**JHA — Issues**

<6 images; 1 table>

**Issues and Challenges in Digitizing Sanskrit Manuscripts**

Jha, G. N.

Heritage computing is a new and emerging area under digital humanities, which promotes use of Information and Communication Technologies (ICT) in all disciplines of the humanities. India’s a country of not only unparalleled diversity but also equally diverse intellectual and cultural heritage preserved in the form of knowledge texts. These are broadly classed into printed and handwritten. By rough estimate, David Pingree puts the number at 30 million. According to Gaur and Chakraborty (2009), who report on a survey done by the National Manuscript Mission (NMM), 67% of all Indian manuscripts are in Sanskrit. Dominik Wujastyk (2011) estimates weekly loss to ‘several hundred’. This gets more serious when taking into account that approximately 95% of the texts are considered scientific. For many, this is a shame considering that we are in a digital age and that India is regarded an ICT giant. The paper therefore discusses in what different ways ICT can be brought to rescue our heritage from a virtual extinction.

The paper discusses the status of Sanskrit manuscript preservation efforts in India and the challenges at hand in using digital technologies. While mostly found in India, many important collections are found in the neighboring South Asian countries and also in other Asian and European countries. The National Manuscript Mission (NMM) set up by the Indian government has started with a liberal definition of the manuscripts but has done little so far to help preserve them. They have done surveys, collected a good number from libraries and private collections in north India, and have catalogued and microfilmed some of them. However, the quality of the catalogues prepared is extremely poor. In the name of digitization, only scanning of some of the manuscripts has been done. The manpower that they claim to have trained in various manuscriptology tasks like reading, cataloguing, microfilming, scanning, and editing is mostly nonexistent. The very nature of government-supported work in India is simply inadequate to carry out such a massive exercise without adequate support from innovations in digital technologies like Optical Character Recognition (OCR), Text Editing, text processing, Machine Translation (MT), etc., which are not easy to develop given the nature of these historical texts.

The paper is structured in the following sections:

(a) Objectives and desirable goals of digitizing Sanskrit manuscripts.

(b) Levels of digitization.

(c) Complexity of the tasks.

(d) Standards, tools, and technologies required.

(e) Work done in India in general and by the author at Jawaharlal Nehru University’s Special Center for Sanskrit Studies.

(f) Digitization and beyond.

(g) Suggestions and conclusion.

**Background**

Sanskrit manuscripts represent the oldest body of heritage literature available to mankind. Sanskrit—the oldest documented language of the Indo-European family—has generated interest among scholars across disciplines all over the world. The value of these manuscripts can be seen from an effort by various groups worldwide in acquiring these and making efforts to read and understand the fundamental texts of knowledge. The government of India set up a dedicated body called the National Manuscript Mission (NMM) to tackle the problem; however, the success is far from being satisfactory due to inherent challenges. A project document of the NMM (Gaur et al., 2003) gives rough statistics as follows:

• Total number of manuscripts in India = 5,000,000

• Indian manuscripts available in European countries = 60,000

• Indian manuscripts in South Asia and Asian countries = 150,000

• Number of manuscripts recorded in catalogues = 1,000,000

• Percentage of manuscripts language-wise:

- Sanskrit 67%

- Other Indian Languages 25%

- Arabic/Persian/Tibetan 8%

An older estimate by David Pingree as reported by Wujastyk (2011) puts the number of manuscripts closer to 30 million. The severity of the problem can be gauged from the fact that even a rough estimate is difficult to come by, due the fact that we do not even have a definition of a Sanskrit manuscript—whether to call it ‘Sanskrit manuscript’ or ‘Indian manuscript’ or ‘South Asian manuscript’. Geographically, they are spread in the whole of South Asia, Southeast Asia, China, other culturally related countries, or any other country. These can be found in government libraries, community libraries, personal collections, temples, etc. and are getting lost in big numbers. Wujastyk (2011) estimates the rate of loss to several hundred per week. The NMM has started with a rather liberal definition of what can be called manuscripts (mss) by the Indian government as being at least 50 years old. They are making efforts to collect copies of the mss and have done surveys in some northern Indian states in approximately 35,000 repositories. They have done cataloguing and microfilming of these mss and have trained manpower for reading and editing these. However, the NMM suffers from typical problems that plague government-funded projects in India. The project has resulted in poor quality of catalogues, missing manuscripts, incomplete folios, and access issues. There has been no work on creating technology and standards and an overdependence on manpower has developed.

