

To: ASA and IMS Boards of Directors
From: Duncan Murdoch, Chair, CIS Management Committee
Re: Reappointment of Abstracting Editor; Budget for CIS
Date: May 24, 2006

Status report

The Current Index to Statistics is currently housed at Carnegie Mellon University (with David James as Database Editor). Operations are running smoothly, and journal article updates are now occurring approximately quarterly. David is planning an effort over the summer to bring our book entries up to date, linking books to their reviews.

Distribution of bibliographic information has changed drastically over the lifetime of CIS. Recently it has become the norm for researchers to obtain copies of articles online, rather than by photocopy from a library. Currently CIS has no automatic links to online content. We had planned to add this over the summer of 2005, but this did not happen. Instead our efforts went towards a proposal to move CIS hosting to Cornell University Libraries. Unfortunately, this turned out to be much more expensive than we had anticipated, and the plan has been shelved. We are currently planning more modest proposals for contracting out particular aspects of the work.

In 2004, the MC agreed to allow free online access for ASA and IMS members. This is now in place from both the ASA and IMS web sites. As expected, we have seen a substantial drop in personal subscriptions due to this change, with a corresponding drop in revenue. However, costs have been lower than expected, so the net result is an increased surplus in 2005 over 2004. We expect 2006 will come in approximately on budget, i.e. with revenue and expenses approximately balanced.

In last year's budget, we planned for an open competition for web development proposals. We carried out this competition, but only had 3 submissions. We do not plan to hold a similar competition in 2007. A full report on the competition is attached.

Reappointment of Abstracting Editor

The term of appointment for our Abstracting Editor, George Styan, is due to expire on December 31, 2006. Styan has expressed a willingness to extend his term by one more year, and the MC has agreed to nominate him for an additional year (i.e. to December 31, 2007). A motion to this effect is attached.

Search for Database Editor

The term of appointment for our Database Editor, David James, is also due to expire on December 31, 2006. We are in the process of searching for a new editor.

Medium to Long Range Plans

It is clear that CIS must add features in order to remain useful. It is no longer sufficient to give users citations of papers and expect them to find the papers themselves; we need to provide means to get to the papers, as competing bibliographic services do.

Fortunately librarians and others have developed standard schemes to facilitate this. CIS can construct simple links to brokers who will redirect them to full text versions of articles for users who are subscribers to the original journal. We are currently in the process of arranging a contract to carry out this work.

CIS will also become more relevant to its users as a trusted reference source. We are planning to implement unique article identifiers, so authors can cite a link to CIS which will give readers access to all the CIS information on an article.

These and similar initiatives require time and energy to implement. We are planning to contract out some of the development work: it is too much to expect volunteers to implement. We expect that these improvements to CIS will make it more attractive to libraries and commercial subscribers. Potentially, its market is much larger than the current subscription base: doubling or tripling our library subscriptions would be an achievable target in 3-5 years, if we offer more value to our users. On the other hand, maintaining the status quo does not appear to be sustainable in the long term.

2007 Budget Proposal

In putting together the 2007 budget, we considered the following on the revenue side:

- We are concerned about competition from free services such as Google Scholar, and our own free access to ASA and IMS members, so we have not raised subscription rates.
- The forecast numbers of subscriptions were chosen to be between the 2005 actual and 2006 budgeted numbers. We expect increases in the long term as we implement and advertize new features such as linking to online content, but are not counting on this increase to be reflected in 2007 revenue.

On the expenses side:

- We have kept the Abstracting Editor's budget fixed at \$30,000. As predicted last year, the Canadian dollar has risen relative to the US dollar; our budget assumes there will not be further increases through 2007.
- We have added \$10,000 to the Database Editor's budget. The 2006 budget of \$25,000 included \$10,000 for the web development project competition which we do not plan to hold in 2007, so this actually represents a \$20,000 increase in the money available to the Database Editor. We expect to use this money to contract out specific web development projects: linking to external content, etc.

- We have deleted the budget item of \$1500 in licensing fees. It turned out that CIS is covered under ASA subscriptions for the services we needed.
- The CIS admin assistant at Harvard has been budgeted at the same amount as in 2006.
- We have increased the computing equipment budget from \$1000 to \$3000. We expect that the new Database Editor will need to purchase equipment to support the operations.
- Other items have been budgeted the same as 2006.

