BIBLIOGRAPHIC KNOWLEDGE NETWORK

Proposal submitted to the NSF Cyber-enabled Discovery and Innovation (CDI) Program, April 31, 2008.

Principal Investigators
Jim Pitman (U.C. Berkeley, Institute of Mathematical Statistics)
Brian Conrey (American Institute of Mathematics)
Gary King (Harvard, Institute for Quantitative Social Science)

Contents

1. Project Summary

This proposal addresses three fundamental problems of knowledge management: the compartmentalization problem (how to break down barriers which separate disciplines), the navigation problem (how to guide students and researchers within and between disciplines), and the maintenance problem (how to provide incentives for individuals and organizations to improve the quality of publicly accessible knowledge). It is proposed to solve these problems by gradually distilling the wealth of heterogeneous data now available in digital formats into an openly navigable network of websites, the *Bibliographic Knowledge Network (BKN)*, each node of which is a website dedicated to a specific topic or field of knowledge. Each participating site will typically be designed as a guide for researchers, teachers, and students in a particular field of knowledge, and maintained by a Virtual Organization (VO) with a commitment to that field. The BKN will be created through the development of software which makes it easy for a large collection of mostly small and distributed organizations to brand, select, maintain, and annotate collections of structured scientific content. That content will be made available in machine-readable formats, to allow connections between ideas in different disciplines to be made using methods of machine learning.

Intellectual Merit. We are only beginning to understand the dynamics of collective knowledge systems, and how their creation and development may be encouraged by social incentives and computational advances. The BKN will provide an experimental testbed for Virtual Organizations (VOs) to experiment with the design of such socio-technical systems, both cooperatively and competitively, within constraints imposed by communication standards. This network will provide general insights about creation and maintenance of collective knowledge systems, and how to expand the scope of the Semantic Web as a general tool for the benefit of society.

Broader Impact. The collective knowledge system emerging from this project will be available beyond the walls of academia, and provide well-organized high quality information to anyone with an Internet connection. The expository components of the system will attract people from all backgrounds to pursue scientific careers, and will allow students at all levels to encounter materials which will lead them to higher levels. The system will add great value to other Open Access initiatives, including the system of interoperable digital repositories, Wikipedia, Open Journal Systems, and free academic search services.

2. Introduction

2.1. **The problems.** What researchers and graduate students need today is a far better means of tracking the developments occurring at an increasing pace in their own and related fields. They need to be aware of emerging new topics, innovative methodologies, progress on problems, and disciplinary realignments and intersections. They need systems which allow them easily to step beyond the boundaries of their disciplines to gain an overview of where ideas are moving, as well as an ability to drill deeply into relevant work. They also need ways to contribute to the maintenance and development of such systems to make them viable for the long term. To be more specific, we identify three key problems facing the advancement of science today, which we plan to address by the proposed *Bibliographic Knowledge Network*. These are socio-technical problems in library and information science whose solution should greatly increase the efficiency of research efforts in all fields, and be especially valuable for interdisciplinary and multidisciplinary research efforts.

The compartmentalization problem. The organizational structure within the sciences of disciplines and sub-disciplines, with corresponding university departments and scholarly societies, does not currently lend itself to the ready flow of information and knowledge across divisions. The resulting barriers are reinforced by discipline-specific information services, such as the Current Index to Statistics, MathSciNet and ZMATH, and the Chemical Abstracts Service, whose current subscription models restrict the discovery and mining of bibliographic data to a limited body of licensed users. Further, the quality of statistical information about research activities in each field is limited by the processing capabilities of its information providers, which are currently primitive, despite the power of modern statistical computing and data visualization tools.

The navigation problem. Students entering a field of study, as well as researchers drawing on ideas and methods from fields other than their own specialty, need guidance in navigating the structure, substance, and direction that mark these fields: how does one track the field's key works, major problems, innovative methods, and breakthroughs; how are these developments related to and taken up in other fields? The answers to these questions are not easily found with currently available tools, although they lie both buried in the literature, and living in the minds of current experts. The problem is how to elucidate this sort of information about the landscape of fields, both from the literature and from the experts, and how to make it available to students and researchers through the user interfaces of bibliographic services.

The maintenance problem. It takes some real work by an author or editor to maintain the quality of even a modest sized list of specialized information, for instance a personal or subject bibliography, or a list of web links. As time passes, new items need to be added in a consistent format, the relevance or importance of various items changes, subject classifications may change, links need checking, and so on. On a modest scale, up to the order of a million bibliographic items, this problem has been traditionally solved by both scholarly and commercial publishers who maintained high quality abstracting and indexing services by labor intensive methods. However, these services exacerbate the compartmentalization problem, and the economic viability of these information silos is now threatened by the appearance of free academic search services (Elsevier's Scirus [63], Google Scholar [31], Microsoft's Live Search Academic [39]). These free services range over hundreds of millions of bibliographic items, with awesome capabilities for simple text string search, while compromising on the quality of data and providing limited capability for more advanced searches by subject, author, or other attributes. As the availability of free search services increases, there is less and less incentive for individuals and small organizations to add value to the bibliographic universe by curation of bibliographic data, leading to a decline in the quality of that data.

2.2. **Proposed solution: the Bibliographic Knowledge Network.** We propose to solve these systemic problems by gradually distilling the wealth of heterogeneous data now available in digital formats into an openly navigable network of websites, the *Bibliographic Knowledge Network (BKN)*, each node of which is a website dedicated to a specific topic or field of knowledge. Each participating site will typically be designed as a guide for researchers, teachers, and students in a particular field of knowledge, and maintained by a Virtual Organization (VO) with a commitment to that field.

To build up the BKN, we will develop, host and distribute software that makes it easy for a large collection of mostly small and distributed organizations to brand, select, maintain, and annotate collections of structured scientific content. That content will be made available in machine-readable formats, to allow connections between ideas in different disciplines to be made using methods of machine learning. We will do this with only a moderate programming effort by exploiting existing open source systems (many of which are developed or maintained by the PI's and key personnel for this project) to handle indexing, storing, delivering, annotating, and analyzing the content.

Initially we will strive to achieve viability of a small number of nodes of the BKN, each associated with a relatively small VO of some tens to hundreds of researchers associated with a conference workshop, or with a small professional society or interest group. Then we expect to demonstrate a network effect as the number of such systems grows, and the value of their interconnection becomes evident. Such systems, considered either individually or collectively as a network, exemplify *Collective Knowledge Systems* [89] which exploit the ideas and methods of Social Networking and the Semantic Web [79, 89].

Once widely-available well-supported software is available for creation of domain-specific knowledge websites, both individuals and organizations will have stronger incentives to contribute: it will be easier to share this content, the content will be more useful (because it is vetted and structured), and the individuals and organizations will get credit for the content (because of the VO's control over selection and branding, as well as added value of the content). In addition, having quality content in the system is a powerful attractor for additional high quality content, and the sponsoring organizations have a lot of high quality content which they can place into the system very quickly. That done, the BKN will already be a valuable addition to the sciences.

We also plan to do more. We will help non-experts find specialized knowledge much more easily, and also reveal patterns of scientific research – collaboration, duplication, connection, gaps. We will do this by applying machine learning, network visualization, and network analysis techniques to utilize the additional information that the BKN software has enabled organizations to add to bibliographic data (such as annotations, links, and structured descriptions of disciplinary knowledge).

2.3. **Organization of the proposal.** Section 3 illustrates capabilities of the BKN by describing some ways in which the system could be used. This project brings together a large number of groups who have significant experience developing tools which address a subset of the needs of the BKN. Those groups, and the resources they bring, are outlined in Section 4. More details about the duties of groups and individuals are given in the Coordination Plan in Section 7. In Section 5 we give more details about our planned approach to developing the BKN, list additional resources, and describe some further features of the BKN.

2.4. **CDI Themes.** This project strongly develops the CDI theme of *virtual organizations*: enhancing discovery and innovation by bringing people and resources together across institutional, geographical, and cultural boundaries. The project participants have already formed a multidisciplinary, international team to collaborate on designing the BKN using a web-based communication system for collaborative document preparation. We plan to create or co-opt hundreds more such teams to use an infrastructure designed for collaborative bibliographic data management, and readily adapted for other forms of collaborative work. So the primary purpose of the VO formation, to advance research and education in its area of interest, will be strongly supported by the BKN. We plan, wherever possible, to identify existing *working units* of researchers who already know each other and have a common set of values which they can express through bibliographic curation [99]. We plan to identify leading researchers or others already engaged in subject-specific website building activity, and to engage them to improve the quality and quantity of their online offerings by joining the BKN. Initially, this *outreach activity* will require sustained effort by the project principals. We expect this to provide a critical mass of largely self-sustaining VOs which will ensure the long-term growth of the BKN.

