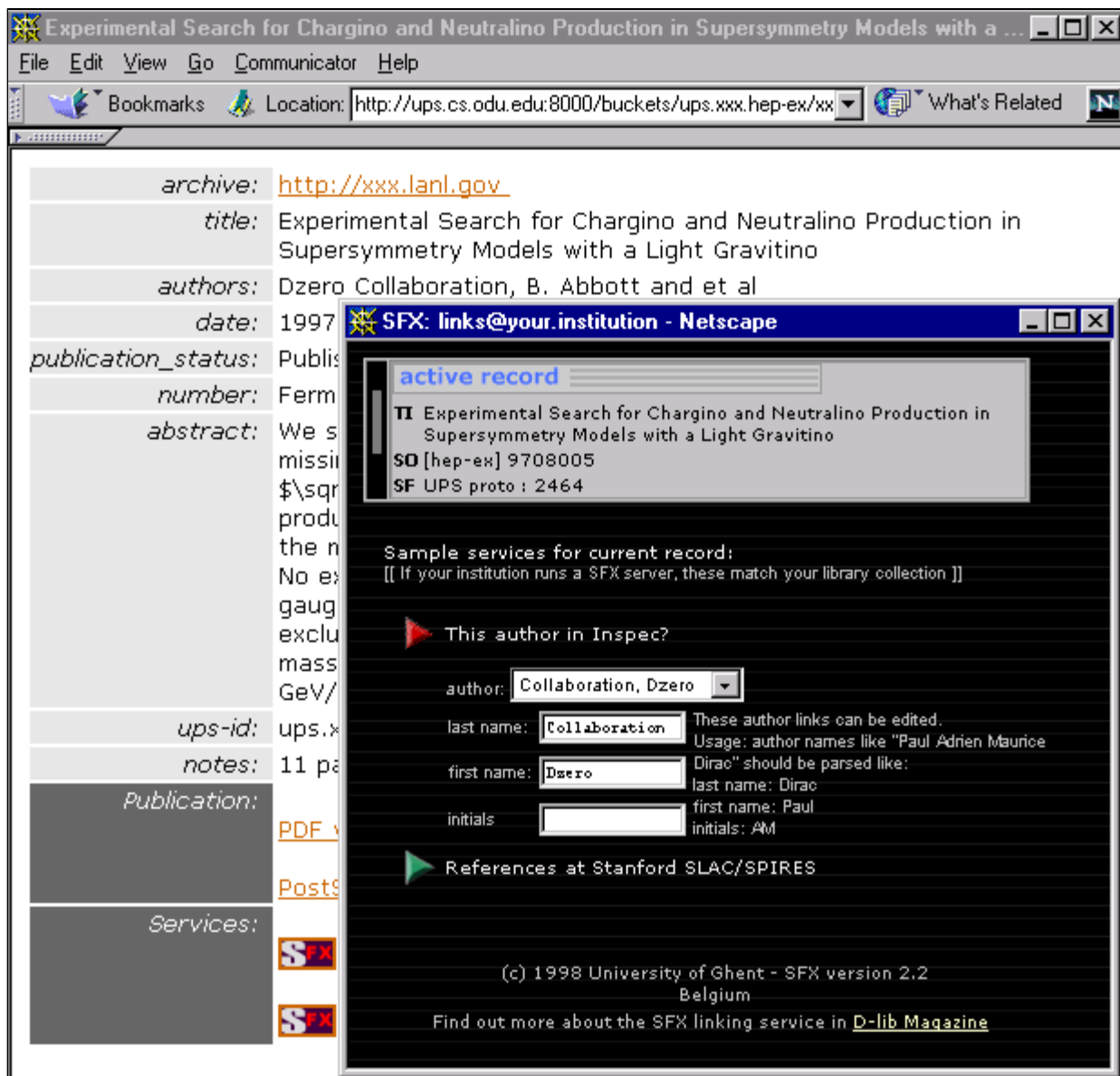


Figure 9: SFX-services for the e-print version of the paper described by the bucket in [Figure 7](#)

