Dutiya-Parakkamabāhu Cullavagga Transcription Project Proposal

Bhikkhu Sujato

August 2018

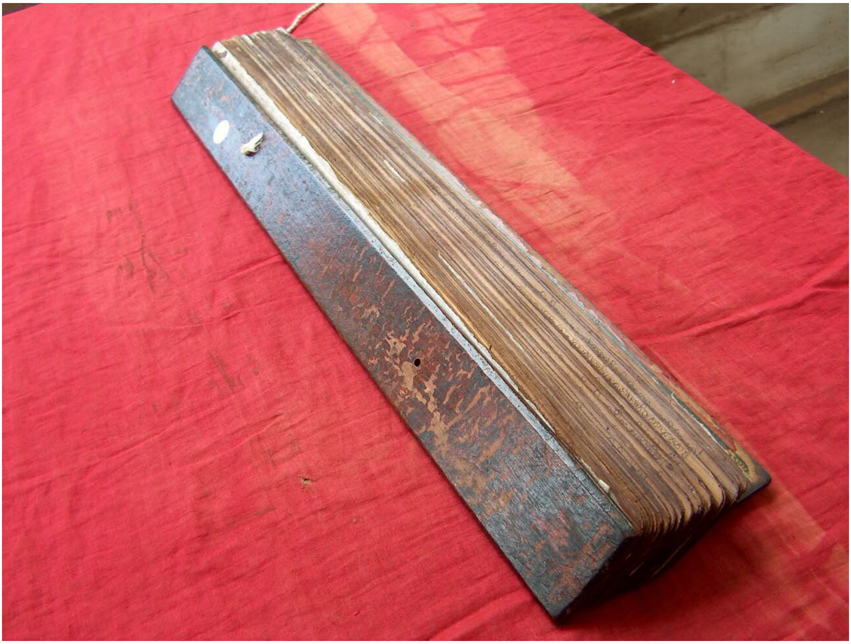


Figure 1: The manuscript in its wooden covers

Contents

1 Introduction

In the National Museum of Colombo there rests one of the oldest, and arguably most important, Pali manuscripts found anywhere in the world. It contains the Cullavagga, which consists of chapters 11–22 of the *Khandhaka* portion of the Pali Vinaya.

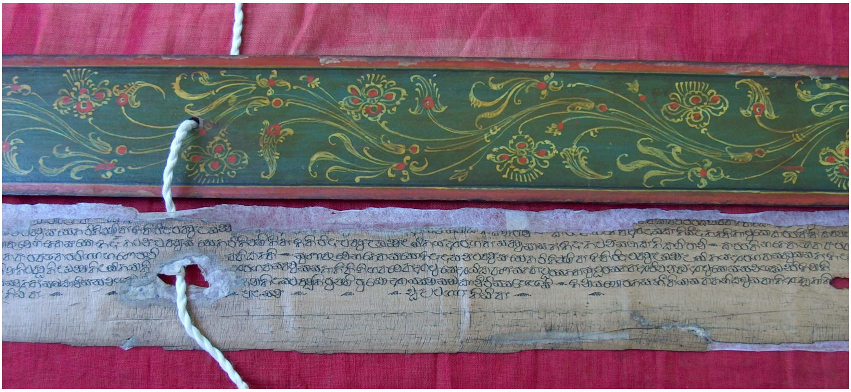


Figure 2: A leaf of the manuscript and the decorative cover

The sign accompanying the exhibit identifies it as object number NM-2995 69L1. The manuscript consists of 143 palm leaves in good preservation. It was acquired by the National Museum on the recommendation of Senarath Paranavitana, who made the initial identification of the king with Dutiya-Parakkamabāhu. The cover of the book is wood painted with depictions of deities, in a style similar to those of the Polonnaruva period. For further details one is referred to H.C.R. Bell, Annual Report 1901, Department of Archaeology.

The manuscript itself includes a note (HPIM4634[1], HPIM4635[2]) to the effect that the manuscript consists of 144 leaves, sized 23×2 ¼. It was purchased for Rs 1000 along with 38 other manuscripts from the late Mr. H.C.P. Bell's estate on 12/1/1938.

[1] </images/dpcv-microfilm/HPIM4634.jpeg>

[2] </images/dpcv-microfilm/HPIM4635.jpeg>

The colophon in the manuscript says that it was copied during the reign of Parākramabāhu, identified as the second of that name, which would place it in the 13th century. This makes it probably the oldest manuscript in Sri Lanka, and one of the oldest in the world. The only older published Pali manuscript I am aware of is from Nepal, dated to the 8th or 9th century, and consisting of a few pages, coincidentally, also of the Cullavagga.

Apart from that, so far as I am aware, all our current sources for Pali texts date from manuscripts of the 18th and 19th centuries. There are old manuscripts in all Theravadin lands, though none so old as the 13th century; but they are rare, understudied, and so far as I know, unpublished.

It is, therefore, imperative that the Cullavagga text be digitally preserved and transcribed. In this project outline I set forth a proposal for how to accomplish this.

The aim of the project is to make available to scholars internationally the contents of this work for research. We propose a comprehensive approach to preparing a digital text. This will include the following items:

- Review the manuscript to ascertain the state of preservation.
- Scan the manuscript into high-resolution images.
- Engage an epigraphic expert to assess the script.
- Carbon date the manuscript.
- Have the manuscript carefully typed and proofread.
- Publish digitally and in print.
- Document the project in academic journals and conferences.
- Publicize the project in popular awareness.

To accomplish this will require the cooperation and goodwill of scholars, governmental organizations, and universities.

I wish to emphasize that I have no experience in dealing with manuscripts, and have no authority in the field. I am simply a monk who has studied Pali and Buddhism. However, for the past 15 years I have run a website called SuttaCentral, which handles over 70,000 Buddhist texts in over 40 languages. Hence I have considerable experience in managing digital texts, so have described the digitization process in some detail.

Everything in this document is simply an initial proposal, and feedback and criticism is welcomed. I would invite all those interested in this project to consider my proposals, and work together to improve them. This is an evolving document, and we will expand and correct it as the project develops.

2 Some details and terminology

So far as I know, the manuscript in question has not become known by any accepted abbreviation. For reference purposes, and especially for handling digital resources, it is useful to agree on a clear and consistent abbreviation and referencing system.