Approvals Requested

On behalf of the MC, I would like to request that the Boards of the ASA and the IMS approve the budget and the extension of the Abstracting Editor's term.

Respectfully submitted by Duncan Murdoch, Chair, CIS Management Committee

Encl: Budget spreadsheet
Report on web development competition

CURRENT INDEX TO STATISTICS

DRAFT 2007 Budget

Calculation of 2007 revenue projections

| | 2004 | 2005 | 2006 | 2007 | 2005 | 2006 | Calculation of 2007 revenue projections | | | |
|---------------------------------------|------------------|--------------------------------------|-----------------|------------------|--------------|--------------|---|--------------|--------------|--------------|
| | <u>Actual</u> | <u>Actual (adjusted)¹</u> | <u>Budget</u> | <u>Budget</u> | <u>Sales</u> | <u>Sales</u> | <u>2007</u> | <u>2005</u> | <u>2006</u> | <u>2007</u> |
| | | | | | Actual | Budgeted | Budgeted | <u>Unit</u> | <u>Unit</u> | <u>Unit</u> |
| | | | | | | | | <u>price</u> | <u>price</u> | <u>price</u> |
| REVENUE | | | | | | | | | | |
| Sales - General | \$20 | \$200 | \$- | \$- | | | | | | |
| Sales - CIS ED 4 Station License | 1,000 | - | - | - | | | | | | |
| Sales - CIS ED 4 Station Upgrade | 625 | - | - | - | | | | | | |
| Sales - CIS ED Personal License | 4,635 | - | - | - | | | | | | |
| Sales - CIS ED Personal Upgrade | 1,035 | - | - | - | | | | | | |
| Sales - CIS ED Commercial Upgrade | 2,160 | - | - | - | | | | | | |
| Sales - CIS ED University License | 5,400 | - | - | - | | | | | | |
| Sales - CIS ED University Upgrade | 1,354 | - | - | - | | | | | | |
| CIS Web - Personal | - | - | 1,950 | 650 | - | 30 | 10 | 65 | 65 | 65 |
| CIS Web - Commercial | 32,311 | 29,597 | 24,000 | 29,520 | 41 | 33 | 41 | 720 | 720 | 720 |
| CIS Web - General | 11,976 | 10,924 | 13,200 | 13,200 | 27 | 33 | 33 | 400 | 400 | 400 |
| CIS Web - University | 63,589 | 62,709 | 54,000 | 59,940 | 232 | 200 | 222 | 270 | 270 | 270 |
| CIS Web - 4 Station | 5,217 | 4,067 | 4,725 | 4,750 | 33 | 38 | 38 | 125 | 125 | 125 |
| Royalties | 93 | 6 | - | - | | | | | | |
| Interest - Other | 853 | 1,428 | - | - | | | | | | |
| Total Revenue | \$130,268 | \$108,929 | \$97,875 | \$108,060 | | | | | | |
| EXPENSES | | | | | | | | | | |
| CIS Abstracting Editor (Styan) | \$26,000 | \$28,000 | \$30,000 | \$30,000 | | | | | | |
| CIS Database Editor (New editor) | 1,696 | 6,071 | 25,000 | 35,000 | | | | | | |
| Database licensing fees | - | - | 1,500 | - | | | | | | |
| Computing Equipment (New editor) | 2,321 | - | 1,000 | 3,000 | | | | | | |
| Subtotal - program | <u>\$30,017</u> | <u>\$34,071</u> | <u>\$57,500</u> | <u>\$68,000</u> | | | | | | |
| CIS Admin Assistant (Harvard Med Sch) | \$13,728 | \$10,975 | \$14,500 | \$14,500 | | | | | | |
| Software Support (Carnegie Mellon) | - | 541 | - | - | | | | | | |
| Postage / Shipping Charges | 67 | 82 | 200 | 100 | | | | | | |
| Telephone | 25 | 0 | 200 | 500 | | | | | | |
| Storage/warehouse | - | - | 500 | - | | | | | | |
| Audit/Accounting Services | 4,700 | 730 | 5,000 | 5,000 | | | | | | |
| Bank Charges | 955 | 735 | 1,000 | 1,000 | | | | | | |
| Bank Credit Card Fees | 2,611 | 2,165 | 3,000 | 3,000 | | | | | | |
| Publications Management Committee | 482 | - | 2,500 | 2,500 | | | | | | |
| Awards/Plaques (for retiring editor) | 68 | - | - | 100 | | | | | | |
| Marketing/Advertising | - | - | 400 | 500 | | | | | | |
| ASA Management Costs | <u>20,700</u> | <u>9,000</u> | <u>12,500</u> | <u>12,500</u> | | | | | | |
| Subtotal - administrative | <u>\$43,336</u> | <u>\$24,228</u> | <u>\$39,800</u> | <u>\$39,700</u> | | | | | | |
| Total Expenses | \$73,353 | \$58,298 | \$97,300 | \$107,700 | | | | | | |
| NET REVENUE (EXPENSE) | \$56,914 | \$50,631 | \$575 | \$360 | | | | | | |
| ASA share | 50% | \$28,457 | \$25,315 | \$288 | \$180 | | | | | |
| IMS share | 50% | 28,457 | 25,315 | 288 | 180 | | | | | |
| | | <u>\$56,914</u> | <u>\$50,631</u> | <u>\$575</u> | <u>\$360</u> | | | | | |