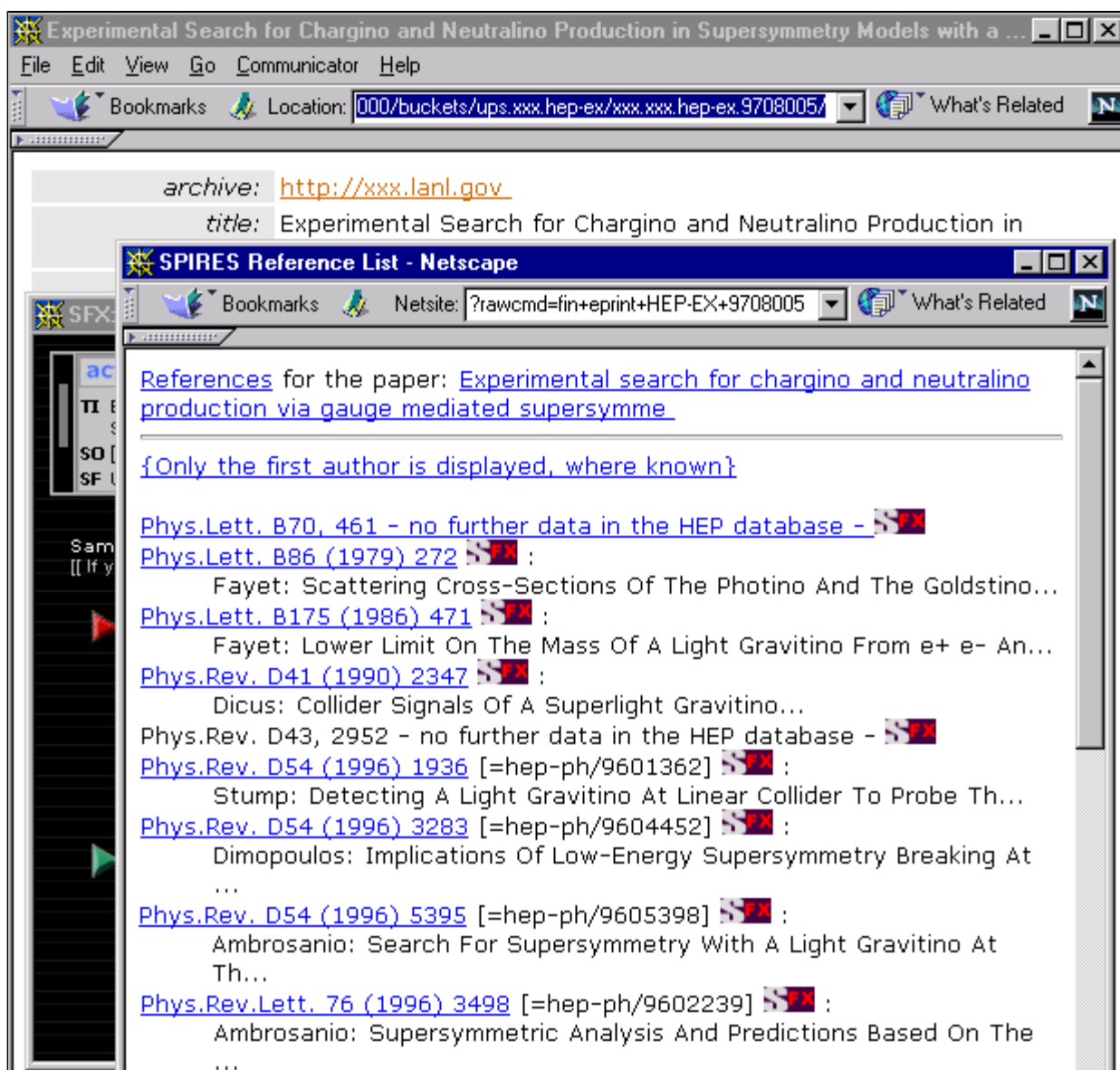


Figure 10: The user follows the *reference* link from the SFX-menu screen of [Figure 9](#)

