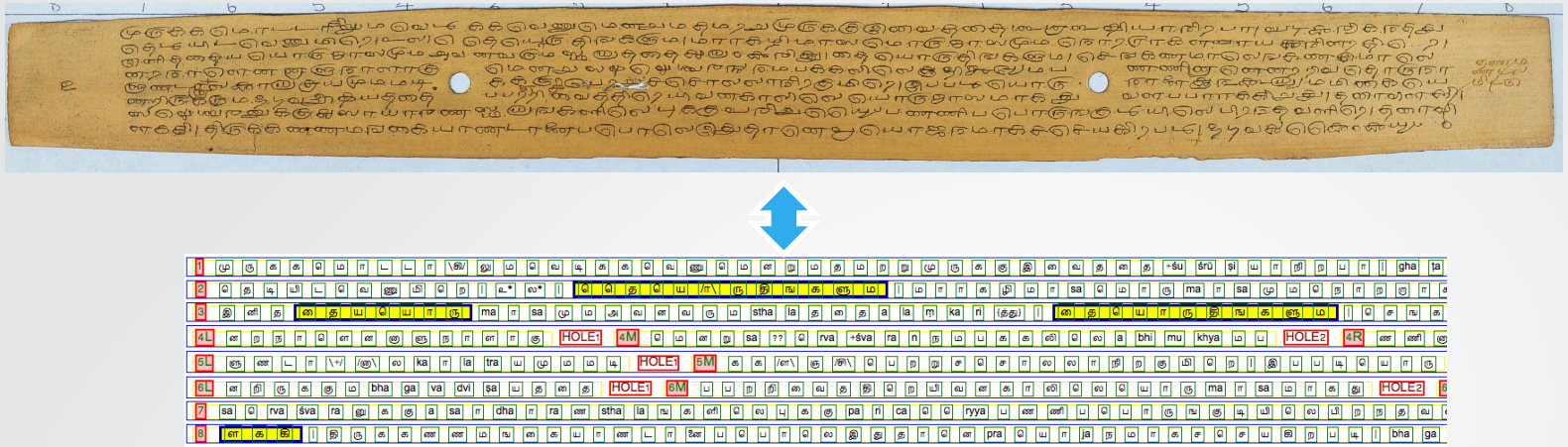


Preserving Provenance

Interlinking Digital Representations of Manuscripts

Vinodh Rajan *vrs3@st-andrews.ac.uk*

Dr Mark-Jan Nederhof *mn31@st-andrews.ac.uk*

School of Computer Science

University of St Andrews

Scotland, UK

University of St Andrews

FOUNDED 1413

University of St Andrews
FOUNDED 1413

Overview

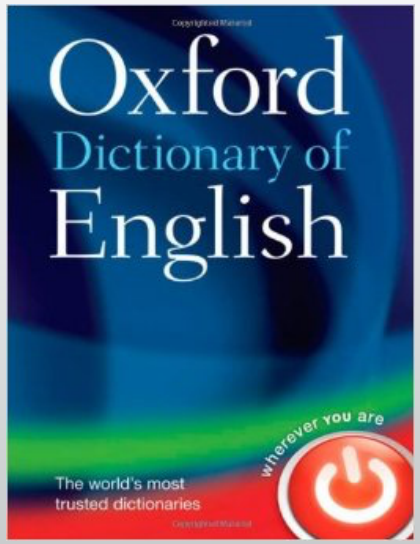
- Data Provenance
- Digital Representations of Manuscripts
- Interlinking the Representations
- PhilologEg
- Case Study: Tamil Manuscripts
- Questions & Discussion

University of St Andrews
FOUNDED 1413

Provenance

The Oxford dictionary defines the word "Provenance" as follows:

"THE **PLACE OF ORIGIN** OR EARLIEST KNOWN HISTORY OF SOMETHING"



University of St Andrews
FOUNDED 1413

Data Provenance

- Provenance can be easily associated with material objects
- What about data?
- How can the "provenance" of digital information be captured?
- What does it actually mean by the term "Data Provenance"?