¹Adjusted to include \$10,927 FY2005 reimbursement request for CIS admin assistant received after books closed. This expense will be reported in the 2006 books, but is reflected in the 2005 column of this spreadsheet for clarity. Also adjusted by DJM to re-label funding for database assistant Paul Juska as a Database Editor expense rather than an Admin Assistant expense, and to delete items with no expense historically or expected.

Results of CIS call for web development proposals

The CIS Management Committee

April 21, 2006

In its budget submission for fiscal 2006, CIS included \$10,000 to fund a competition for web development projects. The RFP for these went out in January 2006, with a deadline for submissions of April 1, 2006. The goal of the program was to add value to CIS. Projects were to be judged by the CIS Management Committee on the criteria:

- Value to CIS of the likely outcome from the project.
- Clarity of objectives, and likelihood of success.
- Matching funds from other sources.

In all, three proposals were received, including one from one of the members of the CIS Management Committee (Jim Pitman). The other two were from Valentinas Kriauciukis and Sigitis Tolusis of VTEX, and Andrew McCallum of the University of Massachusetts. The proposals are attached to this report.

Prof. Pitman excused himself from the discussion of the merits of the proposals. The other four members of the Management Committee (Duncan Murdoch, as chair; Ed Gbur, Jim Gentle for the ASA; John Wierman for IMS) read the proposals. Murdoch discussed them by telephone with Wierman and Gbur, and received comments by email from Gentle.

All three proposals appeared to meet the criteria set by the Management Committee. We ranked the Pitman and McCallum proposals ahead of the VTEX proposal based mainly on two considerations: they were both valuable contributions unlikely to be pursued by CIS without this funding, and their

budgets requested \$5,000 each, allowing both to be funded within the \$10,000 budget for the program.

The VTEX proposal will provide a core service that CIS has identified as being a high priority, namely providing links from CIS results to other sources of information on the Internet about each article: full text, Math Reviews entries, etc. We recommend that it be pursued using other CIS funds. The full CIS Management Committee should decide how best to proceed: whether to use 2006 funds or budget it for 2007.

CIS Proposal: BibServer development

Jim Pitman

March 31, 2006

1 Background

BibServer is a Python program designed to create a network of displays of bibliographic data maintained in BibTeX or logically equivalent format (XML, YAML, ...) by contributing authors and editors. Current code is capable of providing multifaceted displays over all many kinds of bibliographic data, including navigation over sets of people as well as varied sets of bibliographic items. Following are some sample pages:

- Author Index
<http://bibserver.berkeley.edu/cgi-bin/bibserver>
- Individual author listing with search facility
http://bibserver.berkeley.edu/cgi-bin/main?au=UCB_MATH:74
- UC Berkeley Mathematics Faculty Listing
http://bibserver.berkeley.edu/cgi-bin/authors?&Source=UCB_MATH

Proposer is collaborating with a number of others, including staff at VTEX, and Nitin Borwankar, web applications developer, to modularize the bibserver code in an appropriate web application architecture for use by numerous individuals and departments.