Later phases of BKN development will involve also the CDI themes of *transforming data to knowledge*, and *understanding complexity in natural, built, and social systems.* The BKN is specifically intended to distill bibliographic data into a humanly-comprehensible collective knowledge system, according to the design principles laid out by Gruber [89]: (a) user generated or selected content, (b) human-machine synergy, (c) increasing returns with scale, and (d) emergent knowledge. Nodes of the BKN will provide trusted authorities to certify the relevance and importance of data to particular groups or disciplines. This decentralized rating mechanism provides a new model for relevance or importance of data, which like citation data and web link data can be processed to identify the relevance of specific information to particular queries. We expect that systematic data-mining and use of machine learning techniques based on relevance data should assist a process of transformation of bibliographic data into knowledge of the high level structure of scientific disciplines, to be reflected in solutions of the navigation and compartmentalization problems provided by the BKN. We are only beginning to understand the dynamics of collective knowledge systems, and how their creation and development may be encouraged by various incentives. The BKN will provide a testbed for VOs to experiment with the design of such socio-technical systems, both cooperatively and competitively, within constraints imposed by communication standards. In just a few years time, we should understand better through this process how to create and maintain bibliographic knowledge systems and their associated curatorial organizations, just as we now understand how to create and maintain professional journals and their associated editorial boards. This should provide insights about creation and maintenance of collective knowledge systems, and how to expand the scope of the Semantic Web [79, 89] as a general tool for the benefit of society.

2.5. **Transformative Outcomes.** Expansion of the BKN to a large number of fields will accelerate the arrival of a *bibliographic revolution* whereby responsibility for bibliographic control (the organizing and cataloging of metadata associated with publications) will shift from centralized agents, such as the Library of Congress, OCLC [50] and large abstracting and indexing services, to an aggregation of many smaller VOs which will contribute discipline-specific expertise on a collaborative basis. The potential for such a revolution is recognized in a draft report to the Library of Congress released last year [108] by a committee of 14 distinguished librarians and a representative from each of Google and Microsoft:

> The future of bibliographic control will be collaborative, decentralized, international in scope, and Web-based. Its realization will occur in cooperation with the private sector, and with the active collaboration of library users. Data will be gathered from multiple sources; change will happen quickly; and bibliographic control will be dynamic, not static.

We propose the BKN as a means for the research community to directly influence control of its bibliographic data, and to push libraries and their corporate partners towards a more open system of high quality bibliographic services, with substantial domain knowledge freely accessible to anyone with an Internet connection. Such a system will be of enormous value to all involved in research and education. In the long term, systematic application of machine-learning methods developed for the BKN might so reduce the cost of cataloging information that libraries could afford to expose their vast metadata holdings with open access. That would add great value to the BKN by linking it directly to library holdings, and allowing researchers, teachers and students to collaboratively filter and add value to library records in environments managed by their own professional communities.

3. Uses of the BKN

This section offers a "top down" view of our proposed developments of the BKN in mathematical sciences, starting with a high level organization of the entire field, then moving down to specialty areas, and how a new small VO might be created using the BKN.

3.1. **Organizing an entire science.** Zentralblatt Math (ZMATH) [72] is the world's most complete and longest running abstracting and reviewing service in pure and applied mathematics. The ZMATH Database [73] contains more than 2.0 million entries drawn from more than 2300 serials and journals [74] and covers the period from 1868 - present. ZMATH is managed by FIZ Karlsruhe [4], a non-profit service institution whose mission is to make scientific and technical information publicly available worldwide and to provide related services. The entries in the ZMATH database are classified according to the Mathematics Subject Classification [40] (MSC 2000, currently undergoing revision to create MSC2010), a classification of mathematics into about 5000 sub-areas.

ZMATH has agreed to develop a node of the BKN which will provide a high quality placeholder on the web for each of the more than 5000 subjects in the soon to be released MSC2010. This will be accomplished by using information in its own databases, by using BKN tools to link to existing material on the web, and by using BKN tools to make it easy for people to add new information. For example, ZMATH can use information in its own databases, possibly with some editorial help, to find a small set of important books and papers for each MSC entry. This can provide a 'seed' for the information in its BKN node. By providing just a small subset of its bibliographic data in this way, ZMATH will not compromise its business model of selling subscriptions to its complete index of math papers, because there will still be a demand for this more comprehensive service. A list of most-cited books and papers in a subject is a useful resource that will bring notice to the site and help attract users to contribute further content, such as their own selections or rankings. There is typically a wealth of additional online material that can be associated to each MSC topic: glossary and encyclopedia items, specialist databases, problem lists, conference announcements, lecture notes, videos, preprints etc. It is of enormous value in navigating the online universe of mathematics to see this material become tagged by MSC categories to assist its discovery and utilization.

We are delighted with this development, which we hope to replicate in other branches of science, to persuade ZMATH's parent organization FIZ Karsruhe, and other scientific database services, to make their subject ontologies (typically a tree like the MSC2010) available on the web with open access. The basic idea is that professional scientific organizations can and should stand behind their classification schemes, and allow their free use throughout the web to provide what becomes an essentially public domain map of the field, which any one can use to classify or tag items of interest, and anyone can attempt to cross-walk or relate to any other similarly open classification, to chip away at the compartmentalization and navigation problems. Moreover, the agreed terms of use of the ZMATH subject pages, with machine-readable data with a By Attribution License, allows for anyone else interested in creating a subject specific site in some branch of mathematics, say a site focused on probability, or on statistics, to machine-read relevant ZMATH data, and create their own site however they like based on portions of the ZMATH data, provided that links are provided back to the source. The effect we anticipate is that a large number of satellite websites should grow up around the ZMATH site, with benefit both for the visibility of ZMATH services, and for the broader community.

Other services which we plan to integrate into the BKN in mathematics are the MathPeople [41] service developed by Jim Pitman. This is a prototype name reconciliation service, which allows one to record, for example, that the "James W. Pitman" who wrote a particular paper is the same as the "Jim Pitman" who spoke at a particular conference, has a certain home page, or edited a particular book. PI Pitman has been cooperating with ZMATH on the development of automated subject classification and navigation tools in mathematics and statistics. One of these tools under development is MatNav [6], a mathematics subject navigator. A more advanced tool currently under development for use in statistics is the Statistics Topics Index [64] designed by Pitman and built by Jeff Regier, with support of the Institute of Mathematical Statistics. This tool leverages multiple data sources to provide a comprehensive collection of topics in statistics. It is planned to extend its scope, first to all of mathematical sciences, then to other fields, by developing the requisite domain-specific VO support through BKN outreach.

3.2. **Organizing a research area.** The Probability Web [67] is typical of a number of subject specific sites in mathematics which were started by web enthusiasts in the mid 1990's, using hand coded html. This site now shows its age, and is of limited use, because the content is not readily searchable or reusable at the item level. Pitman has permission to redevelop the site and had its content scraped from html and put in a structured file format, from which a searchable dynamic version of the entire site can now be recovered [46]. Work still needs to be done to provide an adequate maintenance framework for this site, which is a simple union of lists of Abstracts, Books, Centers, groups and societies, Conferences, Jobs, Journals, Miscellaneous, Newsgroups and Listservers, People (a list of about 1000 homepages of people in probability, one of the data sources for MathPeople), Subject sites, Publishers and Bookstores, Quotes, Software, and Teaching Resources. All of these lists will be very simply managed with basic BKN software, and with minimal coordination this would allow bidirectional linking between the Probability Web and the broader ZMATH site. The Probability Web would then be supplemented by an automated list of data associated with MSC2010 topics in probability, and the these topics on the ZMATH site could be supplemented in turn by content selected by ZMATH curators from the Probability Web. In this way, the value of both sites is increased by data sharing, and the desired network effect is apparent even for just two communicating sites. We plan to demonstrate proof of concept with ZMATH/Probability Web cooperation, then seek to engage other aggregators of mathematically useful web links to join the BKN. Two other sites where we have agreement in principle to cooperate on such content exchange and BKN development are PlanetMath, a community based encyclopedia of mathematics, and Knot a Braid of Links [35] a "cool math site of the week" service to the mathematics community provided by the Canadian Mathematical Society. To emphasize again the purpose of the BKN, it is to lower the maintenance burden and increase the quality and aggregated value of such community oriented sites. We regard this as extremely important in engaging students and researchers at all levels to explore the online universe of mathematics in particular, and science in general, as the chain reactions gradually continue to break down more walls in cyberspace.

