*Old Dominion University Computer Science Technical Report TR-2000-01, February 2000.*

# The UPS Prototype project: exploring the obstacles in creating a cross e-print archive end-user service

Herbert Van de Sompel
herbert.vandesompel@rug.ac.be
Los Alamos National Laboratory - Research Library, NM USA & Automation Department of the Central Library of the University of Ghent, Belgium

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

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

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

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

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

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

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

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

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

## Abstract

The Universal Preprint Service (UPS) Prototype was developed in preparation of the first meeting of the UPS initiative - later renamed the Open Archives initiative - held in Santa Fe, New Mexico October 21-22, 1999. The purpose of the meeting was to generate discussion and concensus regarding the interoperability of publicly available scholarly information archives. The invitees represented several renown e-print and report archive initiatives, as well as organizations with an interest 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, recommendation systems and linking systems - for data originating from distributed and dissimilar archives easier. To facilitate the discussions, the UPS Prototype was constructed as a proof-of-concept of a multi-discipline digital library of publicly available scholarly material. The UPS Prototype harvested its nearly 200,000 records from several different archives and created an attractive

end-user environment. As such, the UPS Prototype was a demonstration vehicle for other digital library technologies, such as intelligent digital objects (buckets) and reference linking services (SFX). The paper touches on their applicability in an e-print environment.

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

The main aim of the first 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, 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 UPS acronym refers to Universal Preprint Service, which was the original name of the Open Archives initiative. The project is 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 is 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. A companion paper (Van de Sompel, Krichel, Nelson et al 2000) describes the fundamental issues arising in the creation of the prototype, which are related to the nature of the e-print collections involved. That paper also describes the recommendations made by the project team to the Open Archives meeting. This paper concentrates on the issues related to the application of selected digital library technologies in the Prototype.

Herbert Van de Sompel 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. As will be shown, this project presentation had an important impact on 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 UPS Prototype project aimed to create the following 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 that 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 convincing. 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

Such complexity is already apparent in this UPS Prototype project since it sets out to create end-user services for data originating from some major archive initiatives: arXiv.org (commonly known as the xxx.lanl.gov archive), CogPrints, NACA, NCSTRL, NDLTD and 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 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 |
|---|---|---|---|---|---|---|
| | **C**entral<br>**D**ecentral | **A**uthor<br>**O**rganization | **C**entral<br>**D**ecentral<br>**M**irror | **C**entral<br>**D**ecentral | **Y**es<br>**N**o | **Y**es<br>**N**o |
| 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 the 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 |
|---|---|---|---|---|
| | | **criterion** | **documented** | **documented** |
| | **Y**es | **S**ubject | **Y**es | **Y**es |
| | **N**o | **D**ate | **N**o | **N**o |
| | | **A**ffiliation | | |
| 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 |
| **Total** | **235,361** | **191,622** | **187,115** | **112,516** | **22,844** |

**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

| 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**

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.

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 to that, 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: framos@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:9601001

**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 seperate 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)

**author field for a record from arXiv**

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

**ReDIF version of the author information after procedural conversion**

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

## Creation of a UPS namespace of unique identifers

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 1999b) 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 at 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 Manisfestation 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.

| | | | |
|---|---|---|---|
| **version:** | 1.1 | **Template-Type:** | ReDIF-Paper 1.0 |
| **authors:** | Crusio, Wim E. | **Author-Name:** | Crusio, Wim E. |
| **e-mail:** | crusio@citi2.fr | **Author-Email:** | crusio@citi2.fr |
| **catcode:** | bio.bio-socio | **Series:** | cogprintsbio |
| | | **Classification-Ila:** | Sociobiology |
| **abstract:** | 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. | **Abstract:** | 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. |
| **from:** | Wim E Crusio | **X-Cogprints-Submitter-Name:** | Wim E Crusio |
| **title:** | The sociopathy of sociobiology | **Title:** | The sociopathy of sociobiology |
| **userid:** | CrusioW | **X-Cogprints-Userid:** | CrusioW |
| **file-html-main:** | bbsmeal.htm | **File-URL:** | http://cogprints.soton.ac.uk/archives/bio/papers/199805/199805001/doc.html/bbsmeal.htm |
| | | **File-Format:** | text/html |
| | | **File-Function:** | Main file |
| **context:** | pjour | **Manifestation-Type:** | prar |
| **pages:** | 552 | **Manifestation-Pages:** | 552 |
| **pubn:** | Behavioral and Brain Sciences | **Manifestation-Journal-Title:** | Behavioral and Brain Sciences |
| **volume:** | 18 | **Manifestation-Journal-Volume:** | 18 |