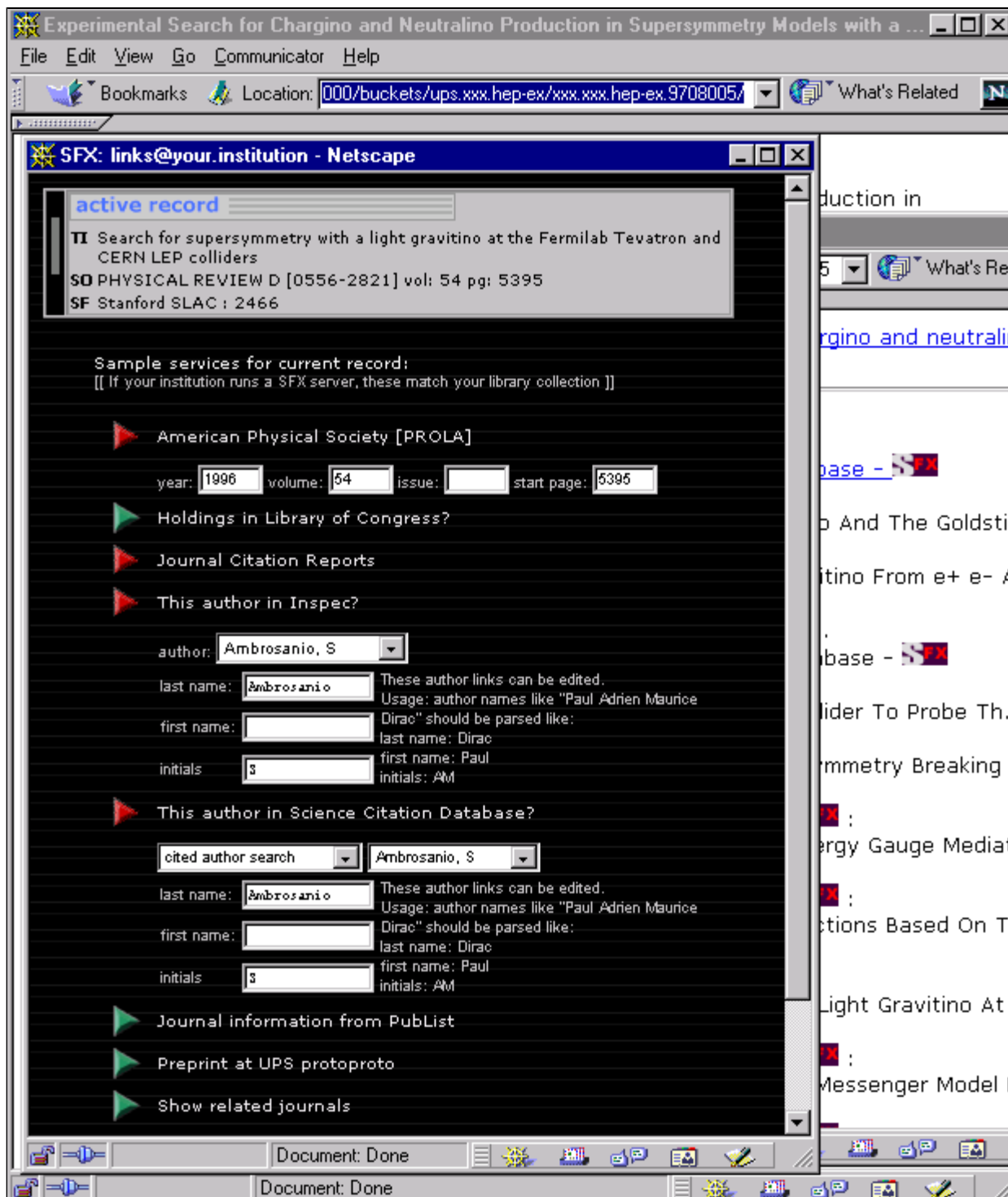


Figure 11: The user requests extended services for one of the citations from [Figure 10](#)