From the user's perspective, there is very little difference between a journal website and an annotated bibliography with full-text links. Using a combination of automated searches and manual intervention, BKN tools will make it easy to create virtual journals that provide specialist views of the literature. Instead of editors, these virtual journals will have curators [99], who are "subject matter experts who dive through mountains of digital information and distill it down to its most relevant, essential parts". To prototype this idea, one of the project participants, Hadley Wickham, will create a Virtual Journal of Statistical Graphics, which will aggregate together links to all articles related to statistical graphics and visualization. Such virtual journals would be naturally associated with interest groups supported by scholarly societies, such as IMS Groups.

3.3. **Creation of new VOs.** New research groups will use BKN software tools to initiate the activities of their VO. This process will occur for all workshops at the American Institute of Mathematics (around 20 each year; see Section 4.1 for a brief description of these focused workshops), and this will provide valuable feedback as we develop the BKN software. A BKN node will typically be set up prior to the workshop as a means of preparing participants for the workshop activities.

The creator of the VO (called the *user* below, who would typically be a domain expert) uses a web browser to access the BKN and clicks a button to set up a new node. The BKN prompts for very basic information, such as the general subject of the VO. In this case the user selects 'mathematics'. The BKN knows about the MSC2010, through the ZMATH node, and it prompts the user to enter the most relevant classifications.

The BKN now offers several options, and the user selects "add people to the VO" as her first choice. This starts a BKN tool which is an interface to the MathPeople node. The user enters a name, and the BKN tool offers a menu of mathematicians having a similar name. The list is relatively short because MathPeople already knows that various appearances of a particular name on the web actually correspond to the same person. One or two clicks enable the user to select the correct name, which the BKN then associates with the VO. The user now has immediate access to a wealth of professional information associated with the person: a home page, a list of published papers, email address, membership in other VOs, and so on. This process of adding a person to the VO takes less time than Googling for the person's home page and setting up a link in a wiki, yet it provides immediate access to a wealth of higher quality information which has typically already been vetted by other BKN nodes.

The user adds a few more members, and then decides to add some references. She chooses the "manage bibliography" tool. This tool already knows the mathematical subject classifications of the VO, so it starts an interface with the appropriate ZMATH node, to pick up a list of important books and papers in those areas. The BKN offers these options to the user, who selects a few of them to associate to her VO.

Some of the members who were added to the VO are members of other BKN VOs, so their books and papers are already available in a format that the BKN bibliography tool understands. So these are offered as options to the user, who selects a few of them. A couple of the papers are particularly relevant to the initial work planned for the VO, so she adds a comment and those papers are placed at the top of the list.

The user knows of some very recent papers that are relevant, so she has the BKN bibliography tool make an interface to the arXiv [11] and selects a few more papers.

In a short span of time the user has made a list of relevant articles and made some comments on them. But it is not just a list of titles and authors: each article entry carries with it the MathPeople information about the author, what other VOs have selected that paper, what papers cite that paper, etc. And all that information comes with about as much effort as typing a standard bibliographic entry.

Next the user starts a problem list which will help direct the initial research efforts of the VO. The "problem list" tool gives her a form with fields to state the problem, give background information, list relevant papers, etc. The BKN automatically keeps track of who wrote the problem, the mathematics subject classification of the VO, and, later, the progression of additions and changes. The user makes a list of 3 problems, and suggest one as a good place to start because it is the simplest unsolved case of the main problem. (Later that day one of the other members realizes that the special case was actually solved in a recent preprint. So he adds the preprint to the list of papers, marks the problem as ‘solved’, and adds a new, slightly more difficult, problem to the list. The user receives an automatic email notification because she selected that option when she created the problem.)

Finally, the user decides to let other people know what she did. So she starts the “communicate” tool and drafts a letter to the other members. She tells them about the VO and ask people to login and take a look. The VO already knows the email addresses, so there is little overhead for the user to send the message. When they log in, members who were not previously known to the BKN are offered a tool that interfaces to the BibServer program [?] for managing their own papers. This will give the VO full information about those papers, and also make it easy for the member to manage their own papers. (Many members of the Berkeley Math Department already use BibServer).

The members are inspired by the usefulness of the information available and the ease of adding and modifying the material. The BKN node becomes the starting place whenever a member wants to know something about the subject, and when they find out about a new result or prove a new result of their own, their first reaction is to add it to the BKN so that it is easy to find again and is integrated with the other material on the subject. Our goal is to create such attractive tools and interfaces, and to provide adequate acknowledgment of curatorial effort, that scientists will want to throw themselves into developing BKN nodes, as people do with Wikipedia. The big difference with Wikipedia will be the reputation of the nodes as subject authorities by association with scholarly societies and research institutes.

3.4. **Conversion projects.** We recognize that there is an enormous amount of material in legacy formats (such as hand coded html, .doc and .tex formats) rather than fielded text, and that it would be profitable to convert selected, selected, intellectually-structured, high-quality collections of such data to BKN-compliant formats. IMS has some expertise in semi-automated conversion of plain text bibliographies (.html, .doc, .tex) into the fielded bibtex format [19], through a combination of machine-parsing and database lookup. These services will be made generally available to BKN nodes to assist with data processing. Selected legacy content of other types will also be converted with support of the BKN project to demonstrate the value of such conversions. Component tools for this purpose will be made as modular as possible to encourage their widespread support and re-use.

To provide a specific example, as an outgrowth of its workshops, AIM is host to more than 100 annotated problem lists. The lists comprise unique, expert, and (at the time of creation) up-to-date summaries of open problems in different fields within mathematics. A spectacular example is the collection of Problems in Geometric Group Theory [56]. This is a wiki and PDF list of approximately 500 problems in 24 categories, initiated during an AIM workshop and further supported by a special semester in geometric group theory at Mathematical Sciences Research Institute (MSRI).

To someone working in the field, each problem list reveals information that is highly regular, structured, and that integrates well with the organization of knowledge in that field. However, there are a number of barriers to those in related fields who would want to discover and use this information, and there are problems with maintaining the lists. Some are in wikis, which are easy to update provided that the appropriate people are aware of the list and motivated to maintain it, but difficult to use as authoritative, citable repositories. And many of the problem lists exist only in LaTeX and PDF format, which allows for the high-quality display of mathematical symbols but makes it more cumbersome to update. These problem lists will be used as a testbed for development of BKN conversion tools.

3.5. **Encyclopedia projects.** AIM has just been awarded a $1.2 million Focused Research Group grant to create an encyclopedia and database of L-functions. L-functions play a central role in modern number theory, and indeed, in much of modern mathematics. The project is to classify L-functions according to their degree, level, and functional equation parameters; to write an article about the genesis of each, to develop algorithms for their swift computation, and to compile basic data such as the initial string of coefficients, special values, and the first zeros. The result will be a giant database and encyclopedia containing everything we know up until now about L-functions. The L-functions project has funding to develop tools and standards for storing and sharing large amounts of numerical data, and we plan to incorporate those tools and standards into the BKN. We will also work with the L-functions group to ensure that their wiki is converted and further developed in a way that allows a smooth transition to a general BKN encyclopedia tool when it becomes available.

More broadly, Pitman is in contact with a number of institutional partners interested in creating an open access encyclopedia of mathematical sciences, starting from data available in existing encyclopedic compilations including PlanetMath [54], the Springer Encyclopedia of Mathematics, MathWorld, and selected entries from Wikipedia, and building out with navigation features provided by PlanetMath and the ZMATH site. Funding for management and content development of such an ambitious project in mathematics is beyond the scope of the present proposal. However, basic glossary and encyclopedia infrastructure provided by the BKN will provide a general platform to support such content development projects in any field of science. A significant stimulus for encyclopedic content development should be provided by the recently announced Google Knol Project [76].

4. Organizational Partners

This project involves four *primary organizational partners* (AIM, Berkeley/IMS, Harvard (Institute for Quantitative Social Sciences), and Stanford/PKP) whose disciplinary bases are mathematics, statistics, social science, and education respectively. Each of these organizations has ongoing projects to improve electronic information resources. They plan to combine their efforts in this domain and administer a collaborative effort of like-minded organizations to develop and share software tools for this purpose.

