The JNU Data Depot

An Essential Facility for Big Data Research
for Higher Education in India—Initial Concept Memorandum

Introduction

The mission of the JNU School of Computational and Integrative Sciences (SCIS) is to "innovate and integrate computational and analytical approaches adopted in different branches of sciences." One of the key obstacles to students and faculty in the school is lack of access to key data sets that are essential to furthering research activities.

The JNU Data Depot is a facility that will further these aims by providing capabilities for big data research. Big data is an evolving branch of computational science which focuses on very large, complex, variable data sets that cannot be analyzed using traditional computing techniques. Big data research requires significant computing power, which is provided at SCIS by the High Performance Computing Facility (HPCF). What is often missing, however, is the data needed to feed into that facility.

Data Management: The JNU Data Depot

The JNU Data Depot is a network attached storage (NAS) computer that is housed in the HPCF. Access to the device is carefully limited to the systems administrators and no interactive computing (e.g., web servers) or other user facilities are provided. High standards of cyber security such as limiting login to SSH only limit access.

A prototype system is now running in the HPCF with 8 8-tbyte disk drives. The system runs the FreeBSD operating system and uses the ZFS file system. ZFS allows multiple drives to be combined with a variety of RAID capabilities for redundancy and performance. ZFS works across a number of operating systems, including consumer platforms such as Mac OS X and most LINUX variants using the Fuse plugin. This means that large multi-disk file systems can be moved from one platform to another, a capability missing in most proprietary RAID systems.

The current prototype system uses FreeBSD as the base Unix operating system and an open source system called FreeNAS for package management and configuration. FreeNAS supports NFS, secure FTP, and other protocols for making data available. In the case of the JNU Data Depot the remote file

access is carefully limited to authorized users and only within the confines of the HPCF.

The current 8-disk prototype system will be replaced with a 32-disk system which occupies 4u of rack space. We are currently encoding data in sets of 8 disks, with the ZFS2 capability, meaning 2 disk drives can fail without loss of data. On the 32-disk system, it is envisioned that 3 banks of 8 disks will be installed for active data with an additional 8 disks left open for data interchange purposes. One advantage of ZFS is file systems can be arbitrarily large, allowing for future growth to much larger data sets while preserving capabilities for replication and interoperability.

Funding for the initial core hardware and disk is being provided as a gift from Public Resource, a US NGO which is working with SCIS on this facility. The Board of Directors of Public Resource include Heather Joseph, the Executive Director of the Scholarly Publishing and Academic Resources Coalition (SPARC) and a member of the PLOS Board of Directors, Prof Pamela Samuelson, renowned copyright scholar and Richard M. Sherman Distinguished Professor of Law and Information Management at the University of California, Berkeley among other dignitaries. Its President is Carl Malamud, recipient of the Berkman Award from Harvard, author of 9 books who was behind the first radio station on the Internet. Public Resource works with public libraries and other institutions in the area of access to knowledge, such as the Indian Academy of Sciences.

Data Sets

A number of data sets will be provided on the facility to facilitate not only big data research but also, in a few cases, bulk data interchange with other institutions. Such exchange and interchange will be solely for educational purposes. We are cognizant that some of these datasets have potential copyright issues if broadly distributed, hence the focus on carefully limited access as the default for purposes of big data research.

To enrich educational content in the project we are collaborating with the US based Internet Archive, which began in 1996 by archiving the Internet itself, a medium that was just beginning to grow in use. It is a non-profit (under 501(c)(3) of US) working with the mission of Universal Access to All Knowledge. Internet Archive is essentially a digital library of Internet sites and other cultural artifacts in digital form. It functions like a physical library and provides access to researchers, historians, scholars, librarians, the print disabled, and the general public. They have 20+ years of web history accessible through the Wayback Machine and work with 450+ library and

other partners through its Archive-It program to identify important web pages. Collaboration with Internet Archive will enable sourcing of content otherwise immediately not available over the internet. Access to the material provided by their library partners will enrich the scholarly corpus available in the Depot.

In cooperation with Public Resource and using the facilities of the Internet Archive, a number of open data sets have been preloaded on the prototype facility. These include the "Hind Swaraj Collection" which has over 1,000 high-quality scans of books, including the complete works of Mahatma Gandhi, Pandit Nehru, and Dr. Ambedkar. This is an example of a dataset which could be made broadly available to other educational institutions using the National Knowledge Network (NKN).

A number of other such public data sets will be available on the system, including the Official Gazettes of India, and books that were scanned by the Government of India and form part of the Public Library of India Collection, which has 4 lakh books in 50 different languages, including very substantial collections in Indian languages such as Sanskrit, Gujarati, Punjabi, Tamil, Telugu, Bengali, Kannada, and many others.

Initial data sets are provided by Public Resource and the Internet Archive. One of the goals of the JNU Data Depot is to use volunteers throughout India who have other large datasets that would be of use to the broader community to supplement the initial data sets.

While those data sets are easy to distribute, other data sets will be carefully constrained and used only for the purpose of big data research. Primary among those is a collection of all scientific journal articles that can be used to do essential (and potentially life-saving) research on the scholarly corpus that cannot be presently done in any facility. By applying big data techniques to these data sets such as the scholarly corpus, as well as collections such as the Gazettes of India, we believe important breakthroughs are possible that would be impossible without access to such an essential facility.

Examples of Big Data Research

A number of examples of possible big data research not previously possible are provided. Actual research projects will be approved by the JNU Data Depot Governance Board, described subsequently.