Data Provenance

- Digital data items are more often than not "derived" data
- Created by a process of derivation from other items (either digital or non-digital) (which is then edited and stored as required)
- The provenance of a data item includes information about the processes and source data items that lead to its creation and current representation

Glavic, B., & Dittrich, K. R. (2007). Data Provenance: A Categorization of Existing Approaches.
Buneman, P., & Davidson, S. B. (2010). Data provenance-the foundation of data quality.

Data Provenance

- Why is data provenance important?
- As more and more derivations occur, errors might slip in later stages
- It must be possible to go back to the source to verify/validate
- To do that , one needs to understand how a particular data got its current form

Buneman, P., Khanna, S., & Tan, W. C. (2000). Data provenance: Some basic issues.

University of St Andrews
FOUNDED 1413

Digital Representations of Manuscripts

- The original manuscripts are usually physical
- Two different digital representations are subsequently derived
  - Image-based representations (Scans/Photographs/Facsimiles)
  - Text-based representations (Transcription/Transliteration/Translation)
- Usually text-based representations take priority – as they're more convenient & practical
- Image representations usually have very low emphasis

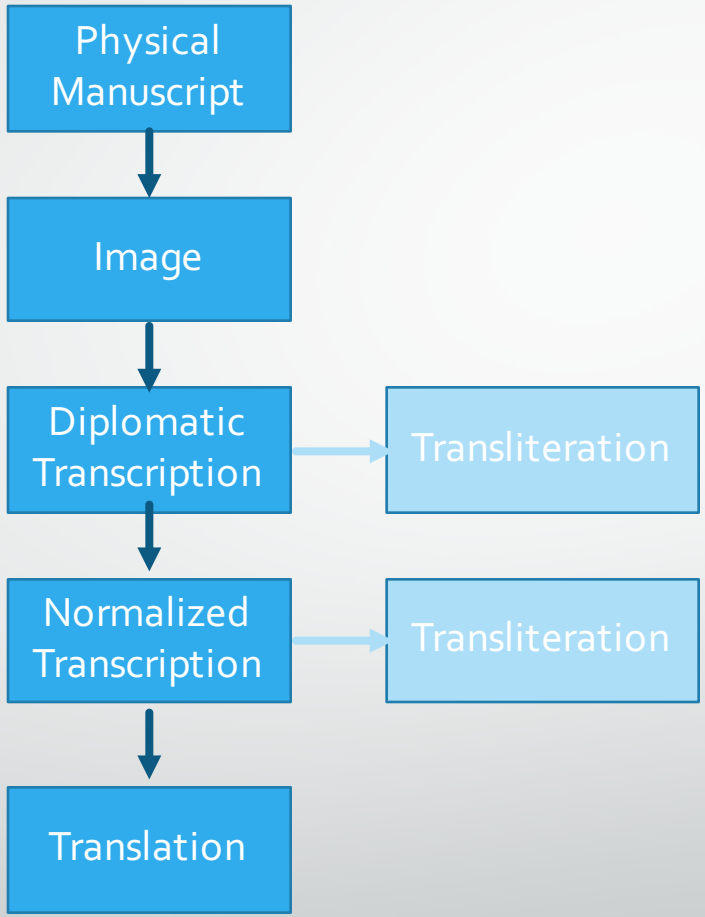
University of St Andrews
FOUNDED 1413

Digital Representations of Manuscripts

- Transcriptions a.k.a textual representations are in a way 'interpretations' of the source artefact
- There may be multiple transcriptions of the same manuscript
- Image representations are free from interpretations and are usually true to the original manuscript
- Image Representations also provide additional contextual information