Each of these four primary organizations has a PI or senior scientist supported on this proposal, as well as a graduate student, postdoc, or professional programmer who will directly support the software and data development of the BKN. Two of these organizations are in fact pairs of organizations A/B, connected because one of the PIs has a both a position at University A and an executive role in a non-profit organization B which is being brought into the project, as indicated in more detail below. In addition there are eight other partner organizations, Metaweb, FIZ Karlsruhe, PlanetMath, RePEc, the Open Library Society, Creative Commons, the R Foundation, and the Journal of Statistical Software, whose missions are directly aligned with the goals of BKN. Their contributions are indicated briefly below, and in more detail in letters of commitment included as an appendix to this proposal

4.1. **American Institute of Mathematics (AIM)** [9]. AIM has a mission to solve important problems through focused collaborative research. CO-PI Conrey and senior scientist Farmer are two of the directors of AIM. The main activity of AIM is a series of focused workshops, each with about 30 researchers from the world community who share a common vision of making significant progress in the topic of the workshop. Activities before each workshop include creating annotated guide to the literature, developing expository articles, compiling a glossary, and making a list of statements from each participant about their perspective on the workshop and identifying the essential obstacles to making progress. During the workshop, participants initially work to create an annotated list of the most important problems, and possible strategies to investigate their solutions; then they split into teams to begin working on these problems. This sequence of events occurs at about 20 workshops per year.

AIM has developed specialized software to assist in the compilation of problem lists, glossaries, and encyclopedic content. This software has some of the functionality required for the BKN, but it needs to be rewritten in a unified way that is better suited for the long-term maintenance of the code and associated data. AIM will employ a computer programmer to assist in this development.

AIM will take a leading role in the early deployment and testing of the BKN software, and in conversion projects. Each AIM workshop will constitute a new VO, and as BKN functionality is developed, the AIM workshops will switch to using the new software. Conrey and Farmer will closely monitor this activity, enabling rapid development and ensuring that the software meets the needs of the users. Conrey will also serve as liaison to the other NSF-funded Mathematics Institutes [49] and will help develop ways that BKN can support the semester-long programs at those institutes.

4.2. **Berkeley/The Institute of Mathematical Statistics (IMS)** [33]**.** PI Pitman is a member of both the mathematics and statistics departments at Berkeley. IMS is a scholarly society with about 4000 individual members, dedicated to the development and dissemination of the theory and applications of statistics and probability. Pitman has been closely involved with the IMS over the last decade, as an editor, as promoter of open access journals, as President in 2006-07, and now as the member of the IMS executive with responsibility for information technology development.

IMS organizes conferences, publishes journals, and maintains the Current Index to Statistics (CIS), which is a bibliographic index to publications in statistics, probability, and related fields, covering the entire contents of over 160 core journals, as well as partial contents of about 1200 additional journals in related fields, and about 11,000 books in statistics published since 1975. By participating in the BKN, IMS will contribute the software and data of CIS as a testbed for automated and semi-automated bibliographic data processing by VOs with an interest in probability and statistics, especially IMS Groups.

Among the senior scientists on the proposal, Stefano Iacus is the current CIS database editor, and Hadley Wickham has developed a new interface to CIS [45], and will take primary responsibility for the user-interface aspects of the BKN.

Pitman has been working over the last few years on development and management of bibliographic services for mathematics and statistics researchers with funding from the Berkeley Center of Pure and Applied Mathematics [14] and IMS. This led to development of the BibServer software system [96], written in python [59], which provides fully automated multifaceted displays of bibliographic data sets of up to about 1000 items. See [32] for a version of BibServer development supported IMS for display of personal bibliographies. Pitman is also the developer of MathPeople [41] service, described in Section 3. Pitman plans to work with graduate students at Berkeley on technical problems associated with BKN development which are described in more detail in Section 5. This work will be advised in part by machine learning experts Michael Jordan and Tom Griffiths. Students will also benefit from contact with other faculty in the Statistics and Computer Science departments at Berkeley. Pitman also plans to continue ongoing work with graduate and undergraduate students, supported in part by other funding, including an NSF VIGRE grant, to integrate BKN services into the curriculum of the Berkeley Mathematics and Statistics departments, through creation and curation of course notes, problem lists, glossary and encyclopedia items, all hyperlinked to BKN services. Pitman and his students will also assist in the development of prototype BKN nodes associated with various departments at Berkeley, and as well as larger aggegations (e.g. Berkeley Mathematical Sciences) and multi-disciplinary research groups (Berkeley Probability Group, and the Berkeley Machine Learning Group).

4.3. **Harvard/The Institute for Quantitative Social Science (IQSS)** [66]**.** IQSS conducts research into methodology for and applications of quantitative social science. And IQSS provides research and development platforms and production services in the areas of information technology (through the Harvard-MIT Data Center), digital preservation (through the Henry A Murray Research archive), and Digital Libraries (through the Dataverse Network Project). The Dataverse Network Project (DVN) [26] [94] provides a gateway to tens of thousands of data sets, including the combined catalog of all major U.S. social science archives. Co-PI Gary King is a professor of Government at Harvard and Director of IQSS.

The DVN software is an open source system, used by over a hundred different research groups, projects, journals, and public interest organizations to provide access to the universe of research data and to facilitate data sharing. The DVN project will significantly contribute to BKN's need for data storage and the interfaces to that data. The IQSS staff is at the cutting edge of computation-intensive statistical analysis of bibliographic data, and we have budgeted for a programmer based at IQSS to contribute to the data-storage aspects of the BKN.

4.4. **Stanford/Public Knowledge Project (PKP)** [57]**.** PKP is a research and development initiative directed toward improving the scholarly and public quality of academic research through the development of innovative online publishing and knowledge-sharing environments. Senior scientist John Willinsky is Professor of Education at Stanford and Director of PKP.

Begun in 1998, PKP has developed Open Journal Systems [53] and Open Conference Systems [52], free software for the management, publishing, and indexing of journals and conferences, as well as Open Archives Harvester [51] and Lemon8-XML [36] to facilitate the indexing of research and scholarship. This open source software is being used around the world to increase access to knowledge and improve its scholarly management, while considerably reducing publishing costs. PKP will support the use of its software by BKN nodes, as well as integrating bibliographic support services into future releases of its software which will be compatible with the needs of BKN nodes. IQSS is also collaborating with PKP on tools for Dataverse development. IMS and PKP will cooperate to prototype these services for the IMS open access journals. PKP will bring its expertise in publishing/indexing systems to the development of the BKN, to provide leadership, design specifications, and quality control for programming the BKN development work. In that process, we will ensure that the resulting open source software is readily distributable.

Willinsky is a well-known advocate for the open access of scholarly information. He will play a major role in determining the overall direction of the BKN and designing the standards against which BKN functionality will be measured. Willinsky will oversee work of a graduate student who will contribute to the development of the BKN by reviewing and comparing the recent developments in related software systems, and coordinating the testing and debugging of software and systems developed through this project.

4.5. **Further unfunded partners who have provided letters of commitment.**

Metaweb Technologies [43] is a private company based in San Franciso which aims to build a better infrastructure for the Web. Their first product, Freebase [5] is an open, shared database of the world's knowledge, supporting a community with the explicit goal of normalizing and cross-linking disparate data sets. Metaweb, through its Freebase service, will provide a centralized backend infrastructure for storing BKN data, as well as many of the user interface tools necessary for importing, managing and viewing the data. Freebase is an open database of structured and unstructured information. What makes Freebase unique is that it allows collaborative definition and maintenance of structured information. Information schemas can be edited, reconciled, and layered onto content drawn from distributed sources. This makes it easy to write BKN tools which share information between different nodes.

FIZ Karlsruhe/Zentralblatt Math (ZMATH). [72] These organizations will participate in the project as indicated in Section 3.1. See also the accomanying letter.

RePEc (Research Papers in Economics) [60] exemplifies many of the features we plan to build into the VOs supporting nodes of the BKN. RePEc has much experience and success with persuading authors and institutions to offer their bibliographic data in machine-readable plain text formats, which should be of great value to BKN development.

The Open Library Society [7] is a not-for-profit corporation, founded by Thomas Krichel, founder of RePEc, to build freely-available digital libraries. This organization will offer BKN nodes use of its index of Academic an Research Institutions in the World (ARIW)[10], and a general AuthorClaim registration service [13], which allows authors to identify their works in various freely available sources [12]. This provides an open alternative to emerging commercial services such as the Thomson Corporation's Researcher ID [61].

The R Foundation [68] is a non-profit organization founded to support the R project for statistical computing. Two members of the R Core development team (Kurt Hornik and Stefano Iacus) have confirmed their willingness to actively participate in the project, in particular turning the Comprehensive R Archive Network (CRAN, the central R related resource base, featuring a rapidly growing repository of currently about 1350 extension packages for R maintained at WU Wien by Kurt Hornik and colleagues), into a BKN node.