Once the code base has stabilized and been made portable, VTEX has agreed to maintain it and use it to support the IMS Bio-bibliographies Project, which was recently approved by IMS Council. As a benefit of membership, IMS will provide technical assistance to encourage its members to create and maintain machine-readable bio-bibliographies to be used for a number of purposes:

1. for members to provide high quality easily maintained web-renderings of their biobib data, such as the bibliographies provided by BibServer
2. to assist automated acquisition of bibliographic content by Current Index of Statistics, in particular for CIS to provide author name authority in its bibliographic records (Math Reviews already does this, and has agreed in principle allow extension of their identifiers to CIS)

3. to assist editing and presentation of biobib data in memorial volumes and festschrifts
4. to assist in creation of an attractive online compilation of collected works of IMS members, such as the Dutch Cream of Science Project

In the first instance, this will be done for IMS Fellows, then expanded to broader membership. Specifically, VTEX will prepare suitable latex/bibtex/xml templates for management of biobib data, so these templates can be used routinely in preparation of memorial volumes and festschrifts. As a benefit of membership, IMS will provide a service to its members whereby VTEX would convert their biobib data from legacy formats into the latex/bibtex/xml standard.

2 Proposal

It is proposed to continue work on modularization of the bibserver code before transfer of the code base to VTEX to manage. Funds are requested to support this programming effort.

- Budget: \$5,000 to support coding work by Nitin Borwankar.
- Timeline: BibServer code base should be transferred to VTEX for use in the IMS biobibs project by June 2006.
- Deliverables: Properly modularized open source python code with typical MVC web application architecture capable of providing current bibserver functionality over a suitably configured MySQL database.
- Value to CIS of the outcome: Individuals and departments will have adequate software, either to manage themselves, or to access by a VTEX operated webservice, to serve their own bibliographic data to the web. Format of this data should be adequately structured to allow CIS to access this data in an automated way, with the option of adding data in the index in a semi-automated way, subject to approval by editors.
- Clarity of objectives, likelihood of success: This proposal a small component of the broader program of bibserver development, for which there is substantial support in the community, as evidenced by IMS and the Departments of Mathematics and Statistics at U.C. Berkeley.
- Matching funds from other sources: \$5,000 towards coding work has already been provided by IMS, which is planning to support VTEX work on the IMS biobibs project. It is expected that further support of customization of bibserver for use by U.C. Berkeley Mathematics department faculty and staff will be provided by the Berkeley Center for Pure and Applied Mathematics, likely around \$15,000. This will provide the administrative interface to allow faculty and staff to update their bibserver entries.

3 Followup

It is hoped that CIS will make a long-term commitment to data exchange with bibserver installations managed by various departments and other organizations with bibliographic data to share, by provision of some financial support to maintain the bibserver infrastructure. A start should be provided by cooperation between CIS and IMS/VTEX on data standards for the IMS biobibs project. CIS has already agreed to allow its data to be used for IMS biobibs displays. The present proposal is a step towards facilitating data transfer in the other direction, to allow CIS to import data from sources maintained by numerous individuals and groups using bibserver software to manage their data.

CIS enhancement with linkserver

Proposal for CIS Management Committee

Valentinas Kriaučiukas and Sigitas Tolušis
(valius@vtex.lt, sigitas@vtex.lt)

VTEX

March 29, 2006

1 Introduction

This is a proposal for the Current Index to Statistics (CIS) following Call for Proposals of the CIS Management Committee.

We propose to enrich CIS database entries with multiple links through *linkserver*, a new kind of web service, something like presented in the demo. The *linkserver* transforms one permanent standard link from a CIS item to the updatable corresponding links to MathSciNet, arXiv, Zentralblatt Math, CrossRef, etc. It allows quick and direct access to the repositories containing full text, abstracts, reviews of articles. New targets can be added in future without change of the parent link from CIS.

2 Requirements

1. The linkserver accepts SICI identifiers (other forms of identifiers can be added in the future).
2. The linkserver expands to links stored in its database (more functionality can be added in the future).
3. The linkserver database will be build for all bibliographical entries of CIS for which the SICI identifier can be constructed.
4. The linkserver will be programmed in Python.
5. The linkserver runs on VTEX supported server (mirroring in the CIS site can be added in the future).