**Desirable Goals**

Given the current scenario in India, a dedicated and broader initiative has to be undertaken with the following strategy:

• A nationwide survey, which should include all possible kinds of collections.

• An international survey in the neighboring countries and in other countries of the world where collections are available.

• Using a right mix of human labor and computing technologies.

• Digitizing, archiving, search, cross linking.

• Enable reading with the help of digital technologies.

• Translation (human and machine).

• Encourage and fund research groups in the country to undertake fundamental research in key texts.

• Promotion (popular media, target younger readers, multilingual delivery, internationalization).

**Digital Humanities and Sanskrit Manuscripts**

Using available computing and archiving standards and techniques would be desirable. Using NLP techniques in text editing and processing could be done as well. However, one has to understand that the mss we are talking about are very complex documents. And therefore we have to take what could be immediate, intermediate, and future steps in this regard. In general the following digitization tasks could be undertaken:

• Study available text encoding standards and see if they can be used for Sanskrit mss. If not, adapt them or create new ones.

• Create scanned images of texts with metadata and convert them into simple e-books/download.

• Create online/interactive catalogues with multilingual/multi-script search.

• Use trained manpower to transcribe texts.

• Use NLP tools like OCR to get transcribed texts.

As far as the standards are considered, we need to look at available text standards like the TEI and see if they fit the requirement. However, what we have is not a simple text document but period handwritten documents with occasional multiple scripts (sometimes unrecognizable) and even including editing remarks, other characters, and graphics. The current document technologies will not be enough in getting these handwritten texts to electronic form. Therefore, we may need new standards and tools for handwriting/image recognition first. As a next step, tools like data input/output mechanisms, editing, spelling and grammar checking, text readers, machine translation, e-learning/multimedia, etc., will be needed.

**Digital Technologies for Sanskrit: The Current Status and Challenges**

This section discusses the current status of digital technologies and linguistic computing in India and how they can be used for Sanskrit mss digitization. Though Sanskrit heritage computing has been intensively done by a few centers, including Jawaharlal Nehru University (JNU), New Delhi, University of Hyderabad, Center for Development of Advanced Computing (CDAC), Bangalore, INRIA Paris, and Sanskrit Library (Brown University), there is some other notable work in content creation:

• Sanskrit Wikipedia, Sanskrit wikipedia (Sanskrit medium Wikipedia; http://sa.wikipedia.org).

• Sanskrit wikisource (Sanskrit e-texts), Sanskrit wiktionary (Sanskrit encyclopedia), Sanskrit wikiBooks (Sanskrit e-library).

• Digital Library of India (DLI) project (http://dli.iiit.ac.in/)—33,820 Sanskrit books (IISc, CMU, NSF, ERNET, MCIT).

• Clay’s project (http://www.claysanskritlibrary.org), JJC foundation, NYU Press.

• IGNCA (http://ignca.nic.in/sanskrit.htm).

The DLI project does Sanskrit and other Indian languages (and some foreign languages as well). Their Sanskrit number is relatively smaller considering the actual size of Sanskrit texts available in the world. If we analyze the statistics on the DLI website, we are not surprised that the number of Sanskrit books scanned are fewer in number than popular languages like English and Hindi.

English 257,810

Hindi 51,777

Sanskrit 33,820

Source: http://www.dli.ernet.in/ (accessed 2 March 2015).

There are obvious reasons for it:

• Most of the Sanskrit texts are still in the manuscript form and need digitizing and editing.

• Their collections are scattered in many places and many countries and even in private collections.

• They may be in many scripts and sometimes in mixed scripts.

• They may be in multiple disciplines for which Sanskrit-fluent experts may not be easily available.

• Sanskrit may still suffer from certain social preferences in which certain sections may not be allowed access to all kinds of texts.

Not only are the Sanskrit manuscripts vast, diverse, and valuable, there are equally diverse problems in taking precise stock of them and evolving the necessary standards and technologies for their preservation and promotion, as we shall see in subsequent sections.