2.1 Naming the manuscript

Since the manuscript is said to have been sponsored by King Parākramabāhu II, it seems sensible to name it after him. The Pali term would be Dutiya-Parakkamabāhu. (See Dīpavaṁsa II 37.84[3] *dutiya-parakkamabhuja*.) Thus I propose the following:

> **Name the text "Dutiya-Parakkamabāhu Cullavagga", with the acronym DP-CV, or dpcv for programming contexts.**

[3] <http://gretil.sub.uni-goettingen.de/gretil/2_pali/3_chron/dipav_2u.htm>

Note that, while the Sanskritic form *parākramabāhu* has become standard, in the Dīpavaṁsa II[4] and the Mahāvaṁsa[5] the form *parakkamabhuja* is more common. *Bāhu* and *bhuja* are synonyms, both meaning "arm". In the Mahāvaṁsa, for example, *parakkamabāhu* appears once only, while *parakkamabhuja* occurs 56 times. Until it is edited, of course, we cannot say what form is used in the manuscript itself. The Sanskritic drift of the king's name is, aptly enough, an example of the kind of problem that we wish to address through studying this manuscript.

2.2 Naming the chapters

The Cullavagga is the second portion of the Khandhaka section of the Pali Vinaya. In the Khandhakas, it is preceded by the Mahāvagga, which consists of 10 chapters. The Cullavagga consists of 12 chapters; hopefully the manuscript contains all twelve, though this remains to be confirmed.

Thus we have:

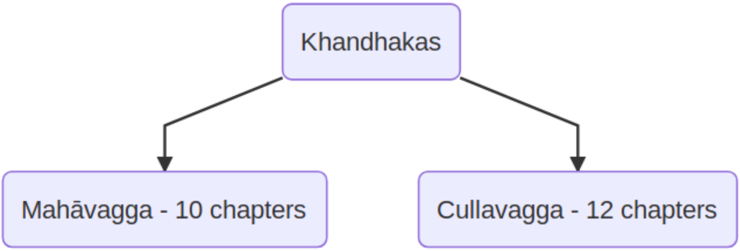


Figure 3: Structure of the Vinaya Khandhakas

[4] <http://gretil.sub.uni-goettingen.de/gretil/2_pali/3_chron/dipav_2u.htm>

[5] <http://gretil.sub.uni-goettingen.de/gretil/2_pali/3_chron/mahava_u.htm>

So for the numbering of chapters, I propose the following:

> **The chapters of the Cullavagga be named Kd 11–Kd 22, or kd11-kd22 in programming contexts.**

This follows the conventions established on SuttaCentral, where Kd stands for Khandhaka, and the chapters are numbered as a simple increment.

Note that the division between Mahāvagga and Cullavagga is not intrinsic to the texts, and does not correspond to any meaningful division of content. It is merely a division of convenience established for recitation and copying manuscripts. Other Vinayas in Chinese, Sanskrit, and Tibetan lack this division and simply organize this portion of the texts as Khandhakas (Skt. *skandhaka*, sometimes also called *vastu*, etc.).

3 Background

The texts of the Pali Tipiṭaka are our most important witness for the original teachings and person of the Buddha. Their significance has always been recognized in Theravada Buddhism, and became known internationally in the 19th century through the work of scholars such as T.W. Rhys Davids[6] in England and Bunyiu Nanjio[7] in Japan.

Pali texts were traditionally passed down on palm (or *ola*) leaf manuscripts. From the late 19th century into the 20th century these manuscripts have been transcribed and published in book form. From the 1990s the books have been digitized, and are now widely available in several editions on the internet. (See Lancaster, "Digital Input of Buddhist Texts", pp 288–296.) All this has greatly increased the availability of the Pali texts and has benefited scholars immensely.

[6] <http://0-www.worldcat.org.novacat.nova.edu/identities/lccn-n50035648/>

[7] <http://0-www.worldcat.org.novacat.nova.edu/identities/lccn-n81033028/>

However, in recent years scholars have paid greater attention to the problem of manuscript authenticity. Almost all of the manuscripts in existence date from the 18th and 19th centuries. This means that the entire Pali tradition rests on physical witnesses no more than a couple of centuries old.

By way of comparison, modern Chinese Buddhist texts mostly stem from the Tripiṭaka Koreana[8], a 13th century edition carved in wood blocks.

Linguistic, historical, and other evidence gives us good reason to be confident that the Pali texts are, on the whole, much older than this, and that they substantially date back to the time of the Buddha or shortly after. (See Sujato and Brahmali, *The Authenticity of the Early Buddhist Texts*.) Nevertheless, we should not neglect the opportunity to further test and refine our understanding of the historical provenance of the Pali texts.

A further problem is that almost all the Sri Lankan manuscripts stem from texts that were imported from Burma or Thailand around the 18th century. Thus, although Burma and Thailand originally received their texts from Sri Lanka, we have little evidence for a Sri Lankan transmission that has not passed through those countries.

For a discussion of issues regarding the state of Pali manuscripts in Sri Lanka, see Bhikkhu Nyanatusita, *Pali Manuscripts of Sri Lanka*[9]. He speaks of the importance of a project to preserve and make available the many manuscripts in Sri Lanka, observing that:

> The digital photographs should not be stored in just one institution, where they might be lost, as happened with the microfilms at the National Archives, but should be made freely available on the internet so that any scholar can access them. Also detailed information (description, history, photographs, list of

[8] <https://en.wikipedia.org/wiki/Tripitaka_Koreana>

[9] </related-essays/Pali_Manuscripts_of_Sri_Lanka.pdf>

> manuscripts) should be given about the monasteries where the manuscripts were photographed.

While we are not prepared to undertake the more wide-scale digital preservation envisaged by Bhikkhu Nyanatusita, by preserving and published DP-CV we can, perhaps, stimulate interest and further action.

One of the few early Pali texts in existence was studied and published by Oskar von Hinüber in his *The Oldest Buddhist Manuscript*[10]. This offers a transcription, description, and detailed study of a 8th–9th century Pali manuscript from Nepal by a leading Indological philologist. Von Hinüber suggests that it may have derived from a branch monastery of the Mahāvihāra on the mainland, possibly at Bodhgaya.

As it happens, the text itself, while only a few pages, is from the Cullavagga. Thus it is a small portion of the same text found in the DP-CV, and when the DP-CV is published it will enable a study of the same passage in manuscripts of the 8th, 13th, and 18th centuries. SuttaCentral has completed a digitization of this text, which we have named the Bendall Cullavagga (`bendall-cv`), which can be viewed on Github[11]. This was intended as a pilot project to test and demonstrate some of the methods we are proposing for DC-CV.