3 Timeline

Step 1. Initial analysis: the agreements on the format of the queries to linkserver, methods to fill the database [2 month].

Step 2. Programming of the linkserver functionality [2 month].

Step 3. Building the linkserver database [2 month].

Step 4. Final launch and final adjustments [1 month].

4 Budget

Step 1: 1320\$.

Step 2: 3960\$.

Step 3: 2640\$.

Step 4: 880\$.

Total: 8800\$.

5 Future enhancements

The linkserver can be used to add more functionality in the future. The natural extension, now missing in CIS, is to have *publicly accessible identifiers* for all items in CIS. This would allow external links to them from external web resources (like linkservers). Then the authors of the publications could be able to link to the corresponding entries in CIS and so provide links (through CIS and the linkserver) to items related to their publications.

In fact, the identifiers should be visible as active web links next to corresponding bibliographical items, when those are shown in screens exposing search results. Clicking on the identifier in the browser, the page of the bibliographical item could be exposed. This page can show at least the item and links provided by the linkserver, but here is easy to add more activity and more content in the future, like an abstract, reviews, bibserver, etc.

CIS Proposal for Web Development 2006: Cross Referencing between CIS and Related Fields and Services

Andrew McCallum
mccallum@cs.umass.edu
<http://www.cs.umass.edu/~mccallum>

1 Summary

We propose to obtain a large collection of bibliographic data from fields neighboring statistics, make it available to CIS and the IMS Biobibs Project in structured form, and perform cross-referencing (name authority) sufficient to integrate it with existing CIS databases. The new data will be obtained by extensive Web spidering for research articles in PDF and Postscript, as well as our existing spidering relationship with the Cornell arXiv.

Whenever we are able to obtain the full-text of these new articles, we will also extract references from the end of the article, extract the separate fields for title, authors, journal, year, pages, etc, and perform reference linking—enabling us to provide lists of articles that cite articles in the CIS or IMS Biobibs.

Furthermore, in a limited number of cases (about half), we will be able to provide homepage URLs for those authors that have them, as well as their email addresses.

We expect to provide about 100,000 papers and 100,000 authors, and potentially much more.

Finally, we will also create links from our own CiteSeer-like computer science research paper search engine into CIS and IMS Biobibs—guiding more Web traffic to statindex.org and imstat.org.

The work will be done at the University of Massachusetts, however, we will coordinate with VTEX for input of data to the IMS Biobibs Project and/or CIS, and for long term maintenance needs to be determined by IMS/CIS.

2 PI Qualifications and the Rexa Project

Associate Professor McCallum directs the Information Extraction and Synthesis Laboratory (IESL) at the University of Massachusetts Department of Computer Science, with 8 PhD students, 2 postdoctoral fellows, and 2 full-time staff software engineers. He is the author of more than 50 research papers on machine learning and natural language processing, and is on the editorial board of the Journal of Machine Learning Research. Prior to his arrival at University of Massachusetts he was Vice President of Research and Development at a 170-person start-up company that used machine learning methods to perform information extraction from the Web.

The project proposed here is enabled by leveraging substantial software infrastructure and bibliometric extraction research being funded by a \$2.6 million NSF ITR project within IESL. The goal of this NSF project is (1) to fund new research in statistical natural language processing targeted at bibliometric data and joint inference across the multiple steps of a language processing and data mining pipeline, and (2) to create an enhanced alternative to CiteSeer and Google Scholar. The Rexa project was begun in 2003; the original NSF ITR goes through 2007; the PI plans to continue the project indefinitely, and has already obtained follow-on supporting grants from NSF and elsewhere.

The new service is called Rexa, and is available at <http://rexa.info>. It currently requires an invitation and password for access. These can be obtained by contacting McCallum.

The core of the research in the NSF project is to intimately integrate information extraction (IE) and data mining—enabling newer, more accurate, more ambitious mining of distributed unstructured textual data sources. Such efforts historically usually have limited success because IE is never perfectly accurate, and data mining is brittle. Our research uses robust probabilistic models, confidence estimation, and joint inference across IE and data mining so that uncertainty and multiple IE hypotheses can be used to improve mining (“bottom-up”), and emerging mined patterns can be used to improve IE (“top-down”). We are making this approach practical in real-world systems—including research in new probabilistic inference and parameter estimation methods along the way.