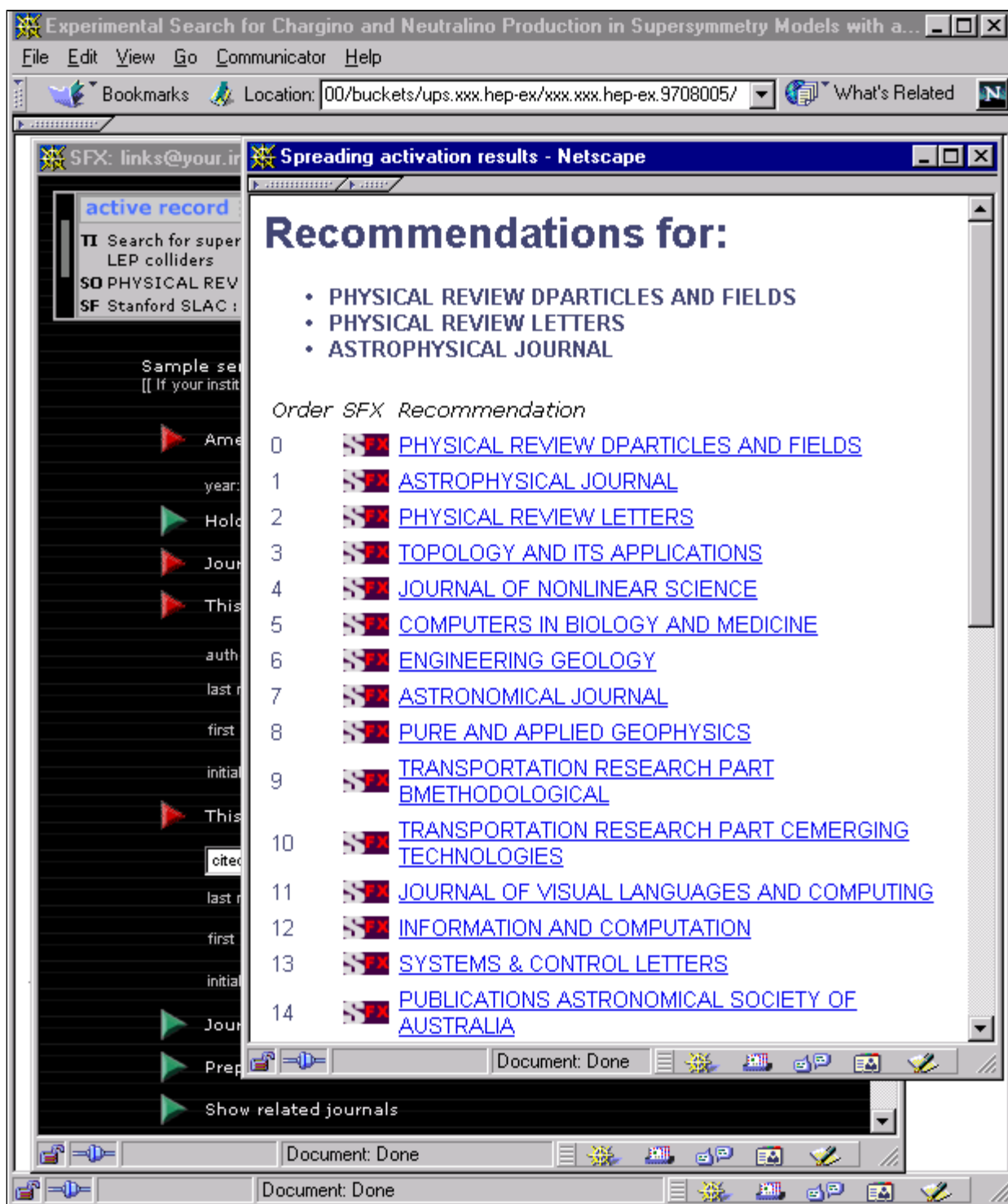


Figure 12: The user checks out the journal recommendations made by the SpreadIt! system by clicking the "Show related journals" menu item

