tinker.

Text Transcription in the Humanities, Arts and Social Sciences: Creating readable and computationally analysable data.

Ben McRae 11/12/2018

Key messages:

The main goal of transcription is to create readable, machine readable or computationally analysable content. This will largely depend on what you are transcribing, but the main aim here is for readable and analysable content. For example if your are working from handwritten manuscripts, creating digital and then readable content from the handwriting will open the options up for further analysis. Similarly if you are working from images or notes transcribing the materials into a codifiable from will help to draw out themes and locate patterns in the research.

"Transcription is the process of taking linguistic or musical material in a different modality (e.g., speech, sign or musical performance) and representing it in a written form. In the digital age, the meaning has expanded to include taking images of written material, and copying it in an encoded (and especially searchable) form."   (2018 Honeyman)

This recipe will give you a clearer understanding of what is involved in the preparation for a transcription project, show you some of the tools available to assist you through the process, and make you more aware of the types of outputs and how this can assist future research.

Learning objectives:

- Clear idea of how to prepare materials for transcription
- Familiarity with the tools supported by the Tinker workbench
- Clear understanding of transcription outputs
- Clear understanding of how to annotate and code transcribed resources for analysis

Ingredients:

Pre-reading:

Schöch, C. (2013). Big? smart? clean? messy? Data in the humanities. Journal of Digital Humanities, 2(3), 2-13.
http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/

Other resources:

Resources to be transcribed

- Text - handwritten or typed materials (eg records, interviews, diaries, stories, etc), audio-visual materials (eg interviews, performances, etc)
- Images - images of text, maps

Pre-workshop requirements:

Prior to attending the workshop participants should download Transkribus and register an account.

From the page	https://fromthepage.com/
Transkribus	https://transkribus.eu/
Digivol	https://digivol.ala.org.au/

Method:

To get your materials ready for transcribing you will need to consider how you will search for and count key terms and themes. You will need to consider glossaries of terms to match synonyms and account for the ambiguities of language that arise.

- Data dictionary - are you using abbreviations of codes to describe elements of the resources? If so, organise these elements to ensure consistency of approach in your transcription.
- Style sheets - a style sheet provides you with a standard set up markups and annotations highlight elements of your transcription. *Note - style sheets are more commonly used in audio transcription but the use of consistent markup methods will greatly improve the quality of your output data.
- Glossary of terms - are you using a defined vocabulary in descriptions? This will help you collate key words and group themes found in the text.

Text or behavioural codes -

- Text codes - (phrases, names or places you wish to highlight and collate, names, places, events, companies or businesses etc. Anything where links between seemingly disparate datasets can be made).

The following links provide further information on setting fields and standardised vocabularies in transcription:

https://www.ands.org.au/guides/vocabularies-and-research-data
https://sites.google.com/site/anecdotestyle/style-guides/general-transcription-style-guide
https://www.transcribe.com/transcription-jobs-online/style-guide/

Once you have your resources ready to transcribe there are a suite of tools available for you to use to assist the transcribing process.

The tools all do the basics of transcription but have very definite pros and cons that can benefit your project. Before you start transcribing, look at the types of output files available from the tools to ascertain the right tool and outputs for your project. Most tools will export in XML, CSV, TXT and RTF file types.

Tool	Server	Pros	Cons
From the page https://fromthepage.com/	Server	Very simple to use Can be used to transcribe any type of document Can be used for both crowd transcription and single user transcription projects, however you need to supply the transcribers	Open sourced Maintained by a very small team Not great for inbuilt markup or hyper dense text
Transkribus https://transkribus.eu/	Locally	Excellent capabilities to transcribe and mark up busy resources Was primarily intended for the transcription of hand written materials stemming from a need to study and digitise historical manuscripts Structured outputs from unstructured texts Will in time learn to machine read the documents	Very tricky to set up Will require time and training to learn all the features of tool
Digivol https://digivol.ala.org.au/	Server	Primarily designed for crowd based transcription Can also be used for individual projects Primarily designed for the description and annotation of images. Digivol is best used when consistent structured transcription of materials is required, the tool was designed to transcribe notes and museum records. Open to both individual and institutional projects	Does not allow for markup

Sample activities:

Activity 1

The objective of this activity is to best match your transcription project with the tools available
look at the the resources and materials you have, think about the data you need and then using the links above and the pros and cons on the table choose the tool you think will best meet your needs.

Activity 2

Upload your resources into your chosen tool and investigate how the tool operates.

Notes:

For further information contact your institutions library or eResearch team.

Next steps:

Once you have transcribed text, it will need to be cleaned, curated and then analysed.