The core of the application in the NSF project is building an enhanced research paper search engine for computer science—an enhanced alternative to Google Scholar, CiteSeer, ACM Portal, etc.¹ There are many improvements (including better inter-

¹The PI believes that—just as it is important that we have many national newspapers, not just one—it is very important that there be multiple research paper search services. Each has different strengths, weaknesses, biases and emphases. So I see these other services as colleagues, not competitors. For example, the PI has co-authored grants with Paul Ginsparg, and stay in contact with Lee Giles.

face, higher accuracy, tagging, etc) but the main enhancement is that Rexa will know about many more first-class, de-duplicated, cross-referenced object types: not only papers and their citation links, but also people, grants, universities, conferences, journals, research groups, topics. We will then be able to mine this inter-connected web of related objects to understand (a) how ideas travel through social networks of researchers, (b) map out influence and impact of papers, people, departments, grants and research areas, (c) map out sub-fields, and automatically create introductory guides, lexicons, related work, (d) extremely effective people search, (e) automatic paper/grant reviewer suggestion, etc. Papers, people, grants and topics are supported now. Our person coreference system recently won first place in an official competition where other competitors included CMU, Columbia, University of Maryland and Fair Isaac.

Rexa's operation includes:

- Obtaining open access full-text papers by spidering web sites related mostly to computer science, but also including some math, statistics, physics, economics, social science, linguistics, biology, medicine, etc. Our highly efficient spider can download over a million PDFs in just a few days.
- Converting PDF and Postscript to and XML representation for plain text.
- Automatically locating research paper "headers" and "references".
- On these headers and references, performing information extraction of about 14 different bibliographic fields (author, title, journal, year, etc) using a highly sophisticated machine learning method based on conditional random fields (Lafferty, McCallum, Pereira, 2001, ICML).
- Performing cross-referencing of article citations and author names using sophisticated graph partitioning methods, parametrized edge weights, and parallel processing.
- Depositing all this information into an SQL database and indexing the text fields with Lucene.
- Making the service available through a Web site based on Java, JDBC, Javascript and AJAX.

IESL makes use of over 100 Xeon-class compute servers to perform these tasks.

The currently-available Rexa system has over 7 million research paper references and over 879,000 de-duplicated authors.

3 Project Description

The CIS has goals that include: (a) gathering additional bibliographic metadata from new collections, (b) adding value to existing entries, including links to other related articles, (c) providing links to external sources of information related to the entries in the index, for example contact information for the authors. The IMS Biobibs project has the same three needs.

We propose a project that will touch on all three of these needs with some sensible initial steps. We hope that this project will be the beginning of an ongoing supportive and collaborative relationship between Rexa and CIS and the IMS Biobibs project.

The Rexa project aims to be a good community citizen—among both researchers and other providers of bibliographic information services. We believe that research communities benefit when such service providers coordinate and work together, while focusing on the needs of their own constituents. Rexa’s focus is on the computer science literature. Of course, there are long-established and growing links between computer science and statistics that make coordination between CIS/IMS and Rexa interesting.

Our proposed CIS/IMS project consists of the following.

- Spidering additional Web sites suggested by CIS/IMS as having open-access research paper content that would be useful to CIS/IMS. This would likely include non-journal as well as journal content. This will undoubtedly include the arXiv at Cornell, especially its PR and ST sections. We have not yet written the necessary code to spider from the arXiv (it has unusual URLs with a certain kind of re-direction), but we already have a relationship with the arXiv, as well as the necessary permissions and access ports. We could obtain more than 100,000 full-text PDFs.
- Extract bibliographic information from these PDFs—from article headers and references—including title, authors (first, middle, last names), journal, conference, series, volume, number, pages, year, editor, location, publisher and institution. No extraction process (whether by machine or by hand) is perfect. Our conditional-random-field-based extractors are currently the top performers in benchmark tasks tested in the community, obtaining an average of over 97% accuracy.

Of course, for those records obtained from research paper headers (as opposed to references sections), we would also be able to deliver the URL of the full text of the article.