| | | | |
|---|---|---|---|
| **year:** | 1995 | **Manifestation-Journal-Year:** | 1995 |
| **number:** | 3 | **Number:** | 199805001 |
| **idcode:** | bio/199805001 | **Handle:** | CogPrints:cogprints:cogprintsbio:199805001 |
| | | **Order-URL:** | http://cogprints.soton.ac.uk/abs/bio/199805001 |

**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-MSC-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 proofs 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 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 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.

### *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 search interface provides a simple and advanced search interface. 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.

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 problems 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 SFX context-sensitive 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 (Van de Sompel and Hochstenbach 1999c), 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 a interesting opportunity for further explorations of this problem domain.

#### The SFX framework

The details of the SFX framework have been described at length (Van de Sompel and Hochstenbach 1999a; Van de Sompel and Hochstenbach 1999b; Van de Sompel and Hochstenbach 1999c) and therefore only an overview of its functionality is given here. The fundamental aim of the SFX linking framework is to present an information entity of a digital library collection in the context of the complete collection. When an information system is interoperable with SFX, it presents each search result with an accompanying SFX-button. Systems that do this are referred to as SFX-aware. When the SFX-button of a search result - called the link-source - is clicked, a list of extended services that are relevant for the given link-source is dynamically generated. Taking a citation in a journal article as a link-source, such services can be a link to the corresponding full-text, a link to an A&I database that has an abstract for the citation, a cited-reference look-up into a Citation Database, a holdings look-up link to an Online Public Access (OPAC) system, etc. But, since SFX is context-sensitive, these 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. If his institution stores the full-text in a private repository the link will point there, not to the publisher's repository. The OPAC service will point at the OPAC of the user's institution. And, presuming the A&I database where an abstract can be found is the BIOSIS database, the service link will point into the implementation of the user's institution, which could be an Ovid, SilverPlatter or whatever version of BIOSIS. In order to achieve this, the SFX linking framework builds on an architecture with two independent components:

- A redirection mechanism that - when the SFX-button is clicked - transports the link-source from its origin to the doorstep of an institutional service component;
- An institutional service component that will use a rule-based approach to deliver context-sensitive services to the user who has clicked the SFX-button. This rule-based approach will take into account the origin database of the link-source, the link-source metadata and a database of potential services subject to boundary conditions as input for this dynamic decision making process.

#### SFX-awareness and buckets

In the SFX experiments that have been conducted previous to the UPS Prototype project, 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 service component;

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.

### 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. Doing so would open up an important research problem regarding the interpretation of subject information and its representation in a rules-based system such as SFX. This clearly is far beyond the aims of the UPS Prototype project.

## The SFX services in the UPS Prototype project

Even given the above restrictions, the SFX linking system introduced for the UPS prototype presents some 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 of the SFX-research, 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 | *author* | Inspec |
| arXiv:hep-th | *reference* | SLAC/SPIRES |
| arXiv:math | *author* | MathSci |
| NCSTRL | *author* | Inspec |
| RePEc | *author* | EconLit |
| RePEc | *author* | ABI/Inform |

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

Unfortunately, the problems regarding the - lack of - metadata quality, is a serious 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 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 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 them. Since such references are both to 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 (Van de Sompel and Hochstenbach 1999c) 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 solution. In those cases where parsing is successful, again, an attractive integration of the e-print environment and the traditional scholarly communication environment results.

## 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 of (Van de Sompel, Krichel, Nelson et al 2000) by means of Lotus Screencam movies and screendumps that show how a user navigates the multidisciplinary UPS Prototype environment.