When reading Pali texts in modern editions, it is quite evident that a process that might loosely be called "Sanskritization" has taken place. We frequently see spelling variations even in common words, such as *supaṭipanna* vs. *suppaṭipanna*, *viriya* vs. *vīriya*, or *byākaraṇa* vs. *vyākaraṇa*. It is generally believed that this process occurred under the influence of medieval grammars such as the Saddanīti, which were based on the Sanskrit grammars. In most cases the spelling variants are inconsequential, and no more affect the meaning than, say, the choice to use UK or American spelling for English. Nevertheless, there are some cases where variant readings can affect the meaning

[10] `</related-essays/Hinuber_Oldest_Pali_Manuscript.pdf>`

[11] `<https://github.com/sujato/bendall-cv>`

of a passage. Thus it is important for scholars to study the nature of these changes, which can only be rigorously understood through studying early manuscripts or inscriptions.

4 Literature Review

This manuscript has been mentioned a number of times in published articles.

The initial report of the discovery is said (in the sign on the exhibit) to have been posted by H.C.R. Bell in the Department of Archaeology's Annual Report of 1901. Unfortunately, this edition of the Report does not seem to be available online. We should ascertain whether a copy of this Report exists at the Department, and if possible have it scanned and uploaded.

The manuscript is discussed in fair detail in P.E.E. Fernando's excellent *A Note on Three Old Sinhalese Palm Leaf manuscripts*[12] (1982). Fernando confirms, on the basis of epigraphic analysis and historical context, Paranavitana's conclusion that the manuscript was copied in the reign of Parākramabāhu II of Dambadeṇiya, and is definitely later than the Polonnaruwa period. He analyzes some details of the script, which may be useful for our typists. He also discusses the identification of the people and places mentioned in the colophon. The Konduruvā forest is in the vicinity of Dambadeṇiya in the North-Western Province, while Beligala is about 25km south-east of Dambadeṇiya.

Fernando discusses the colophon, which poses a number of problems of interpretation, and offers the following translation:

> This is the Pāli book Suluvaga that the Venerable Great Lord Medhaṅkara of the Konduruvā forest caused to be transcribed by the Grand Thera Sumedha of Beligala as a gift to the *saṅgha*, after collating a whole *nikāya*, being satisfied (with regard to its

[12] `</related-essays/Fernando_Three_Old_Sinhalese_Manuscripts.pdf>`

accuracy) after consultation (with competent scholars), with the patronage of King Parākramabāhu, the Sovereign of Laṅkā and the participation of fellow-monks living the holy life, such as *Theras* and *Grand Theras*, for the purpose of transcribing (providing) one book for each monk as a gift to the venerable *saṅgha* living in the island of Laṅkā.

5 Who is involved

A project such as this requires a variety of specialized technical skills and equipment, and can only be accomplished through the cooperation of several parties. We will endeavor to work together with all parties who have an interest in the matter.

I propose that a Memorandum of Understanding (MoU) be signed by the major parties involved in this project.

Here is a preliminary note on relevant parties so far. Note that this is merely a record of initial meetings and does not imply that these people or organizations have made any commitment to the project.

1. **Sri Lankan Department of Archaeology**: The Department of Archaeology is the owner of the manuscript, and the governmental institution most directly concerned with preserving and making available Sri Lanka's ancient culture. We have had a preliminary meeting with Prasanna Ratnayake, the Acting Director General. He indicated that the Dept. would support this project, and may assist in various technical areas.
2. **National Museum of Colombo**: The home of the manuscript since 1901. We have had preliminary discussions with Mrs. Sanuja Kasthuriarachchi, the Acting Director of the Museum.
3. **Sri Lankan Ministry of Buddhasasana**: We have met with Prof. Nimal de Silva, adviser to the Minister, who greeted our project favorably.

4. **University of Sri Jayawardenepura**: Ven. Prof. Medagoda Abhayatissa has lent his support to the project in recognition of its importance for Pali studies.
5. **Bhante Sujato of SuttaCentral**: The author of this document! I have a special interest in the digital aspects of the project, due to my extensive experience managing digital texts for suttacentral.net.
6. **Heshan Karunaratne**: IT support.
7. **Yalith Wijesurendra**: Project coordinator.

6 Funding

Funding for the project has not yet been discussed in detail. SuttaCentral would delighted to sponsor the commercial typing of the manuscript. Other costs, especially those involving the handling and assessment of the physical manuscript, need to be determined with the relevant institutions.

7 Licensing

One of the distinguishing features of the Buddhist tradition is that it has always been freely available. The Buddha did not have the "closed fist of a teacher". Rather, he made his teachings as widely available as possible. The Sangha, as traditional custodians of the texts, have followed this example, with the support of the Buddhist lay community. Hence I would strongly recommend that all resources relating to this project be made freely available with no copyright restrictions.

First to clarify the potential scope of copyright in this project. I am not familiar with the licensing laws of Sri Lanka. However, in most jurisdictions it would be the case that the *images* taken of the manuscript are subject to copyright, as they are considered to be an original creation. However the *text* is not: it is an ancient text in the public domain, and any copyright claim is legally void. Thus the typed digital text, regardless of the work and funds it took to make, would not fall within the scope of copyright, as it is not an original creative work. Assuming the Sri Lankan situation follows these principles, it would be theoretically possible to claim copyright over the images, but not the text.

However, the fact that something *can* be placed under copyright does not mean that it *should*. If the project is placed under a restrictive license, or even if it has no clear license, it will restrict the capacity of scholars to build upon and apply the work. There is a large body of evidence indicating that the influence of work is significantly reduced when it is restricted by copyright. Here are some representative quotes.

Abhishek Nagaraj (UC Berkely): *Does Copyright Affect Reuse? Evidence from Google Books and Wikipedia*[13]

> While digitization has greatly increased the reuse of knowledge, this study shows how these benefits might be mitigated by copyright restrictions. … I find that, while digitization encourages knowledge reuse, copyright restrictions reduce citations to copyrighted issues of Baseball Digest by up to 135% and affect readership by reducing traffic to affected pages by 20%.