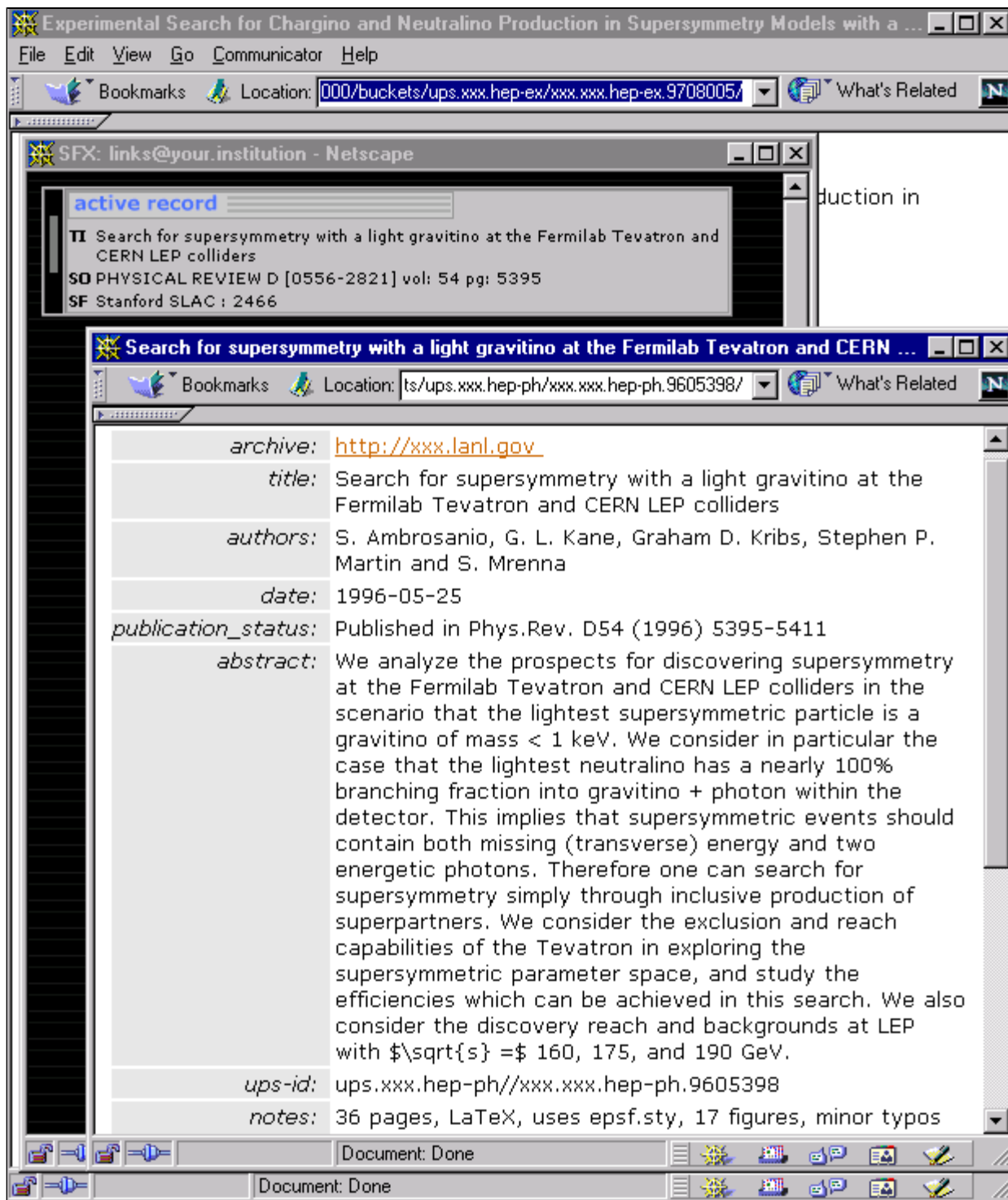


Figure 13: The user follows the Preprint link, which brings him back into the UPS environment

Copyright © 2000 Herbert Van de Sompel, Thomas Krichel, Patrick Hochstenbach, Victor M. Lyapunov, Kurt Maly, Mohammad Zubair, Mohamed Kholief, Xiaoming Liu, and Heath O'Connell. (Although he is a co-author of this story, Michael Nelson is not listed as a copyright holder because he is an employee of the U.S. Federal Government.)

[Top](#) | [Contents](#)
[Search](#) | [Author Index](#) | [Title Index](#) | [Monthly Issues](#)
[Previous story](#) | [Next Story](#)
[Home](#) | [E-mail the Editor](#)

[**D-Lib Magazine Access Terms and Conditions**](#)

[**DOI:**](#) 10.1045/february2000-vandesompel-ups