University of St Andrews
FOUNDED 1413

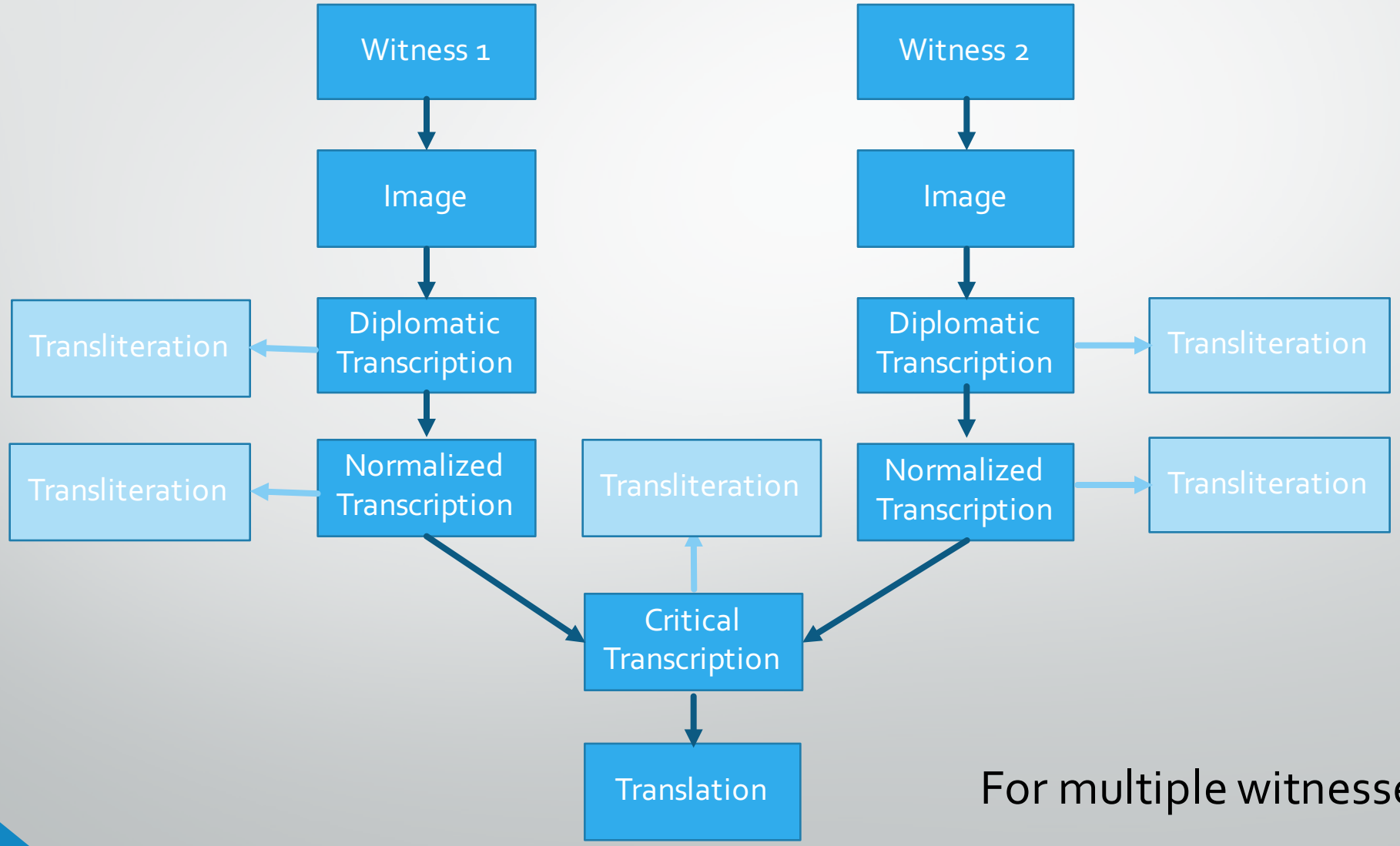
Hierarchy of Manuscript Representations



For a single witness…

University of St Andrews
FOUNDED 1413

Hierarchy of Manuscript Representations



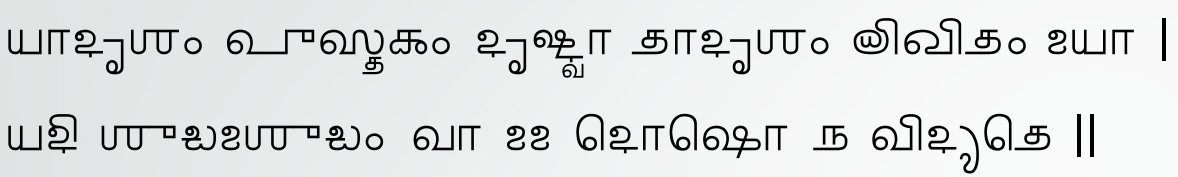
For multiple witnesses…

University of St Andrews
FOUNDED 1413

Need for Data Provenance

- If there is a contentious reading in a translation…
- If we need to confirm a particular reading in a normalized transcription…
- If we have to disambiguate a reading in a diplomatic transcription…
- If we want to decide between two different readings…
- Any moment, we want to check something in any of the layers…

A Scribe's Disclaimer…



yādṛśaṁ pustakaṁ dṛṣṭvā tādṛśaṁ likhitaṁ mayā |

yadi śuddhamaśuddhaṁ vā mama doṣo na vidyate ||

As it was in the text, having seen, so it was written by me

Whether it is correct or incorrect, there is no fault of mine

- Scribal verse from a manuscript of *Karmavibhangopadesha*

Source: Digital Buddhist Sanskrit Canon
(http://www.dsbcproject.org/karmavibhaṅgopadeśaḥ/karmavibhaṅgopadeśaḥ)

University of St Andrews
FOUNDED 1413

Need for Data Provenance

- At some point, there will be a need to investigate the images and the textual representation in relation to each other
- Usually, digital archiving divorces image and textual artefacts
- Hence, matching one against the other, usually requires tedious manual effort
- In case of preparing critical texts, the problem is further complicated due to the presence of multiple witnesses

University of St Andrews
FOUNDED 1413

Need for Data Provenance

- The various representations when taken together provide a rich contextual environment to study/analyze a text
- They're usually seperate and not interlinked i.e individual files
- If integrated, they can aid collation and creation of digital critical texts
- Hence, we need to interlink all these different derivations of manuscripts

University of St Andrews
FOUNDED 1413

Preserving (Data) Provenance

- Through interlinking of the various representations data provenance can be preserved for manuscripts
- Currently, there are no standardized framework to perform the interlinking
- Consequently, there are nearly no tools that can provide an automatic/semi-automatic interlinking of the different representations
- What is required is a integrated environment similar to an integrated development environment (IDE) to interlink texts