Jodie Griffin (Staff Attorney, Public Knowledge): *The Economic Impact of Copyright*[14]

> Current copyright provisions in U.S. law often prevent libraries and archives from preserving the copies of works, particularly when the copyright owner is not know or cannot be found. This ultimately weakens the ability of libraries and other cultural

[13] <http://abhishekn.com/files/copyright_nagaraj.pdf>

[14] <https://www.publicknowledge.org/files/TPP\char`%{}20Econ\char`%{}20Presentation.pdf>

> institutions to preserve the cultural heritage of our society and make that heritage accessible for future generation.

Paul J. Heald (University of Illinois College of Law): *How Copyright Keeps Works Disappeared*[15]

> A random sample of new books for sale on Amazon.com shows more books for sale from the 1880s than the 1980s. Why? This paper presents new data on how copyright stifles the reappearance of works.

The last paper includes a particularly telling graphic. Works published before around 1920, which are freed from the suppressive effects of copyright, are massively more popular. The dip in the graph below indicates the suppressive effect of copyright.

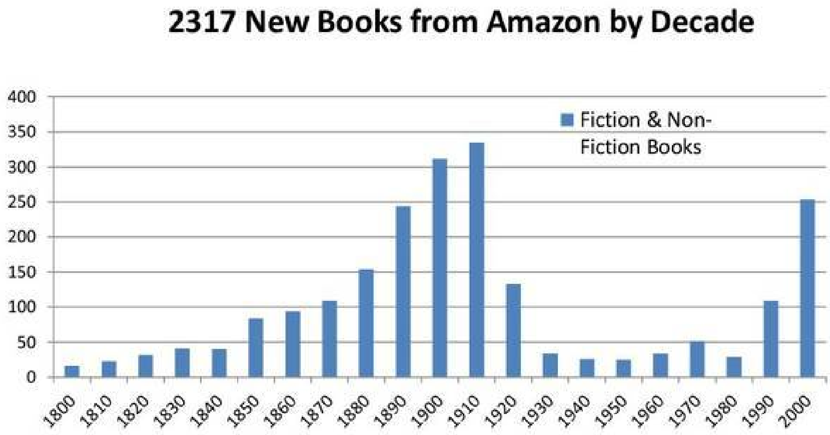


Figure 4: The effects of copyright

I have repeatedly encountered such difficulties in dealing with works on SuttaCentral. It is unfortunately common in the field for

[15] <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2290181>

works to be published under restrictive licenses, even in cases where such licenses have no legal standing. Unclear licensing can be just as destructive. To give an example from Sri Lanka, when seeking to publish the Buddha Jayanthi translation of the Pali Tipiṭaka, I was advised *not* to ask for permission, as there was no clear path towards gaining permission. Eventually we decided to use another translation.

This whole situation is unnecessary. No-one creates Pali editions motivated by profit, so there is no reason to apply a license. The Buddha was right: making the teachings freely available for all creates the greatest benefit for the greatest number, and helps ensure the survival and spread of the teachings. We should follow his example.

> **All text, images, and other resources created by this project should be explicitly dedicated to the Public Domain via Creative Commons Zero (CC0).**

8 How to proceed

Here I will outline what I envisage as the scope of the project. This will have three aspects:

- Divide the project into task modules, each under the responsibility of a specific partner.
- Describe a flow or process for completing each module.
- Present a model for digital management of the project.

9 Task modules

In order to achieve the project efficiently and in good time, it is important to break the project down into clear subtasks, each with an agreed body or person to take responsibility. Here is the initial proposal of this.

9.1 Physically examine the manuscript

The manuscript appears to be in good condition, but it must be carefully examined before taking any action. This must be done by an experienced expert in the field. Any recommendations must be followed so as to ensure the manuscript is not damaged by the digitization process.

From the extant microfilms, it appears as if the majority of the text is quite readable. The top and bottom lines on some leaves are damaged, as are other miscellaneous cases. I would guess that 80%–90% of the text is recoverable.

> **Responsible party:** Department of Archaeology together with National Museum.

9.2 Scan the manuscript

The manuscript must be scanned carefully by qualified staff using good quality modern equipment.

There is a microfilm of the manuscript, but this is not a good solution. Microfilm is an outdated technology, and it is difficult to find equipment to use it. In addition, while the microfilmed images are reasonably good, they will not be of quality comparable to that of modern scans. In the images[16] folder of the Github repository for this project you can see a set of images made from the microfilms, made by Bhikkhu Nyanatusita. While the images are of good quality, the limitations of the source are readily apparent. Compare the quality of a microfilmed image with that of a manuscript image.

Worse, several of the microfilm images appear to be so badly corrupted as to be virtually unreadable.

In this kind of work it is essential that the typists and proofreaders have access to the highest possible quality images so as to make out

[16] `<https://github.com/sujato/cullavagga/tree/master/images>`

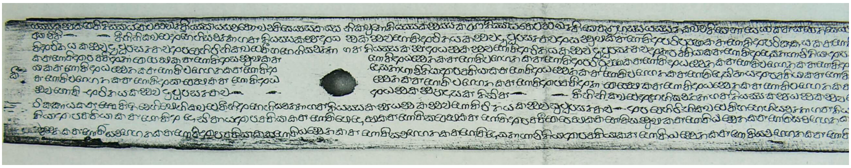


Figure 5: Microfilm image of one leaf of DP-CV

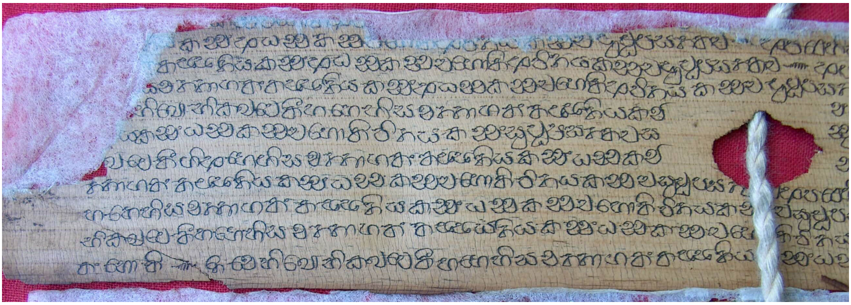


Figure 6: Photograph of one leaf of DP-CV

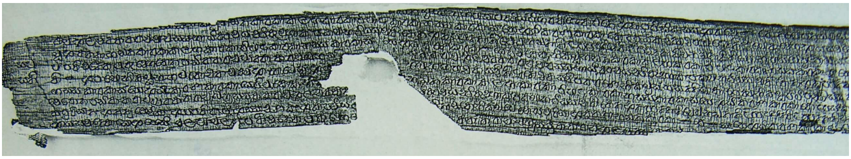


Figure 7: Unreadable leaf of DP-CV

the many obscure or damaged characters. Thus I recommend that, while the microfilm images may be useful for getting an initial idea of the text, the full text should be properly scanned.