Meta-research is the process of examining extant research using statistic analysis and data extraction. An example would be a genomics research institute examining all journal articles which contain results about particular types of genes. It may be recalled that SCIS was part of one such initiative of

mapping the TB Genome for the first time in the world, in collaboration with CSIR Institute of Genomics and Integrative Biology, Indian Institute of Science and hundreds of researchers from various universities collaborating online as part of the Open Source Drug Discovery (OSDD) programme. We have learnt from that experience that while many of these types of relevant information may be captured from articles that are in a few well-known journals, a manual examination of those articles for particular types of results is tedious and error-prone. Moreover, while one can identify a few journals which have many relevant articles, a more thorough and accurate scan of all journal articles will provide a more comprehensive research results. The JNU Data Depot will allow researchers to mine a more comprehensive scholarly corpus looking for all articles on a particular topic.

A second field of research is mining the scholarly corpus for relationships among different types of research. A prominent example of this type of research is examining citations in all articles to develop new and more meaningful metrics about the importance of each article within the overall corpus of scientific articles or within the context of a specific domain of knowledge. That type of research is not possible without access to the full corpus of articles.

A third area of research is computational linguistic analysis of the corpus. Analyzing articles within a particular domain of knowledge over time can provide new insights to the development and evolution of fields of science. Advanced natural language processing can help identify the presence or evolution of key concepts in different domains of knowledge.

A fourth area of research is to apply big data techniques for the improvement of different fields of computing. For example, Optical Character Recognition, particularly for non-English languages, requires extensive statistical training based on actual data sets. Access to large sets of real-world texts can be used to improve not only OCR but perhaps computer-aided translation and other evolving services.

A fifth area of potential research is to analyze the scholarly corpus to aid in improving the quality of education within India. For example, the University Grants Commission has directed an increased focus on the detection and prevention of plagiarism. This facility will provide access to the full corpus in its possession, perhaps in a some form of compressed form and using advanced data matching techniques might yield more accurate detection of plagiarism.

Monitoring and Auditing

The JNU Data Depot will be overseen by Professor Andrew Lynn of the School of Computational and Integrative Sciences who will serve as the Principle Investigator (PI). Professor Lynn will convene a number of advisors who will assist him screening any proposals for analysis of data in the depot and for any bulk distribution of public data, as well as providing advice on issues such as intellectual property, systems administration, and data sourcing. Professor Lynn, with the assistance of the advisors, will also be responsible for establishing and maintaining operational controls relating to the security of the data.

Access to all data sets will begin by being carefully secured from any access by unauthorized users. In particular, the system will be on a private network with all ports shut off from access with the exception of two ports for systems management using "SSH" keys for access. The current prototype system has three authorized users: one doctoral student under the supervision of Professor Lynn, Professor Lynn himself, and Carl Malamud. Only local access is provided and the system is shut off from the broader JNU network.

Doctoral students and professors will apply to use the data sets by submitting proposals to Professor Lynn which will be vetted with assistance of the advisors. Once a research activity that makes use of the data has been vetted, the doctoral students and professors will be required to access the data by connecting their computer to a private IP network at SCIS, where they will be able to read, process, and cache data for the purpose of analysis.

Enabling Access over NKN

Select data sets that have been cleared for public distribution from the JNU Data Depot will be made available on a private IP address on the NKN, which will be accessible to other institutions on the network. This will enable the JNU facility to be accessed by other institutions.

For example the Hind Swaraj collection having complete works of Mahatma Gandhi, Pandit Nehru, and Dr. Ambedkar will be accessible to other social science institutions. Other institutions like the CSIR IGIB which have big data projects on scientific and health data will have access to the JNU collection. This is expected to generate collaborative projects between academic and scientific communities.

Thus the JNU Data Depot will address one of the key objectives of the NKN which is sharing knowledge over the network, which is happening only on a limited extent at present. Since NKN is available even to universities in remote

areas, the JNU Data Depot's scholarly corpus will be like a virtual library enabling students in distant part of the country to access the scholarly literature and databases available therein. The JNU Data Depot over NKN will thus be an essential facility for big data research for higher education in India. At some near future, we also expect to provide audio book facility to the print disabled making JNU a pioneer in this area.

It is reiterated again that all institutions that are provided access will have to abide with the stringent security protocols that will be put in place.

Summary

The JNU Data Depot provides an essential facility missing in the current JNU computing infrastructure, and indeed missing in India. Big data research on the scholarly corpus and other data sets is exceedingly difficult today and this facility will fill in that gap. Important research not possible today will become possible with access to this data.

Research such as that made possible by the JNU Data Depot is essential to important research, particularly in fields such as medicine that have huge implications for the discovery of real solutions to real problems and the quality of life in India. An essential facility such as this also provides for meaningful education to the students in institutions such as JNU, with connectivity to other institutions enabling JNU to fulfill an important value embedded in the missions of the educational institutions of India and in the Constitution of India.

Making the scholarly literature and databases available over NKN to permit big data research is core to the mission of the JNU School of Computational and Integrative Sciences and will provide the school with a leadership role in provision of knowledge access over NKN making this kind of research capability available in other institutions of higher education throughout India. SCIS also expects to provide some audio books and resources over the NKN to the print disabled students in JNU and other institutions connected to JNU Data Depot meeting an important social mission. This facility provided by SCIS over NKN will function as an essential facility for big data research for higher educational institutions in India.