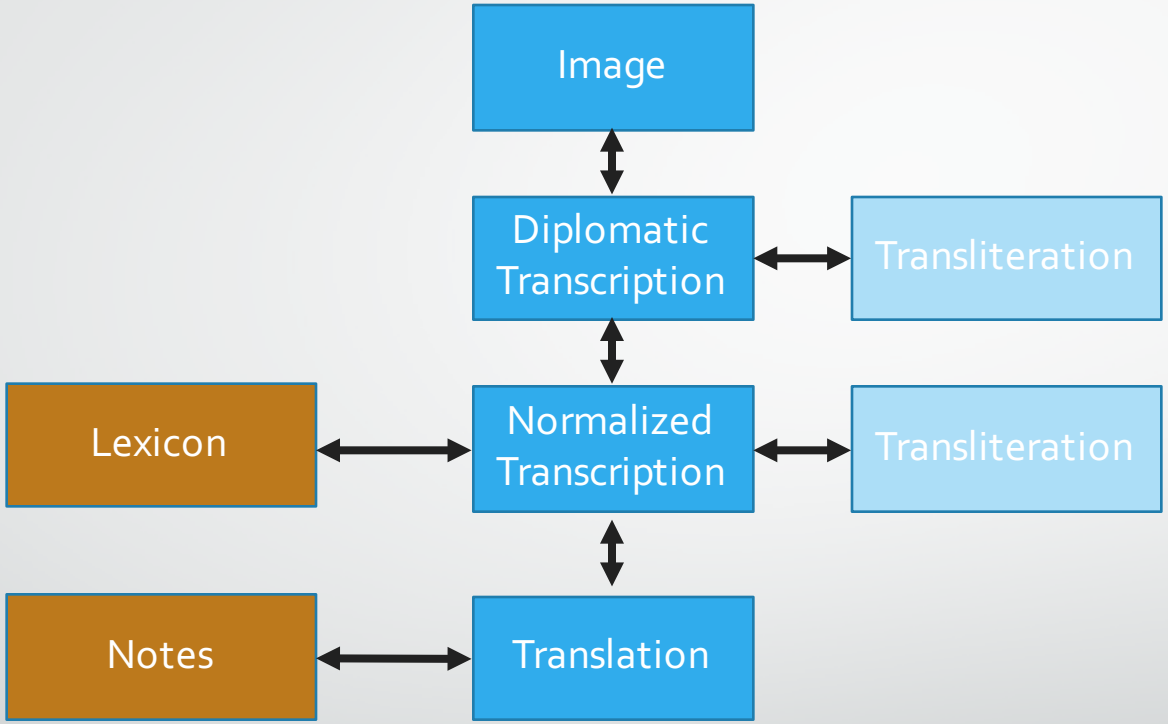
University of St Andrews
FOUNDED 1413

Interlinking Representations

- Adjacent layers are interlinked to each other
  - Image <-> Transcription <-> Translation
- Appropriate meta-data can also be interlinked
  - (e.g) Orthography/Lexical information could be added to transcription
- If we need to check on any successive layers
  - Say we need to double check a translation, the source of the word/phrase can be traced back to the diplomatic transcription and/or image

University of St Andrews
FOUNDED 1413

Interlinking Representations



University of St Andrews
FOUNDED 1413

Research Objectives

- Developing data formats for representing various representations and their interlinks
  - TEI is too verbose, too big and too complicated!
- Developing graphical tools for viewing/editing/linking representations of a text (and also the associated meta-data).
- Importing/Exporting to other formats for representing text representations, including TEI
- Developing tools for lay end-users to interact with the final interlinked text

University of St Andrews
FOUNDED 1413

Research Objectives: Data Formats

- Develop individual XML schemas for the various representations
- Develop notations/schemes to represent the interlinks
- The schemas should be script/language independent
- Develop graphical tools to interact/create/edit the XML files

University of St Andrews
FOUNDED 1413

Research Objectives: Integrated Environment

- An integrated environment to display/edit the different manuscript representations (and also the meta-data)
- The representations will be linked to each other using automatic/semi-automatic methods
- Provide various levels of abstraction and different modes of visualization
- It will have in-built relevant graphical/text processing modules
- It will offer a complete environment to manage all representations of a text

University of St Andrews
FOUNDED 1413

Research Objectives: Tools of Non-Experts

- The integrated tools evironement is targeted mainly for scholars (and other serious users)
- There is a need to develop a tool aimed at non-experts
- Allow non-experts to interact/explore the final interlinked text
- Provide a light-weight version to non-expert users
- It can be used for educational/pedagogical purposes

University of St Andrews
FOUNDED 1413

Research Objectives: Import/Export

- Like it or not TEI (and it's subsets/derivatives) are popular
- There are also other popular custom data formats such as EpiDoc
- We should be able to export or import to all these various data formats if require
- Conversion to other standard data formats such as JSON