The Journal of Statistical Software [34], sponsored by the American Statistical Association, is an open access electronic journal which publishes articles, algorithms, code snippets, book and software reviews. This journal will develop a BKN node by contributing its data and metadata, the expertize of its editorial staff, and a forum for discussion and publication of software developed around the BKN.

PlanetMath [54] is a virtual community which aims to help make mathematical knowledge more accessible. PlanetMath's content is created collaboratively: the main feature is the mathematics encyclopedia [55] with entries written and reviewed by members. IMS has been supporting the development of NNexus: an automatic linker for collaborative web-based corpora, [47] which is a generalization of the automatic linking engine used by PlanetMath. NNexus is the first system that automates the difficult process of linking disparate encyclopedia entries into a fully-connected conceptual network [48]. PlanetMath will develop a BKN node associated with PlanetMath data, and assist the BKN by contribution and maintenance of the NNexus system for use by BKN nodes.

Creative Commons [24] is a nonprofit organization which provides free tools that let authors, scientists, artists, and educators easily mark their creative work with the freedoms they want it to carry, allowing users to share, remix, reuse – legally. ccLearn [20], a division of Creative Commons which is dedicated to realizing the full potential of the Internet to support open learning and open educational resources (OER), is contributing its Universal Education Search [69], which aims to build a scalable, extensible, federated search for all educational resources on the web, for use by the BKN. We also expect BKN nodes to make use of Creative Commons By Attribution license.

5. Technical Approaches

To the greatest extent possible, the BKN will make use of existing computational resources, including existing data storage and software projects. Much of the programming involves writing code that interfaces between existing software and databases. We have allocated more than 30% of our budget (which translates to 7 person-years) for professional programmers to develop this code. This effort by experienced professional programmers developing production code will be assisted by development of prototypes and algorithms by graduate students, and by the considerable programming experience of the PIs and senior scientists, all of which will be coordinated by Project Manager Jack Alves. See the Coordination Plan in Section 7 for more details.

5.1. **Data sources.** We mention here some of the major open repositories of bibliographic data in various fields which are publicly available for acquisition and filtering by BKN nodes, and their approximate numbers of items: arXiv (474K) [11], RePEc (700K) [60] CiteSeer (767K) [21], PubMed (15M) [58], DBLP (1M) [27], CCSB (2M) [65]. This list is by no means exhaustive. Quite differently structured large collections are available via the social bibliographic services CiteUlike [22], BibSonomy [16] and Connotea [23]. In addition, many BKN nodes should have access to subscription services such as CIS, ZMATH and MathSciNet, with some agreement about the extent to which data from these sources can be further propagated to the network. A typical BKN node will only care about data from a small number of sources, and it will be possible customize data feeds to allow curators to easily select and mark up entries from these feeds. The U.C. Berkeley Library has agreed to make MARC data from its collections available to the project. We are also working with CrossRef, a publisher association which has offered to encourage its members to make their full text data available for text-mining by our project under suitable conditions.

5.2. **Annotated List management.** BKN software will provide participating Virtual Organizations the *capability to easily create and maintain annotated lists of information.* An organized list of references (books, papers, conference proceedings, etc), along with *an indication of what the experts think is important* is of great value to anyone entering an area. By providing sophisticated tools to organize and display information, a high level picture can emerge that is not easily discernible from traditional media such as books and papers. We take a broad view of bibliographic data to include lists of many things which are central to the activities of Virtual Research Organizations:

- People (represented by their names, homepages, publications, ...). Who is doing research in this area? What papers have they written, what talks have they given, and what conference do they attend?
- Problem lists. Research in mathematics and many other fields is driven by problems. Annotated problem lists, giving a clear statement of the problem, references to the literature, suggestions from experts on the difficulty and likely avenues for success, etc, are an ideal way to keep track of the forefront of research.
- Results and key experiments. What are the key results/theorems in the field, and where do they appear? It is a valuable activity to extract and organize the main results from the literature. Traditionally this has been done by authors writing research monographs. Modern collaboration tools allow this process to be aided by a virtual organization.
- Names of topics, subjects and fields, along with some ontology (structure) on these names, with a connection to authoritative glossary entries and encyclopedia articles.
- Data sets (speciality of DVN), and Software (speciality of R project)
- Relations between objects of various types

and so on. This list of types is essentially unlimited, and each type with an essentially unlimited number of attributes, a structure which is well respected by the bibtex format for bibliographic data. Some XML expression of bibtex should be adequate as a data exchange format for such lists. Two key requirements are that we expect entries to be *annotated*, meaning marked up with expert commentary, and to contain *meta-information.* For example, an encyclopedia entry includes who wrote it and what general subject it covers. A research paper includes who wrote it and what VOs have linked to it. Such meta-information is critical for the creation of a system which can sustain growth, and to create the network effect whereby the information becomes easier to explore, and more valuable, as it grows.

An additional requirement is that all information should be *disambiguated* or *reconciled* with authoritative databases as well as possible. For example, if a VO wishes to add an article to a bibliographic list, it will first check whether that article already exists in data accessible to it from its own or other data sources, and whenever possible import the title, publisher etc. from an authoritative source, and record locally a link back to that source. This will sufficiently facilitated by the BKN software that the process will be easier for users than manual entry of the bibliographic data. Typically, the user should be able to just cut and paste a reference, or fragments of a small number of fields, and the software should be able to offer offer a limited set of matches from a database. Reconciliation of items on a large scale will be further facilitated by Freebase, MathPeople, AuthorClaim[13] and available machine matching techniques [2, 92, 86, 93, 95]. A key requirement, critical to the long-term maintenance of the data and software, and for communication, is that all data be made machine-readable.

5.3. **Available tools and formats.** We mention here a number of currently available tools and formats for authors and editors to search and browse data from remote aggregations, including corporate web searches, library searches, and publisher websites. Most such tools are currently connected to just one or two of these kinds of sources. Effort will be required to adapt one or more of these to make a generic tool suitable for BKN nodes to acquire and store data. Data conversion on the fly to bibtex or comparable standard for weblinks (XBEL: XML Bookmark Exchange Language [71]) is currently supported by various web scraping tools. Among those with some capacity for tagging or classifying are del.icio.us [28], Linkagogo [38] and other social bookmarking services for general web links, while CiteUlike [22], Connotea [23], BibSonomy [16] and Zotero [75] handle more traditional bibliographic data. BibSonomy integrates with a large number of desktop bibliography managers, and BKN should do so too. BibSonomy Groups [17] appear to be close to adequate for a small VO to aggregate and tag modest numbers of traditional bibliographic items or web links in bibtex format. But Bibsonomy has no provision of name authority or subject ontology, which will be a BKN requirement, so an additional workflow tool is required to reconcile names or add subject tags. Another available tool for assembling, organizing, and sharing collections of data about resources is CWIS [25], which was created to help build collections of Science, Technology, Engineering, and Math (STEM) resources and connect them into NSF's National Science Digital Library [44]. Some of the features of CWIS which BKN should incorporate include: resource annotations and ratings (a la Amazon [8]), keyword searching (with phrase and exclusion support a la Google [30]), fielded searching, recommender system (a la Amazon [8]), OAI 2.0 export, RSS feed support, integrated metadata editing tool, user-definable schema (comes with full qualified Dublin Core [29]) and prepackaged taxonomies. See also [84] for an extensive list of open standards and software for bibliographies and cataloging, and [15, 18, 3, 105] for more recent developments.

5.4. **Machine Learning.** To develop and sustain interest in community controlled bibliographic service with open access, especially in the computational statistics community, we propose to use CIS to showcase the best available tools for computation-intensive text processing, machine learning and statistical visualization of bibliographic data. Specifically, we plan to stimulate the development of new statistical, data mining and visualization methods to answer important questions: What articles have I forgotten to cite? Who are the key players in a subject? What articles are related to this article? What other application areas use this technique? How are research interests in a discipline evolving? And to simplify difficult tasks: automated subject classification; suggest keywords for a new article; see the context around an article (ancestors, descendants, siblings); summarize an author's body of work; help applied disciplines identify useful theory, and recognize when problems they encounter may have been already solved in other areas; disambiguate author names; use high-level subject classification to focus searches.