The images must be saved in as high resolution as practical. The basic image format should be lossless, which rules out `jpeg`. `png`, which uses lossless compression, is a better option.

The images must be saved using a sensible naming system. Since there is only one manuscript, I propose that we use a simple increment to number the pages: `dpcv-001`, `dpcv-002` and so on. This might seem like a trivial detail, but maintaining a clear and consistent naming convention from the start can avoid a lot of headaches later on.

In addition to scanning the images, it would be advisable to also print a few sets of images. These should be printed professionally, with high resolution on good quality paper using ring-binding or similar. Such printed copies, which need not be great in number, would be useful for typists, rather than having to look at a screen; and in addition, print is a more lasting media than digital.

Responsible party: Department of Archaeology.

9.3 Engage an epigraphic expert to assess the script

The script is early modern Sinhalese script. It is for the most part readily legible, and anyone familiar with modern Sinhalese should be able to read it, with a little work to become familiar with a few archaic forms. However, it is important that the exact nature of the script be defined as clearly as possible.

The epigraphy has already been ably discussed by Fernando, and perhaps this may be adequate. However, I would suggest that an expert in Sri Lankan epigraphy should be engaged to review Fernando's work and further study the script.

The examination should aim to provide an independent estimate of the date of the script, reviewing Fernando's conclusions. Even though we have no reason to doubt the date of the manuscript as given in the colophon, epigraphic evidence can provide important corroboration.

In addition to clarifying the date of the manuscript, epigraphic examination should ascertain any further features of interest in the manuscript, especially such details as may be of use to our typists. We assume that, with a little training, typists and proofreaders will be able to read the manuscript. The epigraphic assessment can verify this and advise as to any potential issues of confusion. Ideally, the epigraphic expert would meet with the typists to train them.

Responsible party: Department of Archaeology

9.4 Carbon date the manuscript

As noted above, there is at present no reason to doubt the dating of the manuscript. However historical dates are always uncertain and it is wise to confirm them using multiple independent methods wherever possible. Carbon dating is a widely used and relatively reliable method of dating ancient artifacts.

In 2007, radiocarbon dating of Gāndhāri manuscripts in Kharoṣṭhī script placed the oldest of them around 75 CE, making them the oldest extant Indian manuscripts. This project, as well as giving a firm historical context, attracted a great deal of publicity, raising the profile of Buddhist manuscript study. (See Mark Allon, et al.)

I propose that we carbon date the DP-CV, with the aim to confirm or refute the ascription of the manuscript to the 13th century. Together with the colophon and the epigraphic assessment, this will provide a solid basis for dating the manuscript.

The commercial cost of carbon dating is about US $500–$1000, so it is not prohibitive. One possible location to use is the Australian Nuclear Science and Technology Organisation (ANSTO)[17] at Lucas heights, who have done the carbon dating for Gandhari manuscripts, organized by Mark Allon. There is a good possibility that testing can be done here under a grant program that would cover the cost.

Responsible party: Department of Archaeology, possibly together with Mark Allon at Sydney University through ANSTO.

[17] <http://www.ansto.gov.au/ResearchHub/OurInfrastructure/acceleratorsciencecentre/Radiocarbondating/index.htm>

9.5 Have the manuscript typed and proofread

This is the most critical part of the project, and the one that will take the most time and attention. It is crucial that the manuscript be typed with the utmost fidelity. The typists must avoid any temptation to reconcile or correct the readings. It is precisely the differences between DP-CV and modern texts that is of interest.

In typing, it is not necessary to know the language. One of the issues I have encountered in doing this sort of work is that experienced scholars are usually busy and have many projects. So if we expect them to devote many hundreds of hours to a project like this, we may have to wait a long time. Fortunately, much of the work can be done at a reasonable price by less experienced workers. We can then use the scholars' time efficiently by having them review the final work. Thus I propose:

> **Typing to be undertaken by a commercial typing firm, with proofreading done by Pali scholars.**

In a project such as this it is common to have the initial typing done *twice* independently, and the two versions compared. This, as scholars of the Theravada tradition would know, echoes the technique that legend says was used by Acariya Buddhaghosa. It is said that after writing the Visuddhimagga, the *devas* took it away, and he had to write it again; and then a third time also. Finally all three versions were compared side by side, and not one character was found to differ.

We are lucky these days to have very powerful "diffing" software, which can do the job of the *devas*, examining two texts and pinpointing any differences between them. I will describe below how this can be accomplished from a technical perspective.

From a project management perspective, I propose the following:

1. Type the manuscript twice.
2. The two versions are diffed, the differences compared with the manuscript images, and the two versions merged as one.

3. The merged text is further diffed against a modern edition of the same text, and any remaining typos or issues checked once more against the manuscript.

This should produce an accurate text. Once an accurate version is achieved, the final text can be supplied to scholars for perusal at their convenience. Any future corrections can be incorporated at any stage.

The target script for the typing should be latin, as this has become the international standard. The ISO 15919[18] conventions should be followed. For the convenience of the typists, it may be easiest to type in Velthuis[19] and convert to ISO 15919.

9.6 Publish digitally and in print

The primary medium for Pali publishing these days is online. Digital text is highly suited to Pali, as it enables us to deal with the large quantity of often obscure works. We will publish the text on SuttaCentral. Our platform will enable a reader to compare the text with other editions, and to see the images next to the text.

In addition, anyone else who wishes to publish the digital text may do so.

If anyone wishes to produce a paper edition, we can support this. SuttaCentral provides a workflow for exporting texts to PDF suitable for high-quality printing.

Since ISO 15919 supports Pali precisely, it is possibly to losslessly convert it to other scripts, especially those of Indic derivation. We do this automatically on SuttaCentral. So anyone who wants to publish in Sinhala or Devanagari script may do so.