- Spider for BibTeX files as directed by CIS/IMS (whether from BibServer efforts, or from elsewhere), and include their bibliographic records in the subsequent processing.
- Perform research paper reference linking (co-reference, or name authority) between extracted bibliographic records and CIS/IMS bibliographic records. We plan to begin with author records for 900 IMS fellows and 5000 IMS members. This reference linking would enable:
 - the identification of which records are new, and could be added to the CIS,
 - the identification of records already in the CIS, for which we may now know an open-access URL for the article’s full text,
 - the augmentation of existing CIS records with fields that may have been missing from CIS records,
 - lists of research papers that cite papers in the CIS,
 - lists of papers that are cited by papers in the CIS.
 - URLs for many of these papers.

Many of these newly provided links will be between statistics and computer science articles; some will also include papers from statistics, math, social science and other related disciplines.

- Add to Rexa hyperlinks to CIS and the IMS Biobibs Project, so that authorized CIS members can access CIS pages, and all users can access IMS Biobibs pages, directly from relevant Rexa pages. This will drive traffic to statindex.org and imstat.org and help users in neighboring fields discover these sites, and encourage new subscriptions to CIS and new membership of IMS.

A significant consequence of CIS/IMS/Rexa collaboration on Biobib data associated with authors who are IMS members should be the provision of name authority in CIS for such authors. Future work could include provision of further name authority e.g. to coauthors of IMS members, to ASA members, and beyond, gathering biographical information on authors (including institutions, electronic and U.S. mail contact information, former institutional affiliations with dates, advisers, etc).

Future work could also include topical analysis to automatically assign new keywords to articles, automatically place them in a topic hierarchy, augment existing topic hierarchies, and perform analysis of how topics are related to each other.

4 Deliverables and Timeline

The project will take about 9-12 months.

Months 1-3 Have dialog with CIS/IMS about choice of open-access web sites to spider. Perform extensive web spidering. Determine which PDFs are research papers and which are not. Report to CIS/IMS on results, obtaining additional and revised list for further spidering. Develop new code to access the arXiv. Spider requested .bib files, and implement methods to incorporate these records into further processing steps.

Months 4-7 Make improvements to pdf2text, segmentation and extraction code to handle any new formatting issues. Set up and run segmentation and extraction code on CPU server farm. Deposit data into SQL database. Run research paper co-reference linking on server farm, and depositing results into SQL database.

Months 8-11 Run machine learning methods for author homepage-URL-finding. Devise format for delivery of data to CIS/IMS. Create method and write code to incorporate hyperlinks to CIS/IMS into the Rexa search engine. Deliver resulting data to CIS/IMS. Aid CIS/IMS technical personnel in interpreting and incorporating the data, in coordination with VTEX for long-term maintenance of incorporation process.

5 Budget

The project will require significant effort on the part of a research staff programmer. Fortunately the efforts of this person will significantly leverage existing resources from our past research, our existing software, and available computer hardware and storage resources.

We have also obtained leverage in the form of 50-50 matching funds from IMS.

The budget below, for approximately \$10,000, covers just one-half day per week of our staff programmer for 10 months. It includes his salary, standard required UMass benefits, and the standard required departmental IT support fee (system maintenance, email, network bandwidth, backups, etc, per head-count, pro-rated for his percentage time on this project).

Our proposed billing is for \$5,000 from CIS, plus \$5,000 matching funds from IMS.

| | |
|---|-------|
| Programmer salary 8.4% of calendar year | 4,543 |
| Fringe benefits | 1,574 |
| Computer maintenance | 252 |
| Total direct costs | 6,369 |
| Indirect costs (57%) | 3,630 |
| Total Cost | 9,999 |

Our hope is that this project will form the beginning of relationship involving CIS, IMS and Rexa that will continue in the years ahead, and that there will be further funded opportunities together in the future.

ECP Volume 10, 2005

The figures for submissions and publications over the last five years for ECP are as follows.

| Year | Submissions | Volume | Papers published |
|------|-------------|--------|------------------|
| 2005 | 71 | 10 | 30 |
| 2004 | 67 | 9 | 20 |
| 2003 | 51 | 8 | 21 |
| 2002 | 34 | 7 | 22 |
| 2001 | 34 | 6 | 12 |

The mean time to final acceptance for papers published in 2005 was 176 days.