The section "The recommendations made to the Open Archives group" of (Van de Sompel, Krichel, Nelson et al 2000), shows the recommendations that the coordinating trio presented to the Open Archives group at the Santa Fe meetings, as a result of the insights that were gained regarding the creation of a cross e-print end-user service.

The project also identified several issues related to the technologies that have been used during the UPS Prototype project. A summary of those issues is given as a conclusion of this paper.

## Conclusion

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 have been translated into 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. A careful reading of the Santa Fe Convention, that is the result of the discussions at that meeting, will reveal that the Open Archives group has adopted several elements of these recommendations.

The UPS Prototype Project also resulted in interesting new insights regarding the digital library technologies that have been used. Scaling problems arose from the use of the NCSTRL+ environment. One was caused by a conflict between Dienst, requiring an archive to reside under a single directory, and the Solaris operating system, allowing a maximum of 32,767 directories under a single directory. A workaround has been adopted, but the core of the problem remains unsolved. Another was related to the Dienst search engine that does not scale beyond approximately 20,000 records. While the alternative engine - freeWAIS-sf - can index beyond

this figure, it has other limitations that seriously decrease the effectiveness of searching. Lessons have been learned and appropriate actions have already been taken by means of the introduction of an Oracle-based storage system that should lead towards a resolution of these issues.

It has not been possible to illustrate many of the interesting features of the SODA architecture in the project. Still, a proof of concept is obtained regarding the possibility to deploy buckets in a large-scale digital library. The project showed that further research was required to minimize the footprints of buckets. Again, actions have already been taken to address this issue in new bucket versions. Since it is essential to retain the ability to have all the necessary code for a bucket to function by itself, the optimization strategy is proceeding along the lines of factoring code from buckets that are contained within an archive but leaving the code intact when they are mostly standalone objects.

The storage of two concurrent metadata representations in the UPS buckets was beneficial for the creation of the end-user services. It is also seen as an interesting illustration of their aggregative capabilities and the importance thereof. The ease with which services like SFX were attached to buckets is encouraging in light of hooks for other anticipated services such as the Bucket Matching Service, as part of the Bucket Communication Space (Nelson et al. 1999). The self-contained aspect of buckets will ensure that such services will remain available in the buckets, regardless of the service provider accessing the buckets and the location of the bucket.

The decreased performance of the SFX linking service, caused by the lack of metadata quality could have been foretold, since SFX depends on metadata to function. In cases where the metadata quality was decent, the service confirmed the flexibility and attractiveness that had already been demonstrated in other experiments. The project has demonstrated that SFX can be a valuable tool to integrate the e-print environment with the established scholarly communication mechanism. Because of this quality, SFX may find its way into the JISC/NSF OpCit project (Harnad 1999) that actually aims at such integration. The project has also identified a new research area related to SFX by concluding that the actual solution is not fit to deal with data originating from multi-disciplinary 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/].

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].

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]

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].

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.

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 April . [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-ieeedl-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].

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

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

Paskin, Norman. 1999b. DOIs used for reference linking. Washington & Geneva. [http://www.doi.org].

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].

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/].

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].

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].

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].

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: an experimental end-user service across e-print archives. D-Lib Magazine 6, no. 2.[http://www.dlib.org/dlib/february00/vandesompel/02vandesompel-ups.html].

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

## Acknowledgments

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

- Paul Ginsparg - arXiv.org archive
- Stevan Harnad and Robert Tansley - CogPrints archive
- Michael L. Nelson - NACA collection
- Carl Lagoze - NCSTRL initiative
- Ed Fox, Anthony Atkins - NDLTD initiative
- Thomas Krichel - RePEc 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 for making their system SFX-aware
- Lieve Rottiers at the Ghent Library Automation team for arty work
- Jenny Walker at SilverPlatter for data-access

Many thanks for sponsoring and support:

- Deanna Marcum and Rebecca Graham at the Council on Library & Information Resources and the Digital Library Federation
- Donna Berg and Rick Luce at the LANL Research Library
- Thomas Krichel, the WoPEc project of the JISC funded e-Lib program
- Richard Johnson and Alison Buckholtz at SPARC & ARL

Herbert Van de Sompel wishes to thank the Belgian Science Foundation for a special PhD grant