[18] <https://en.wikipedia.org/wiki/ISO_15919>

[19] <https://en.wikipedia.org/wiki/Velthuis>

9.7 Academic documentation

As the project is of considerable interest to Pali and other Indological scholars, it should be described in international journals. Such essays should document various aspects of the project, including:

1. Linguistic discussion.
2. Historical context.
3. Project methodology.
4. Artwork on the covers.
5. Investigation as to the provenance of the manuscript, and circumstances of its finding and preservation.

9.8 Publicity and popular awareness

The National Museum and/or other bodies may wish to publicize the work to raise awareness of Sri Lanka's textual history. This might consist of improvements to the museum display, website documentation, leaflets, and so on.

10 Flow for completing each module

Here I set out a process for ensuring the efficient completion of the tasks. The principle is that, so far as possible, tasks may be completed in parallel. This means that, unless it is really necessary, none of the team members has to wait for other team members to finish their work. This process is inspired by task management in the modern "Agile" process for teams, which empasizes "release early and often".

Many of these tasks are of intrinsic value, even if later tasks are not completed. For example, having scanned images available is useful in and of itself, regardless of whether we have a digitized text. So there is no need to wait for the digitization process to be completed before releasing the images. Similarly, carbon dating or epigraphic assessment are of intrinsic interest. These details can be released, also, in a timely manner. Making information publicly available in this way increases interest in a project and demonstrates a commitment to accountability and transparencey.

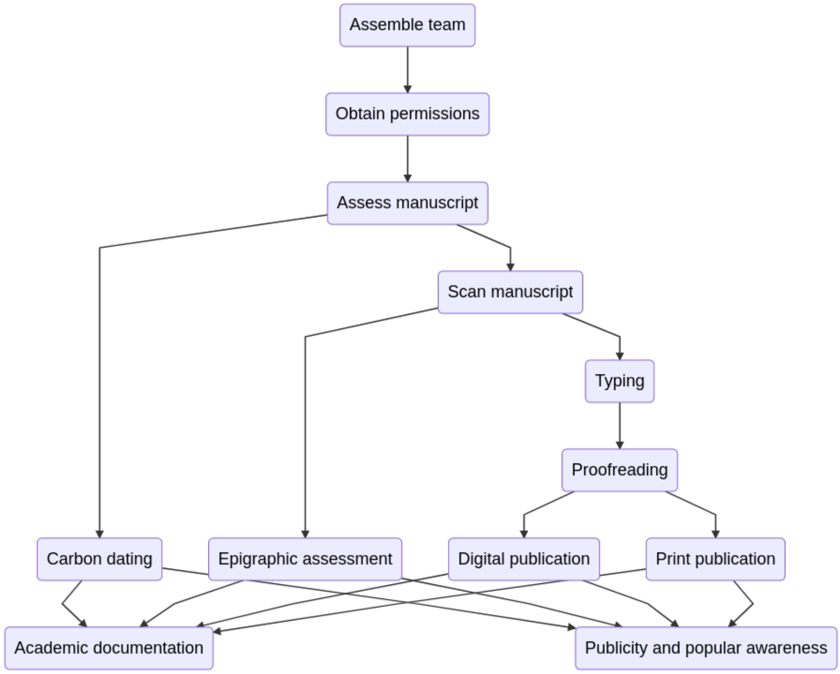


Figure 8: Task module flow

11 Digital Strategy

We live in a digital age and the primary product of our work is digital media. Thus the manner in which we undertake our project should reflect best practices for digital media. While this may seem obvious, it is, unfortunately, very much the exception in the field.

Consider our source manuscript. It tells us it was copied by Sumedha Mahāthera on the instructions of the Medhaṅkara Mahāsāmi as part of a donation of giving one book to each monk in Sri Lanka with the patronage of King Parākramabāhu.

Now consider the state of many of our digital texts. Who were they typed by? Who organized and checked the work? Who was the sponsor? What was the method by which the work was done? We often know none of these things. The texts are produced, ascribed to an institution, and published in a digital form with little or nothing to tell us how they were made. We have better information about a 13th century manuscript than we do about most of our digital texts.

Even worse, as a result of poor digital management strategies, it is unfortunately common for years of work to be lost. The Pali text used on SuttaCentral, for example, was rescued from a project by the Dhamma Society of Bangkok. Sadly, the parent organization fell apart and their online presence disappeared. The text, to which they had devoted years of proofing and preparing, vanished. It was only by sheer luck that we located a source that had been copied by a monk. In another case, years of work annotating and preparing a digital version of a Pali dictionary in a German University was permanently lost, as the IT manager resigned and left without sharing the passwords. More recently, a project to scan and collate manuscripts in Sri Lanka by a Thai organization resulted in many scans being made, but none of them are publicly accessible. For all intents and purposes, they may as well not exist. Prof Lewis Lancaster's article "Digital Input of Buddhist Texts" contains several more such sad stories of work gone to waste.

We can do better. What, we should ask, is the gold standard for creating and maintaining digital projects? In my view, the answer is this: it is what programmers use for their own projects. What we are producing is, in fact, just code. At its core it is binary data, no different from what programmers produce. Thus I propose that we use the state-of-the-art version control offered by Git to manage our digital assets.

11.1 Use Github for all assets

For many years, SuttaCentral has managed its code and source texts at Github. Github is a public cloud service for hosting code and files using the open-source Git version control system. Github is extremely stable and secure, being used by major software companies such as Google, Microsoft, and Facebook for their own code.

The great advantage of using Github is that it preserves every detail about *what* changes are made, *who* changed it, *when* they changed it, and (optionally but recommended) *why* they changed it. These are recorded in an indelible and publicly accessible form on Github. In a Github repo, *nothing is ever deleted*. So if any mistake is made, no matter how small or large, we can always revert it back to the prior state. This means that we can preserve a forensic level of detail and accuracy in all our doings, setting the highest standard for reliability and transparency.

Take the Bendall-CV project undertaken by SuttaCentral as an example. You can see the full list of commits here[20]. Each of these tells you what work was done by whom and when. Clicking on any of these, to pick a random example[21], shows you exactly what changes were made in that commit. You can also see a handy summary of contributers to the code[22], graphs showing the project timeline[23], and so on.