For some of these problems there are computational tools available, such as those involved in the CiteSeer [21] and Rexa [62] projects, and the Bayesian hierarchical models used for document topic and citation analysis [80],[81]. Other problems may require innovations in machine learning. All of these problems will require research effort to apply available tools effectively to BKN data. The BKN will record relationships between entities such as authors, articles, and topics. These entities and their relationships correspond to the nodes and edges of some graph. If a suitable graphical model is trained on this data, it will yield a low-dimensional, generative representation of the data. This representation should offer solutions of many of the abovementioned problems. The best collaborative filtering systems today make predictions by combining the outputs of numerous algorithms, in order to model numerous and potentially-disparate aspects of the process that generated the data. Each individual algorithm is computationally tractable, and thus computing the combination of these algorithms is also computationally tractable. Neighborhood-based models – typically computed by memory-based algorithms – model the data accurately at high resolution, but often fail to capture general trends. Conversely, matrix factorization techniques often capture the global structure of the data, but neglect to capture the tight couplings between close neighbors. Bayesian networks also tend to to be better at modeling the data's global structure, but their mechanisms differ from those of matrix-factorization techniques. Thus, they model a different aspect of the data, and their output is a key addition to a combination (linear, or otherwise) of the outputs from other methods. Finally, Restricted Boltzmann Machines are a particularly high-performing yet computationally-efficient technique for modeling tabular datasets. We would also use their outputs in making our final predictions. Some recent work in this vein is [78, 85, 101].

5.5. **Data Visualization and User Interfaces.** High quality user interfaces, including visual representations, are essential to attract and engage potential contributors and editors. Two current prototypes enhance bib-data with faceted navigation and search. Pitman's BibServer system [96] supports for faceted navigation (by year, by author, by journal, etc.), appropriate for 100–1000 bibs. As the number of entries grows, faceted search becomes more important, and is demonstrated in Hadley Wickham's CIS interface [45] to over 250,000 records. In both of these prototypes there is still much scope for incorporating the latest research in the design of faceted systems [90] and in embedded visualisations for navigation [107].

A number of existing tools that visualize 1000s–10,000s of bibs simultaneously, in an attempt to show general trends [82, 83, 87]. We will encourage developers of these tools to make them available for use by BKN nodes, and attempt to integrate them as plugins or otherwise in BKN software. For small numbers of bibs, being able to see neighboring bibs will be useful. Notions of neighborliness will be guided by statistical/machine learning models, and we plan to generate maps that show nearby bibs to help navigate bibliographic domains.

Other output from the machine learning models will offer fertile ground for developing novel visualizations. It is difficult to represent the bibliographic network structure directly, so low-dimensional summaries of interesting structure will be particularly useful for visualization. We plan to explore new visualizations using the R statistical programming environment [98], particularly the ggplot2 package [106], which facilitates the rapid development of new graphics.

5.6. **Ontology.** The tension between traditional ideas of ontology, meaning logical relationships among predefined-defined terms (e.g a taxonomy, perhaps supplemented by extra links and relations) and the more recent notion of a *folksonomy* defined by collections of tags that individuals might apply to bibliographic items [77, 100, 102] is far from resolved. We expect that the development of bibliographic knowledge systems will involve combinations of these methods to achieve best results, in ways that are not yet well understood. We plan to incorporate existing ontologies and folksonomies in some standard format, likely SKOS [1] or OWL [70], for search and classification of holdings and to connect DVN to the existing ontology/folksonomy tools. We will also encourage BKN nodes to develop their own ontologies in a fairly free form way, using the Freebase architecture. By doing so we hope to interest researchers in ontology management to make their tools available for use by BKN nodes, which could then experiment to develop various navigation and display schemes based on these tools. Preliminary experiments suggest that very useful navigation schemes can be built quite easily by combining standard classification schemes like the MSC2010, Medical Subject Headings [42], Library of Congress Headings [37], and softer relationships such as WikiRelatedness [104, 97] and using Google distance to weight approximate ontology matches [88].

Major conceptual challenges arise from the fact that any useful subject ontology is not static but necessarily dynamic and evolving over time. We would like to see the BKN will expose the evolution of intellectual landscapes over time, and address the difficult issues which arise in comparing subject ontologies generated by different but related scholarly communities. We want to escape from the restrictive idea of a "Tree of Knowledge", and to find adequate digital representations more like a "Coral Reef of Knowledge", whose growth over time should be visualizable with suitably animated graphics. We recognize these as difficult problems, largely beyond the scope of the present proposal. However, once the BKN is sufficiently developed, it should provide a springboard for researchers to approach these challenging problems.

6. Conclusion

6.1. **Catalyzing participation.** The greatest challenge taken on by BKN is unlocking information and knowledge: extracting it from authors and from literature databases, and structuring it in a way that invites participation and enhancement of a collective knowledge system. There are very significant legal and economic barriers to opening data sources in a meaningful way. BKN's novel solution is a loose federation of interconnected resources. BKN will provide ramps for organizations to collaborate in small steps. But the end result for researchers will be a great leap. Researchers will have the ability to contribute subject-specific search relevance. This creates a viable new model for trusted authorities to emerge. Rather than hoarding data, organizations will offer their "views" of available information through insightful visualizations.

6.2. **Sustainable outcomes.** A main outcome of this BKN project will be a web service based on open source technologies that the IQSS and IMS use to run their production services of DVN and CIS respectively. All bibliographic and statistical software will be developed with an open source license, using open data standards to encourage widespread adoption and support by multiple professional communities. We expect participating professional societies will sustain the BKN for the long term by a business model of memberships by individuals and organizations, and that research institutes will further contribute some small fraction of their overhead from various funding agencies. All content deposited through the DVN for public dissemination during the grant period will be permanently archived through the Henry A. Murray Research Archive, which is the permanent repository for quantitative and qualitative research at the IQSS, and is a member of the Data-PASS alliance, a data-preservation partnership of major social science data archives in the U.S. The proposed extensions of the DVN will provide a framework for outreach to both end-users seeking information and content directly, and VOs seeking to provide such information while maintaining control and branding.

7. Coordination Plan

7.1. **Roles and Responsibilities.**

7.1.1. *Major Institutional Roles.* Berkeley provides overall scientific direction of the research effort, and leads high level outreach to major partners. Berkeley will lead the research into application of machine learning to knowledge mapping and discovery. Berkeley will also coordinate integration of CIS and other IMS information resources into the BKN.

AIM will develop the content portal integration software tools. These tools will integrate the services provided by the more general systems of the other partners (IQSS, PKP, MetaWeb, and the R foundation) to make it possible for virtual organizations to manage structured knowledge; to publish and maintain authoritative subject hierarchies (ontologies) and terminologies (controlled vocabularies) to be used in classification of this structured knowledge; and provide workflows for vetting and authoritative publishing of this structured knowledge. In addition, AIM will lead outreach in the mathematical sciences to encourage VO participation in the BKN.

IQSS will work on a number of extensions to the Dataverse Network Systems in support of the project. These include extension to ingest, preserve, download and document bibliographic networks; interface to expose data and publication citations as networks; extension to metadata to support richer citation among curated datasets and articles; interfaces to enable social citation of DVN content using selected external social citation tools; and (in conjunction with the R project) visualization and analysis of bibliographic databases stored in DVN. Most of this work is not closely coupled to the integration work done by AIM and can proceed in parallel with it.

PKP will support the use of its software by BKN nodes, as well as integrating bibliographic support services into future releases of its software which will be compatible with the needs of BKN nodes. IMS and PKP will cooperate to prototype these services for the IMS open access journals, and to integrate these bibliographic services with other IMS bibliographic projects. PKP will bring its expertise in publishing/indexing systems to the development of the BKN, and provide leadership, specifications, and quality control for BKN programming that will be funded by NSF.

Metaweb, through its Freebase service, will provide a centralized backend infrastructure for storing BKN data, as well as many of the user interface tools necessary for importing, managing and viewing the data.

7.1.2. *Individual Roles.* PI, Jim Pitman, Determination of priorities for overall research effort, coordination with IMS, direction of research at Berkeley, high level outreach and negotiation of data and expertize exchange with various partner organizations.

CO-PI, Brian Conrey - Overall project administration, oversight of the Project Manager, supervision of research effort associated with AIM workshops, outreach to create BKN nodes in the mathematics community.

CO-PI, Gary King - Supervision of research effort at IQSS, integration of BKN nodes with Dataverse, outreach in the quantitative social sciences community.

Senior Personnel Micah Altman - Co-direction of research effort of IQSS, and expert consulting on bibliographic standards, digital libraries, and high performance computing.

David Farmer - Coordinate the incorporation of legacy material into the BKN, and the evaluation of BKN software by VOs.

Tom Griffiths and Mike Jordan - Supervision of research in machine learning by Berkeley graduate students, outreach to engage the machine learning community in analysis of BKN data.