**A Summary of Work Done at JNU**

The work at JNU started in 2002 with creating a database of the Amarakosha—the oldest thesaurus of Sanskrit. It is called Multilingual Online Amarakosha (OMA) and was funded by the University Grants Commission (UGC) under a University for Potential for Excellence (UPOE) program. It has a Java-JSP front-end and SQL server as a back-end. The system has the following features:

• Facility for online data entry/editing by language experts.

• Includes Sanskrit, Hindi, Kannada, Punjabi, Bangla, Oriya, Assamese, Maithili, and English. Scope for other major Indian languages.

• Multilingual data is being stored in Unicode.

• Stores up to 50 synonyms with category, gender, number information, and detailed glosses.

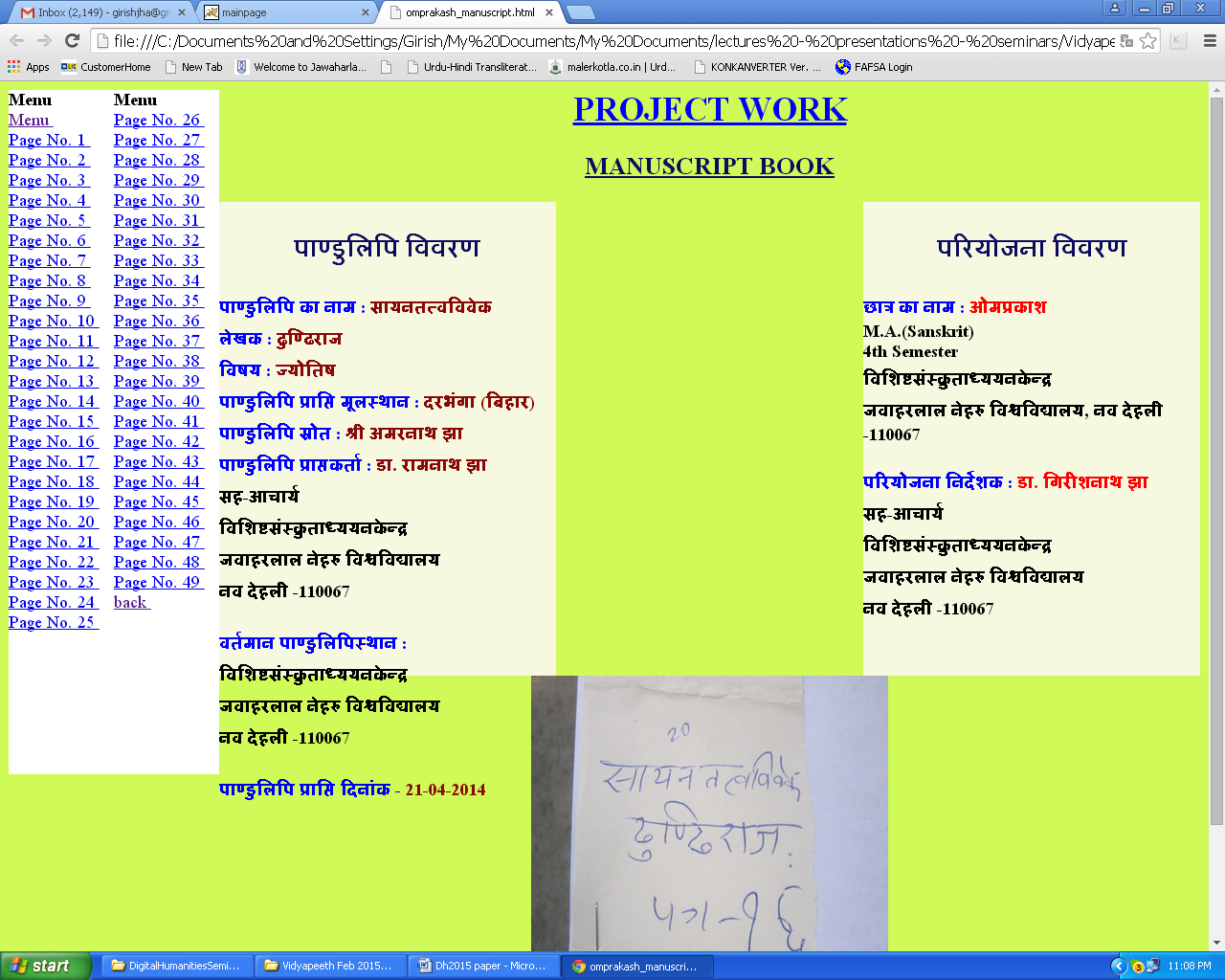
• Cross-referencing among synonyms.