Github is free and simple to use. Anyone with access to the Github account can add or alter assets. No special software is needed; it can be done simply via a browser. For team members, it will be useful to set up a local version of the Github repository. (In Git terminology,

[20] <https://github.com/sujato/bendall-cv/commits/master>

[21] <https://github.com/sujato/bendall-cv/commit/e3378b9866938c592f59593838a576189656f874>

[22] <https://github.com/sujato/bendall-cv/graphs/contributors>

[23] <https://github.com/sujato/bendall-cv/graphs/commit-activity>

a “repository” or “repo” is a folder that contains a set of files for a project.) Local tech support can set this up; because we’re using the same tools that programmers use themselves, pretty much every programmer will know how to do this.

Since Github is designed as a workshop for creating programs, any data and files may be exported (“forked”) into other applications. Thus the data may be made use of by many different projects. In addition, this encourages the creation of multiple redundant copies, thus ensuring the survival of the information.

Thus I propose:

> **All digital assets of the project—including text, images, reports, and essays—be maintained in a single Github repository.**

11.2 Segmenting the text

Segmenting of text refers to breaking it into meaningful semantic segments, for example, a sentence. This aids looking up the appropriate segment, and also helps computers process the difference between texts.

Segments will almost always be the same in different editions, since variations are usually at a smaller scale (a word or letter).

SuttaCentral uses a segmented text based on the Mahāsaṅgīti (MS) edition of the Sixth Council. Let us use this as the reference edition. Here is a sample of a segmented text in SuttaCentral’s system, taken from the first discourse of the Saṁyutta Nikāya.

```
#. </h1></div><p>
#. <a class="pts1ed" id="pts1ed1.1"></a>
#. <a class="pts2ed" id="pts2ed1.1"></a>
#. <a class="sc" id="sc1"></a>
msgctxt "sn1.1:1.1"
msgid "Evaṃ me sutaṃ—"
```

```
msgstr "So I have heard.”

msgctxt "sn1.1:1.2"
msgid "ekaṃ samayaṃ bhagavā sāvatthiyaṃ viharati"
msgstr "Once the Buddha was staying in Sāvatthī"
```

This contains a rich set of data that is associated with each segment.

1. HTML markup, for presentation on the web.
2. Reference details for multiple published editions.
3. An ID number (`msgctxt`) that uniquely identifies each segment.
4. The Pali text itself.
5. An English translation. (Other languages can be easily added.)

Further information can be easily added, such as variant readings or notes on the text.

Normally, it is a lot of work to create and coordinate such rich data associated with a text. Using inappropriate tools such as word processors, there is no simple way to import the data from one edition to another. However, for us this is trivial. To create our DP-CV edition, we first strip all the data from the reference MS edition, keeping only the Pali text. This gives us the following:

```
Evaṃ me sutaṃ—
ekaṃ samayaṃ bhagavā sāvatthiyaṃ viharati
```

Even though there is no metadata at all, we can keep the segments coordinated with the original file by simply counting the line numbers: one line = one segment.

When our typists are working on the DP-CV, they will keep one window on their computer to see this reference edition. Each time they come to a new segment, they mirror the reference edition by pressing`Enter` to create a new line. At the end of the project we can import the new text into the same framework as the reference edition. By this extremely simple method we can integrate our new edition with all the rich associated data from the reference edition.

11.3 Text input

The text of each manuscript is to be inputted as accurately and literally as possible. There is no editorial input; no corrections, identifications of variants, restoration of text, and the like. Texts are romanized, but without adding punctuation, capitals, or any other features. This keeps the job of the typist as simple as possible.

Some editorial intervention will be necessary, however, in the case of unclear or missing portions of the manuscript. Unclear *akṣaras* can be marked with [brackets] or some similar convention, while missing *akṣaras* can be indicated with `+`.

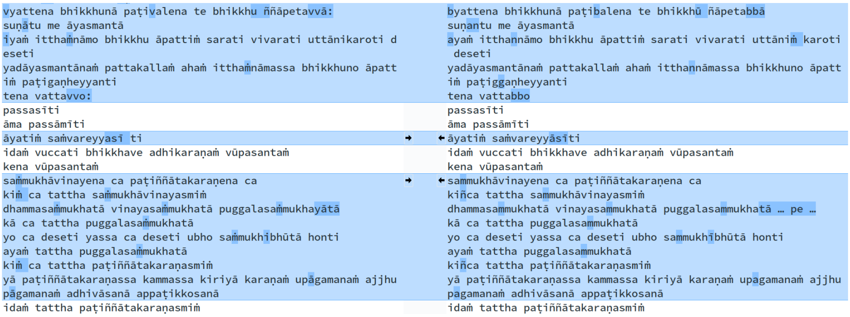
I recommend using the open-source text editor Atom for this process. It can be used on any operating system, and comes with native Github integration (in fact it is built by Github). Whenever the typists save their work it will be automatically synchronized with the central repository on Github. SuttaCentral has created a plugin to enable easy typing of Pali on Atom.

Here is how the text is created.

1. Create a new text file with the ID of the appropriate chapter, let us say `kd1`.
2. Type exactly what is on the page.
3. At the end of each line in the manuscript, insert an arbitrary glyph, let us say `$`.
4. At the end of each page in the manuscript, insert a (different) arbitrary glyph, let us say `#`.
5. Text is segmented by adding a new line (i.e. hit `Enter` for each new segment.)

That's all that is required to create a new digitized edition. There is no need to write line and page numbers in the text; they can be calculated at the end of the project.