Kurt Hornik - Development of BKN node for the R project, engagement of the R community, work on visualization and analysis of bibliographic networks.

Stefano Iacus - Integration with Current Index to Statistics.

Hadley Wickham - Development of statistical graphics and data visualization tools, applications to CIS data.

John Willinsky - Coordination with the Public Knowledge Project, development of workflow systems.

7.2. **Project Management.** A steering committee comprising the PI's and selected senior personnel will be responsible for high level objectives and technical decisions related to recommending projects, data formats, web-development frameworks, and software components. The committee will meet periodically as indicated below.

AIM will hire Jack Alves as a 1/4 time Project Manager to provide both continuous management of its software development and outreach, and coordination across partners. (see budget line B2 of the AIM budget; Alves is budgeted at an average of $30,700 per year, with a weighting toward the initial part of the proposal). We have allocated more than 30% of our budget (which translates to 7 person-years) for professional programmers to develop code for BKN applications (see budget line G3 of the AIM budget). This effort by experienced professional programmers developing production code will be assisted by development of prototypes and algorithms by graduate students, with advice from the PIs and senior scientists. All of this effort will be coordinated by the Project Manager.

Partner activities are loosely coupled, and can proceed largely in parallel, but will be synchronized by the project manager, and through the use of collaboration tools (see below).

The overall BKN project will start with a review phase facilitated by the Project Manager. Leaders will review high level responsibilities and technical compatibility requirements. In parallel work will begin to recruit consultants, setup shared systems, and re-evaluate open software for use in BKN projects.

Project teams will be encouraged to use techniques of Agile Development [91, 103] which promote series of short development cycles. At every stage, the Project Manager will facilitate an evaluation that may include: refining requirements, priority changes, assignment of responsibilities, and process adaptations.

7.3. **Software Development Plans.** The initial design decision is to choose a data exchange format which meets the approval of all the partner organizations in this proposal. For the purposes of efficiency and sustainability, and to ensure that all tools meets the highly specific needs of the various types of VOs, the software applications for the BKN will be largely independent. The applications will interact by sharing structured data, so once a data exchange format is determined, code for different aspects of the BKN can be developed in parallel. John Willinsky from PKP, Gary King from Dataverse, and Jim Pitman from IMS, each have extensive dealings with the problems of data maintenance and storage and will lead the development of the BKN data exchange format. However, all PIs, senior personnel, and managers will be involved in this critical first stage of the project.

We anticipate approximately one month to develop a first draft of this standard, and another two months for initial testing by each organization followed by revision to a first working draft. Another step, to be taken in parallel to developing the data exchange format, is to determine what existing software will be used directly by the BKN. As much as possible we will make use of existing software, typically by writing code that interfaces between the needs of VOs and the capabilities of the existing software.

Once the draft exchange format is available, it will be possible to begin writing applications. We plan two initial efforts in this direction.

First, providing adequate support for conversion of legacy data formats, to provide significant amounts of content for prototype BKN nodes. Second, development of tools for the BKN. We will start by installation and thorough testing of Pitman's MathPeople and Stat Topics services and their extensions to other fields, to provide basic support for services built around people and their subject interests, which will be the primary drivers of VO formation. Main tasks will be 1) to ensure that VOs are provided with adequate workflow tools to connect people and subject services to databases they manage, and 2) to connect these databases to faceted displays of bibliographic data using BibServer or similar rendering system.

Monitoring and feedback of how organizations use these tools will guide further development and deployment of tools for the BKN. Initially, development will be restricted to a limited number of bibliographic types, some modest extension of bibtex types [19], but with an extensible data structure and a flexible rendering system which allow for easy addition of new types, and of new attributes for each type.

We expect to have a reasonably smooth development process in place by the end of the first year. This will enable the parallel development of new tools to support the creation and maintenance of data by VOs, and adequate data communication services, which is the primary goal of the first phase of the BKN. The second phase involves machine learning and the development of visualization tools to explore the large structure of research, and facilitate cross-domain search and navigation. That work will ramp up during the second half of the project.

7.4. **Coordination Mechanisms.** Organizations will use a common set of tools for project coordination and dissemination:

- For external coordination a public BKN website, hosted by AIM, will be updated frequently with information about new features, data, announcements, and links to project sites.

- For internal coordination, IQSS will host a shared wiki, blog, mailing list, issue tracker for features and bugs, and content management system to manage project deadlines, milestone lists, documents and decisions. This is not budgeted explicitly, as IQSS has this capability in place. Project Manager will take responsibility for maintenance of the content of this site.

- For software development, organizations will take responsibility for maintaining the software and data they contribute. BKN will encourage formation of open communities for ongoing support of software and data. All shared software releases will be stored with source code in a common public repository (such as sourceforge or google code), supporting version tracking, and public bug submission and tracking.

Each BKN sponsored development effort will assign a representative to communicate project status. The Project Manager will publish monthly progress reports covering all projects. The Project Manager will facilitate periodic topic specific teleconferences that cover near-term issues that require coordination. These might be weekly, bi-weekly, monthly or quarterly as appropriate. The calls are intended for people directly involved with the specific topic. Similarly, leaders will participate in periodic teleconferences to evaluate progress toward milestones, and make decisions about high level BKN issues. Once a year participants will be invited to an in-person workshop at AIM to review objectives, demonstrate progress, and brainstorm new BKN tools and techniques. (see budget line F of the AIM Budget, for $25,000 a year).

References

[1] SKOS Simple Knowledge Organisation System. http://www.w3.org/TR/skos-primer/.

[2] Alias detection in link data sets. Master's thesis, March 2004. Technical Report CMU-RI-TR-04-22.

[3] Sharef (Shared References), 2007. http://dret.net/projects/sharef/.

[4] FIZ Karlsruhe, 2008. http://www.fiz-karlsruhe.de/company_profile.html.

[5] Freebase, 2008. http://www.freebase.com/.

[6] MatNav, 2008. http://bibserver.berkeley.edu/cgi-bin/matnav.

[7] Open Library Society, 2008. http://society.openlib.org/.

[8] Amazon, 2008. http://www.amazon.com/.

[9] American Institute of Mathematics (AIM), 2008. http://www.aimath.org/.

[10] ARIW: Academic & Research Institutions in the World, 2008. http://ariw.org/.

[11] arXiv, 2008. http://arxiv.org.

[12] AuthorClaim: freely available sources, 2008. http://authorclaim.org/collections.

[13] Authorclaim registration service, 2008. http://authorclaim.org/.

[14] Berkeley Center of Pure and Applied Mathematics, 2008. http://math.berkeley.edu/.

[15] Bibshare, 2008. http://bibshare.dsic.upv.es/.

[16] BibSonomy, 2008. http://www.bibsonomy.org/.

[17] BibSonomy Groups, 2008. http://www.bibsonomy.org/groups.

[18] Bibster, 2008. http://bibster.semanticweb.org/.

[19] BibTeX entry types. Wikipedia Entry, 2008. http://en.wikipedia.org/wiki/BibTeX#Entry_Types.

[20] ccLearn, 2008. http://learn.creativecommons.org/.

[21] Citeseer, 2008. http://citeseer.ist.psu.edu/.

[22] CiteUlike, 2008. http://www.citeulike.org/.

[23] Connotea, 2008. http://www.connotea.org/.

[24] Creative Commons, 2008. http://creativecommons.org/.

[25] CWIS, 2008. http://scout.wisc.edu/Projects/CWIS/.

[26] Dataverse Network Project (DVN), 2008. http://dvn.iq.harvard.edu/dvn/.

[27] DBLP: Digital Bibliography and Library Project, 2008. http://dblp.uni-trier.de/.

[28] del.icio.us, 2008. http://del.icio.us/.

[29] Dublin Core, 2008. http://www.dublincore.org/documents/dcmi-terms/.

[30] Google Refined Search, 2008. http://www.google.com/help/refinesearch.html.

[31] Google Scholar, 2008. http://scholar.google.com/.

[32] IMS Biographies - Bibliographies Development Site, 2008. http://www.vtex.lt/st/biobibs/.

[33] Institute of Mathematical Statistics (IMS), 2008. http://imstat.org/.

[34] Journal of Statistical Software. http://www.jstatsoft.org/, 2008.

[35] KaBoL: Knot a Braid of Links. Website maintained by the Canadian Mathematical Society, 2008. http://www.math.ca/Kabol/.

[36] Lemon8-XML, 2008. http://pkp.sfu.ca/lemon8.

[37] Library of Congress Subject Headings, 2008. http://lcsh.info/.