• Search capability in the supported Indian language.

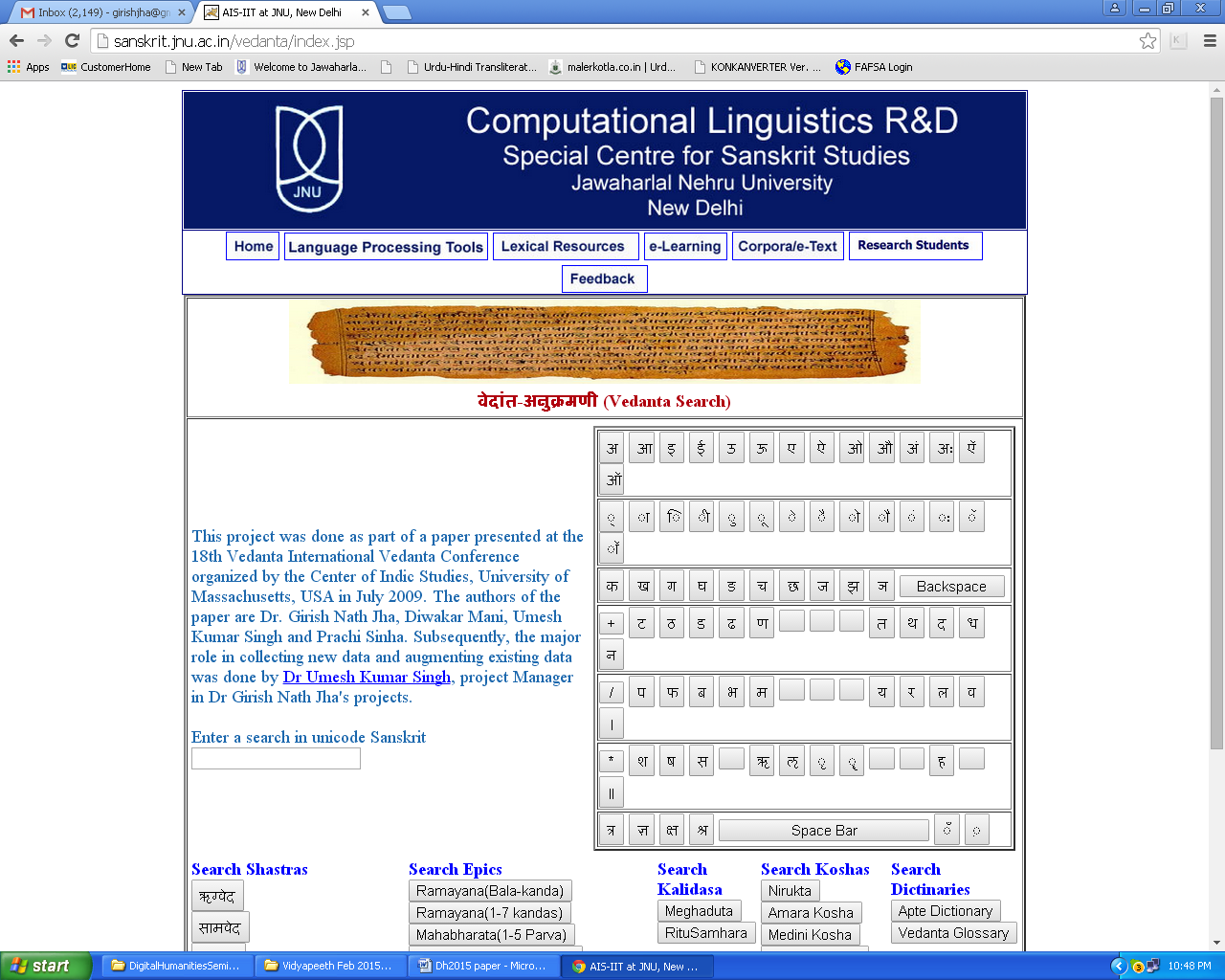
• Ontology display.