As mentioned above, I recommend having the source text typed twice by independent workers. Then the two versions can be "diffed" and the differences resolved. If you're not familiar with how diffing works, here is a sample from two editions of the Cullavagga: on the left the 9th century Nepalese manuscript (Bendall CV) and on the right the text from the Mahāsaṅgīti edition.



```
vyattena bhikkhunā paṭivalena te bhikkhu ññāpetavvā:
suṇātu me āyasmantā
iyaṁ itthaṁnāmo bhikkhu āpattiṁ sarati vivarati uttānikaroti d
eseti
yadāyasmantānaṁ pattakallaṁ ahaṁ itthaṁnāmassa bhikkhuno āpatt
iṁ paṭigaṇheyyanti
tena vattavvo:
passasīti
āma passāmīti
āyatiṁ saṁvareyyasī ti                                          →
idaṁ vuccati bhikkhave adhikaraṇaṁ vūpasantaṁ
kena vūpasantaṁ
saṁmukhāvinayena ca paṭiññātakaraṇena ca                        →
kiṁ ca tattha saṁmukhāvinayasmiṁ
dhammasaṁmukhatā vinayasaṁmukhatā puggalasaṁmukhayātā
kā ca tattha puggalasaṁmukhatā
yo ca deseti yassa ca deseti ubho saṁmukhibhūtā honti
ayaṁ tattha puggalasaṁmukhatā
kiṁ ca tattha paṭiññātakaraṇasmiṁ
yā paṭiññātakaraṇassa kammassa kiriyā karaṇaṁ upāgamanaṁ ajjhu
pāgamanaṁ adhivāsanā appaṭikkosanā
idaṁ tattha paṭiññātakaraṇasmiṁ
```

```
byattena bhikkhunā paṭibalena te bhikkhū ñāpetabbā
suṇantu me āyasmantā
ayaṁ itthannāmo bhikkhu āpattiṁ sarati vivarati uttāniṁ karoti
 deseti
yadāyasmantānaṁ pattakallaṁ ahaṁ itthannāmassa bhikkhuno āpatt
iṁ paṭiggaṇheyyanti
tena vattabbo
passasīti
āma passāmīti
← āyatiṁ saṁvareyyāsīti
idaṁ vuccati bhikkhave adhikaraṇaṁ vūpasantaṁ
kena vūpasantaṁ
← sammukhāvinayena ca paṭiññātakaraṇena ca
kiñca tattha sammukhāvinayasmiṁ
dhammasammukhatā vinayasammukhatā puggalasammukhatā … pe …
kā ca tattha puggalasammukhatā
yo ca deseti yassa ca deseti ubho sammukhībhūtā honti
ayaṁ tattha puggalasammukhatā
kiñca tattha paṭiññātakaraṇasmiṁ
yā paṭiññātakaraṇassa kammassa kiriyā karaṇaṁ upagamanaṁ ajjhu
pagamanaṁ adhivāsanā appaṭikkosanā
idaṁ tattha paṭiññātakaraṇasmiṁ
```

Figure 9: Ratana Sutta diff

The diff engine compares the two texts and highlights any differences. This is made easy when using a segmented text, as the differences are kept to each pair of lines.

In this example, you can see several examples of the kinds of problems we are aiming to illuminate through comparing manuscripts. There are a fair number of differences, although they are mostly minor spelling variations. Sometimes these are consistent (eg. *vattavvo* vs. *vattabbo*), with readings found nowhere in modern Pali. Such cases may reflect a dialectical evolution within Pali, or perhaps a geographical difference. In other cases (eg. *kiṁ ca* vs. *kiñca*) we see the same kinds of variations found within modern texts. In still other cases (eg. *paṭigaṇheyyanti* vs. *paṭiggaṇheyyanti*) the difference appears to be a simple error on the part of the scribe.

So the job of the proofreader will be to resolve each case where the digital texts diverge, referring back to the original manuscript as the final authority.

I recommend that this process be done twice: once by comparing the two typed files of the DP-CV edition, and again by comparing the merged DP-CV file against the reference Mahāsaṅgīti edition.

Once the typing is finished, we calculate the line and page numbers of the manuscript based on the inserted glyphs. Then this reference data can be added to the set of data for all the editions. The newly typed text can be imported into the metadata structure as seen above, and then used for display, reference, translation, and so on.

12 Beyond this project

Once the project is completed and the text digitized, the text may be used in various applications. While the text is reasonably safe on Github, we should try to make it available on sites such as GRETIL[24] and the Internet Archive[25]. The more widely disseminated the text is, the harder it is for it to become entirely lost.

The text and images may be consumed in applications such as SuttaCentral or any other interested project, and we should try to support such efforts.

There are many more manuscripts in Sri Lanka, which are also crying out for similar digitization. SuttaCentral's area of interest is the early canonical texts, and there are at least two such manuscripts of comparable value to DP-CV.

- The National Museum houses a manuscript of the Saṁyutta Nikāya, whose colophon says that the original copy was copied in 1412 CE by Maṅgala Thera of Sunētra Dēvi Pirivena of Pepiliyana. (Nyanatusita, 371)

[24] <http://gretil.sub.uni-goettingen.de/>

[25] <https://archive.org/>

- A 13th century manuscript of the Mahāvagga—the companion text to the Cullavagga—is at the Vidyalankara Pirivena in Kelaniya. (Nyanatusita, 369)

Taking these together with the DP-CV—and assuming they are reasonably complete—they make up over a third of the early Pali canonical prose texts. To establish such a sizable portion of the Tipiṭaka on such early manuscripts would be a significant contribution to Buddhist studies.

There are a number of manuscripts of a similar age, which are also deserving of preservation. Since these are of commentarial literature they are beyond the scope of SuttaCentral. Nonetheless, we would be happy to offer technical support for any such efforts.

I believe that the methods set forth here will achieve an accurate and usable text with a reasonable amount of effort, and would encourage all parties involved to continue the work.

13 Bibliography

- Allon, M., Salomon, R., Jacobsen, G., Zoppi, U. (2007). *"Radiocarbon Dating of Kharosthi Fragments from the Schøyen and Senior Manuscript Collections"*. In Jens Braarvig (Eds.), *Buddhist Manuscripts in the Schøyen Collection*, vol. iii: pp. 279–291. Oslo, Norway: Hermes Publishing.
- De Silva, W.A.. *Catalogue of Palm Leaf Manuscripts*, Colombo 1938, No. 2363
- Fernando, P. E. E. 1982. *"Note on Three Old Sinhalese Palm Leaf manuscripts"*, The Sri Lanka Journal of the Humanities 8, no. 1/2: 146–157.
- Lancaster, Lewis. *"Digital Input of Buddhist Texts"* in Damien Keown, Charles S. Prebish, *Encyclopedia of Buddhism*, pp 288–296, Routledge 2013.

- Nyanatusita, Bhikkhu *“Pali Manuscripts of Sri Lanka”*, in *From Birch Bark to Digital Data: Recent Advances in Buddhist Manuscript Research*. Papers Presented at the Conference “Indic Buddhist Manuscripts: The State of the Field”, Stanford, June 15–19, 2009, edited by Paul Harrison and Jens-Uwe Hartmann, and published by Österreichische Akademie der Wissenschaften Wien in 2014.
- Sujato, Bhikkhu and Brahmali, Bhikkhu. *The Authenticity of the Early Buddhist Texts*. Oxford Center for Buddhist Studies, 2014.