[38] Linkagogo, 2008. http://www.linkagogo.com/go/Home.

[39] Live Search Academic. Microsoft http://search.live.com/academic/, 2008.

[40] Mathematics Subject Classification Scheme, 2008. http://www.zblmath.fiz-karlsruhe.de/MATH/msc/index.

[41] MathPeople, 2008. http://mathpeople.statbib.org/.

[42] MESH: Medical Subject Headings. National Library of Medicine http://www.nlm.nih.gov/mesh/MBrowser.html, 2008.

[43] Metaweb Technologies, 2008. http://www.metaweb.com/.

[44] National Science Digital Library, 2008. http://nsdl.org/.

[45] New Interface to CIS, 2008. http://cis.statbib.org/.

[46] New rendering of The Probability Web, 2008. http://bibserver.berkeley.edu/cgi-bin/probweb/view.

[47] NNexus: an automatic linker for collaborative web-based corpora,, 2008. http://mini.endofinternet.org/publications/nnexus/tkde-revision-submit.pdf.

[48] NNexus details are here., 2008. https://imstat.updatelog.com/W993911.

[49] NSF-supported Mathematics Institutes, 2008. http://www.mathinstitutes.org/.

[50] OCLC: Online Computer Library Center, 2008. http://www.oclc.org/about/default.htm.

[51] Open Archives Harvester, 2008. http://pkp.sfu.ca/harvester.

[52] Open Conference Systems, 2008. http://pkp.sfu.ca/ocs.

[53] Open Journal Systems, 2008. http://pkp.sfu.ca/ojs.

[54] PlanetMath, 2008. http://planetmath.org/.

[55] PlanetMath Mathematics Encyclopedia, 2008. http://planetmath.org/encyclopedia.

[56] Problems in geometric group theory. Test site at http://aimbri12.securesites.net/pggt, 2008.

[57] Public Knowledge Project (PKP), 2008. http://pkp.sfu.ca/.

[58] Pubmed, 2008. http://www.ncbi.nlm.nih.gov/pubmed/.

[59] Python, 2008. http://www.python.org/.

[60] RePEc (Research Papers in Economics), 2008. http://repec.org/.

[61] Researcher ID, 2008. http://www.researcherid.com/.

[62] Rexa, 2008. http://rexa.info/.

[63] Scirus. Elsevier http://www.scirus.com/, 2008.

[64] Statistics Topics Index, 2008. http://stat.regier.ws/.

[65] The Collection of Computer Science Bibliographies, 2008. http://liinwww.ira.uka.de/bibliography/.

[66] The Institute for Quantitative Social Science at Harvard University (IQSS), 2008. http://www.iq.harvard.edu/.

[67] The Probability Web, 2008. http://www.mathcs.carleton.edu/probweb/probweb.html.

[68] The R Foundation, 2008. http://www.r-project.org/.

[69] Universal Education Search, 2008. http://learn.creativecommons.org/projects/oesearch.

[70] Web Ontology Language (OWL), 2008.

[71] XBEL: XML Bookmark Exchange Language, 2008. http://pyxml.sourceforge.net/topics/xbel/.

[72] Zentralblatt Math (ZMATH), 2008. http://www.zblmath.fiz-karlsruhe.de/MATH/home.

[73] ZMATH Database, 2008. http://www.emis.de/ZMATH/.

[74] ZMATH List of Serials and Journals, 2008. http://www.zblmath.fiz-karlsruhe.de/MATH/serials/index.

[75] Zotero, 2008. http://www.zotero.org/.

[76] Google Blog: Encouraging people to contribute knowledge, December 13, 2007. http://googleblog.blogspot.com/2007/12/encouraging-people-to-contribute.html.

[77] H. S. Al-Khalifa and H. C. Davis. Folksannotation: A semantic metadata tool for annotating learning resources using folksonomies and domain ontologies. In Proceedings of the Second International IEEE Conference on Innovations in Information Technology, Dubai, UAE, pages 1-5, 2006.

[78] R. M. Bell and Y. Koren. Scalable Collaborative Filtering with Jointly Derived Neighborhood Interpolation Weights. In *Seventh IEEE International Conference on Data Mining*, 2007.

[79] T. Berners-Lee, J. Hendler, and O. Lassila. A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, May 2001., 2001. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

[80] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested Chinese restaurant process and hierarchical topic models, 2007. arXiv:0710.0845

[81] D. M. Blei and J. D. Lafferty. A correlated topic model of science. *Ann. Appl. Statist*, 1:17–35, 2007.

[82] C. Chen. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006.

[83] C. Chen, Y. Chen, and R. Maulitz. Understanding the Evolution of NSAID: A Knowledge Domain Visualization Approach to Evidence-Based Medicine. *analysis*, 7:8.

[84] B. D'Arcus and J. J. Lee. Open standards and software for bibliographies and cataloging, 2004. Annotated list of web links. http://wwwsearch.sourceforge.net/bib/openbib.html.

[85] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, pages 271–280. ACM Press New York, NY, USA, 2007.

[86] J. Davis, I. Dutra, D. Page, and V. Costa. Establishing identity equivalence in multi-relational domains. *Proceedings of the International Conference on Intelligence Analysis*, 2005.

[87] N. Elmqvist and P. Tsigas. CiteWiz: a tool for the visualization of scientific citation networks. *Information Visualization*, 6(3):215–232, 2007.

[88] R. Gligorov, W. ten Kate, Z. Aleksovski, and F. van Harmelen. Using Google distance to weight approximate ontology matches. *Proceedings of the 16th international conference on World Wide Web*, pages 767–776, 2007.

[89] T. Gruber. Collective Knowledge Systems: Social Web Meets Semantic Web, 2007. http://tomgruber.org/writing/collective-knowledge-systems.htm.

[90] M. Hearst. Design recommendations for hierarchical faceted search interfaces. In *ACM SIGIR Workshop on Faceted Search*, August 2006.

[91] J. Highsmith. *Agile Project Management.* Addison-Wesley, 2004.

[92] P. Hsiung, A. Moore, D. Neill, and J. Schneider. Alias detection in link data sets. *Proceedings of the International Conference on Intelligence Analysis*, 2005.

[93] D. Kalashnikov and S. Mehrotra. A Probabilistic Model for Entity Disambiguation Using Relationships. *SIAM International Conference on Data Mining (SDM). Newport Beach, California*, pages 21–23, 2005.

[94] G. King. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. Sociological Methods & Research, Vol. 36, No. 2, 173-199 (2007).

[95] P. Pantel, A. Philpot, and E. Hovy. Matching and integration across heterogeneous data sources. In *dg.o '06: Proceedings of the 2006 international conference on Digital government research*, pages 438–439, New York, NY, USA, 2006. ACM.

[96] J. Pitman. BibServer. `http://bibserver.berkeley.edu/`, 2003.

[97] S. P. Ponzetto and M. Strube. Knowledge derived from wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212, 2007.

[98] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2007. `http://www.R-project.org`.

[99] S. Rubel. The digital curator in your future, 2008. Blog Post `http://www.micropersuasion.com/2008/02/the-digital-cur.html`.

[100] M. Ruiz-Casado, E. Alfonseca, and P. Castellso. From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach. 1st Workshop on Semantic Wikis: From Wiki to Semantics, at the 3rd European Semantic Web Conference (ESWC 2006). Budva, Montenegro, June 2006., 2006.

[101] R. Salakhutdinov, A. Mnih, and G. Hinton. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798. ACM Press New York, NY, USA, 2007.

[102] P. Schonhofen. Identifying Document Topics Using the Wikipedia Category Network. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 456–462, Washington, DC, USA, 2006. IEEE Computer Society.

[103] K. Schwaber. *Agile Project Management with Scrum.* 2004.

[104] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *21. AAAI / 18. IAAI 2006.* AAAI Press, july 2006.

[105] B. Trushkowsky, K. Campbell, and J. Forbes. An architecture for a collaborative bibliographic database. In *TAPIA '07: Proceedings of the 2007 conference on Diversity in computing*, pages 1–4, New York, NY, USA, 2007. ACM.

[106] H. Wickham. *ggplot2: An implementation of the Grammar of Graphics*, 2008. R package version 0.6 `http://had.co.nz/ggplot2/`.

[107] W. Willett, J. Heer, and M. Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. In *IEEE Information Visualization (InfoVis)*, 2007.

[108] Working Group on the Future of Bibliographic Control. Draft Report to the Library of Congress. November 30, 2007. `http://www.loc.gov/bibliographic-future/news/draft-report.html`.