This system can be tested at http://sanskrit.jnu.ac.in/amara/index.jsp.

Any digitizing of manuscripts will need lexical resources of this kind. We have collected some rare manuscripts of Sanskrit from the Mithila region and some have been digitized by the students. A sample manuscript ebook is shown below:



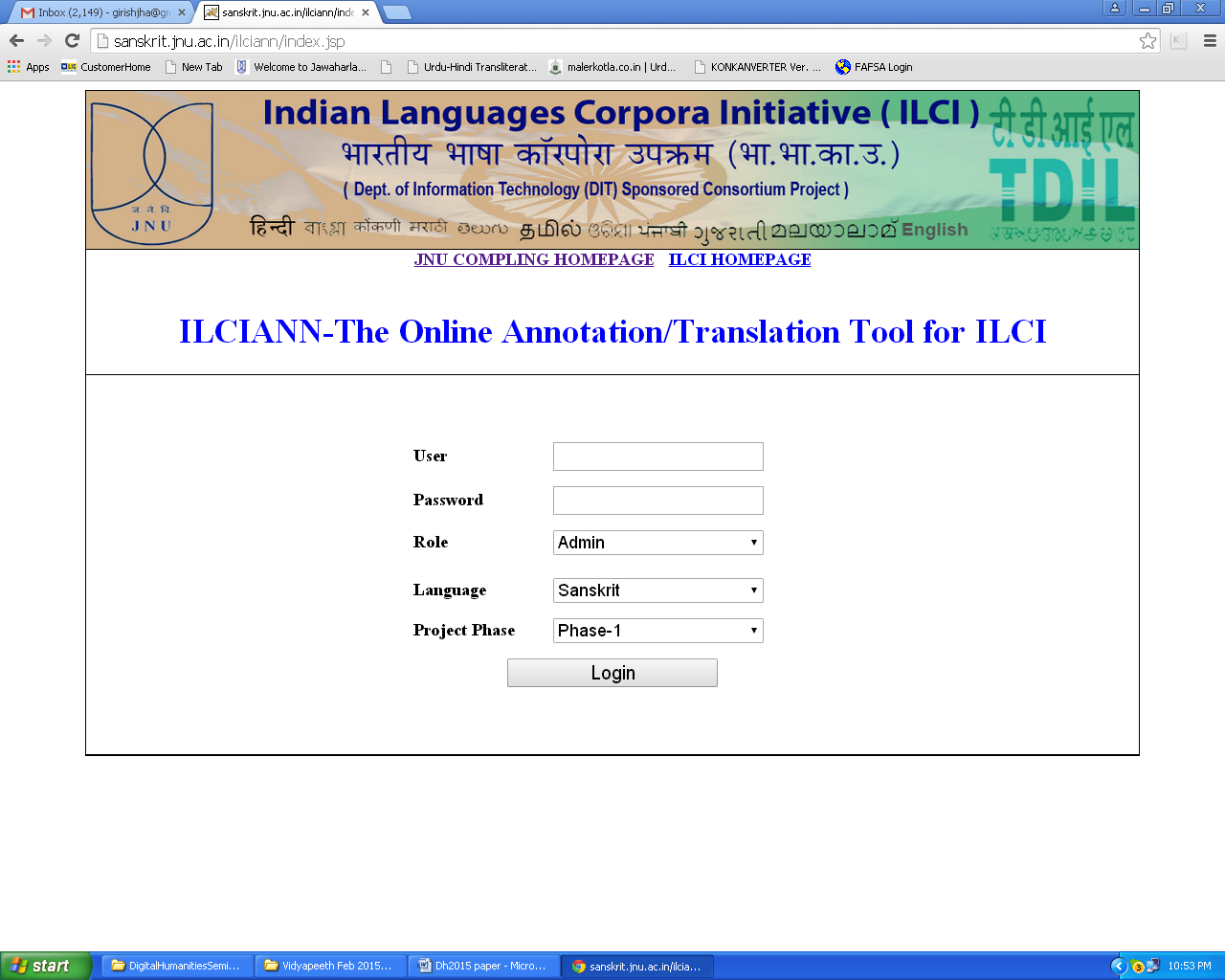
Besides rare manuscripts like the one above, we have several important Sanskrit texts have been digitized in the form of searchable indices.