At the end of 2005 Ted Cox and I retired as editors of EJP and ECP. We wish to thank the Associate Editors and referees for their work in the continued success of these journals.

Martin Barlow
Editor

Report EJP

In 2006 the number of submissions increased in the first six months to 87 papers. This increase has not lead to an increase in good papers so the perspective will be that acceptance rates go down and since Associate Editors and referees have a higher load, it will be a difficult task to keep the median time for a decision low.

There seems to be the need for a journal publishing correct and well-written articles which do not have a high level of novelty of the mathematics but might still be useful as reference.

Andreas Greven

EJP Vol 10 2005

82 papers submitted in 2005, up from 71 in 2004
46 papers published in 2005, up from 28 in 2004

Of papers published in 2005:

mean time to first report: 247 days

median time to first report: 174 days

mean time to final decision: 262 days

median time to final decision: 250 days

Ted Cox

Report from the JCGS Management Committee for IMS, June 2006

The JCGS Management Committee met at the JSM, August 2005. Here is a summary of the meeting and issues that arose that may need to be addressed at the Joint Meeting of CoP and ASA Editors.

- The current Editor, Luke Tierney, reported on the state of the journal:
 - The last year saw 176 submissions and 51 articles published. The increase of submissions was dramatic, up from 110. The quality is not diminishing which suggests we may want to consider increasing the page limit.
 - Over the past few years there have been approximately 45-50 papers are published each year. The increase in submissions for 2005 of approximately 15 submitted papers per month, produces an acceptance rate of between 25-30%.
 - The backlog of articles is about 9 months, and median time to first review is about 4.5 months.
 - The editor's budget request for the upcoming year will depend on the new editor. For 2005 the budget included \$4000 request from the editor, and this was raised to \$4600 for 2006, plus an additional \$1000 for th editor changeover.
 - Luke is spending about 15-20 hours per week on editorial work, which is a substantial commitment. Ways to reduce the burden may need to be addressed for a new editor to accept the role. Allentrack, or the addition of an editorial assistant might help.
 - There is continued concern about the small number of graphics submissions.
 - There was little change in the editorial board. There are about 35 associate editors.
- One major activity in the past year was to begin the process of finding a new editor (2007-2009) to begin reviewing new articles in July 2006, taking over from Luke Tierney. David Van Dyk (UC-Irvine) has been approved as the new editor by ASA, IMS and Interface. As part of the negotiations with David the Management Committee agreed to provide an editorial assistant and to switch to an electronic review management system.
- The budget was discussed. A major part of the next year's budget is likely to be to cover the costs of switching to AllenTrack. Whether the switch happens depends on the new editor. The cost of color also was discussed. To be a successful graphics journal more color is needed. A low cost color solution may be worth investigating. Total revenue for 2005 was \$106,500 and expenses \$103,000, yielding a net profit of \$3,500.
- There was a discussion on continuing to raise the profile of JCGS, in relation to graphical and computational articles.
- Double-blinding of ASA articles made the discussion again this year. The general opinion is that double-blinding is increasingly unrealistic today. Authors' identities can often be easily found using the Internet. Double-blinding is the policy of the ASA. Although it might be difficult to strictly enforce it does convey a message to reviewers that author's identities should not be a consideration in their review.
- The relationship between JSS and JCGS was discussed. ASA has been considering making JSS an independent journal. The board approved this in the latter part of 2005.
- JCGS received a new cover last year, one that prominently displays the logos of the three sponsors.

Probability Surveys Annual Report 2006

Volume 2 (2005) contained 13 papers (549 pages), an increase over Volume 1 (2004) which contained 6 papers (392 pages). As of May 12 2006, Volume 3 contains 6 papers (229 pages) and 7 more papers are under review. This seems to be an acceptable albeit slow rate of growth.

As well as traditional survey papers, we are soliciting write-ups of research summer school lecture notes, and in 2005 published one such (by Le Gall).

Acceptance percentages are not comparable to those of other IMS journals for various reasons. Of papers formally submitted in 2005, there were 4 papers rejected as "more like new research than a survey paper". Such rejections are generally made very quickly, directly by Editor or after consultation with an Associate Editor. I also receive (and welcome) informal enquiries about suitability of draft papers, which obviates later need for rejection of formal submissions.

David Aldous