University of St Andrews
FOUNDED 1413

PhilologEg

- Originally developed for processing Ancient Egyptian texts
- Provides options to view/edit/store different layers of manuscript representations
- Enables the users to interlink the different layers
- The interlinked text are visualized in the special manner called “Interlinear Form”

University of St Andrews
FOUNDED 1413

PhilologEg: Data Formats

- Each representation is saved as it's own file
- The files can be edited independently
- The interlinks are saved as anchor points referring to the files that represent various layers
- Uses XML and plain-text format to encode the information

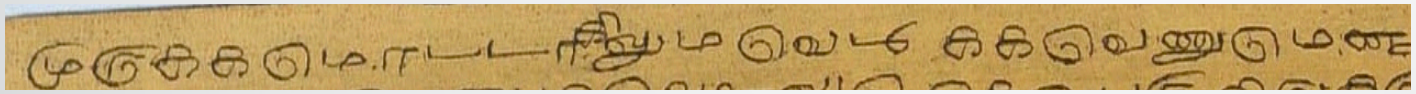
University of St Andrews
FOUNDED 1413

PhilologEg: Interlinear Form

- Interlinear Form is a special way to visualize the interlinks between layers
- The text in the different representations are "anchored" to one another as required
- In the interlinear form, the various representations are aligned to each other
- The software intelligently decides on line-breaks, spacings etc to maintain the correspondences
- The links can also be explicitly seen

Demonstration: Tamil Manuscript

- We have taken a Tamil manuscript (A Manipravala commentary of Tiruvāymoḻi) and have encoded the different layers in PhilologEg