There are several important lexical resources in the searchable mode and cross linking. The list is given below:



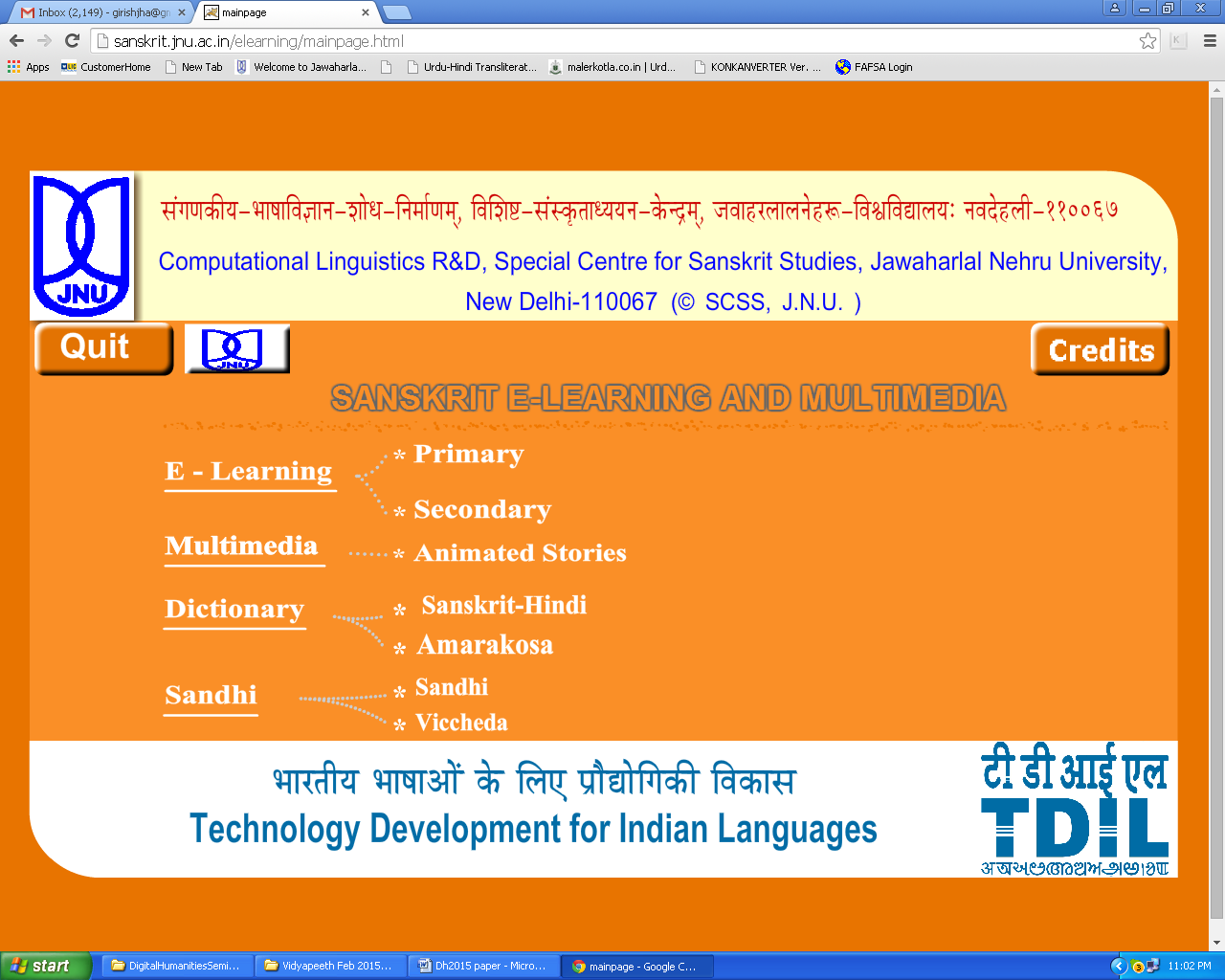
A multilingual resource creation platform called Indian Language Corpora Initiative (ILCI) is being used to develop resources for 17 Indian languages including English but excluding Sanskrit due to less perceived market demand for it.