Image

முருகமொடடாகிலுமவெடிககவெணுமென

Diplomatic Transcription

முருக்கமொட்டாகிலும்வெடிக்கவேணுமென

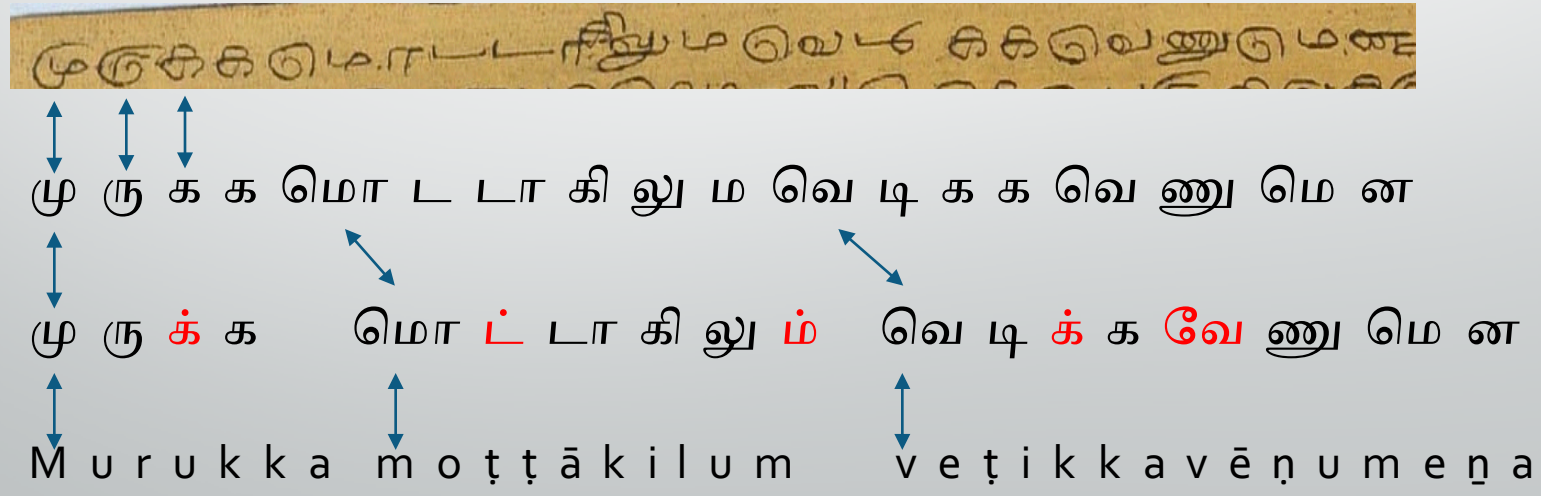
Normalized Transcription

murukkamoṭṭākilumveṭikkavēṇumeṉa

Transliteration

Demonstration: Tamil Manuscript

- Align diplomatic transcription with image
  - Perform character-wise linking
  - Automatic detection of characters in manuscript
- Align diplomatic transcription with normalized transcription
- Align normalized transcription with transliteration



Demonstration: Tamil Manuscript

- The Line 3 of the transcription reads:

இனித்தையொரு ஈஹஸமும் அவனவரும் ஸ்஥லத்தை சுஅலங்கரித்து

iṉittaiyoru māsamum avaṉavarum sthalattai alaṁkarittu

(Hence, in the month of *Tai* avaṉavarum(?) decorating the place [...])

It'd make more sense if the line read "*அனைவரும்*" (aṉaivarum) meaning "everyone" instead.

Let's go back to the source and see what the manuscript says...

University of St Andrews
FOUNDED 1413

Demonstration: Tamil Manuscript

- Pre-Modern Tamil script is very ambiguous
- The person transcribing usually disambiguates the words
- Diplomatic transcriptions are seldom created
- It'll be extremely useful to link the resulting transcriptions with the original manuscript images

Demonstration: Tamil Manuscript

- They system can also be used to auto-generate glyph mappings

மு

ரு

க

க

வ

ம

ர

ட

ட

University of St Andrews FOUNDED 1413

Web Visualization

- The interlinear form can be exported as a web page for non-expert users to interact with the text

https://vrs3.host.cs.st-andrews.ac.uk/demo/

- The entire text is compiled into a web page using only client-side technologies such as HTML, CSS and JS
- Allows interaction with the original manuscript image

University of St Andrews
FOUNDED 1413

To summarize….

- Custom data formats to store the representations
- A method to align different representations of manuscripts
- A method to visualize interlinked texts
- Ability to add and interlink meta-data
- Ability to auto-recognize characters in manuscripts
- Ability to export to other formats such as PDF

University of St Andrews
FOUNDED 1413

Questions

- What are suitable data formats for representing and interlinking various strands of information related to a textual artefact?

- How can the various strands of information be overlaid and displayed to the user to encourage exploration and dissemination of knowledge?

University of St Andrews
FOUNDED 1413

Discussion

- What are all the different layers that could be foreseen?
- How should different layers be interlinked?
- How can the interlinks be visualized innovatively?
- What are the difficulties in manuscript processing?
- Script specific-features that needs to be foreseen?

University of St Andrews
FOUNDED 1413

Discussion

- What is expected in such an integrated environemnt?
- What kind of features do expert users expect?
- What kind of features do non-expert users expect?
- What should be automated and what should not?
- Any other comments/suggestions?

University of St Andrews
FOUNDED 1413

Acknowledgements

- Dr Jean-Luc Chevillard and Dr Eva Wilden for providing the necessary data for the demonstration

𑌇𑌮𑌸𑍍𑌮𑌿𑌂 𑌸𑌤𑌿 𑌇𑌦𑌂 𑌹𑍋𑌤𑌿
𑌇𑌮𑌸𑍍𑌸𑍁𑌪𑍍𑌪𑌾𑌦𑌾 𑌇𑌦𑌂 𑌉𑌪𑍍𑌪𑌜𑍍𑌜𑌤𑌿 [...]

Imasmiṃ sati idaṃ hoti
Imassuppādā idaṃ uppajjati [...]

This being, that becomes;
From the arising of this, that arises [...]

- Assutavā Sutta

Thanks & Questions?

University of St Andrews
FOUNDED 1413