There are a range of text processing tools for Sanskrit, including the Text to Speech (TTS) Named Entity Recognition (NER), Sandhi processing, and morphology analyzers and generators. The links to these applications can be found at http://sanskrit.jnu.ac.in/index.jsp. In addition, a text converter program converts Sanskrit texts into nine other scripts, The interface can be used at http://sanskrit.jnu.ac.in/ile/index.jsp.



For the promotion of Sanskrit texts, there is an e-learning application targeting children and younger students. The program has content for the Indian government-run Central Board of Secondary Education (CBSE) syllabus for Sanskrit. In addition, there are animated stories and language learning tools



Two more initiatives have been ongoing and making significant progress—one sponsored by the Center of Indic Studies, University of Massachusetts Dartmouth (UMASSD) to bring out fundamental sciences in ancient Indian texts, and the other to promote learning of India’s ancient scripts like Sharada, Brahmi, and Grantha.

**Suggestions and Conclusion**

Since Sanskrit manuscripts are found in many countries, they can be termed an international heritage. This section suggests setting up a dedicated international body with the following potential activities:

• **Sanskrit Manuscripts Recognition Consortium (SMaRC),** with members from the traditional institutes, technology institutes/groups, and computational linguists.

• Major goals could be manuscript OCR, proofing, content creation, dissemination, and collaboration with various groups or institutes that have digitized/microfilmed Indian manuscripts to bring all of them on one platform.

• Each partner institute/organization will have a well-crafted task set to accomplish.

• Follow a sound methodology, including the best experts in the world on the advisory committee.

• Create a *śāstra* panel with members from the best of our experts on manuscripts (in each area of traditional knowledge).

• Create trained manpower by partnering with Sanskrit institutions and revising curriculum to include applied courses.

**References**

**Chandra, L.** (2006). Rare Indian Manuscripts in Asian Countries. *Tattvabodha,* vol. 1, NMM.

**Gaur, R. C. and Chakraborty, M.** (2009). Preservation and Access to Indian Manuscripts: A Knowledge Base of Indian Cultural Heritage Resources for Academic Libraries. *Proceedings of International Conference on Academic Libraries (ICAL-2009)*, 5–8 October 2009, Delhi University Library System, University of Delhi (North Campus).

**Jha, G. N.** (ed.). (2010). *Sanskrit Computational Linguistics*, Springer Verlag, Germany.

**Jha, G. N.** (2014). *Sanskrit Manuscripts Digitization and Beyond: The Current Status in India*, presented at *WAVES-2014*, Maharishi University of Management, Fairfield, Iowa, 3 August 2014.

**Mandal, S. and** **Maiti, S. K.** (2006). *Scientific Analysis of Traditional Indigenous Methods of Manuscript Preservation*, *Samrakshika*, Series 1, NMM.

**Marcantonio, A. and** **Jha, G. N.** (eds) (2010)., *Perspectives on the Origin of Indian civilizations*. DK Printworld, Delhi.

**Pathak, K. N. and Jha, G. N.** (2013). A Generic Search for Heritage Texts. In Vetulani, Z. and Uszkoreit, H. (eds), *Human Language Technologies as a Challenge for Computer Science and Linguistics, Proceedings of the 6th LTC*. Fundacja, Uniwersytetu im. A. Mickiewicza, Poznan, Poland.

**Winternitz, M.** (1996 repr.). *A History of Indian Literature*. Vols 1–3. Motilal Banarasidass, Delhi.

**Wujastyk, D.** (2014). Indian Manuscripts. In Quenzer, J. and Sobisch, J.-U. (eds), *Manuscript Cultures: Mapping the Field*. Studies in Manuscript Cultures 1. Berlin: De Gruyter.