Enhancing Digital Skills Applied to History, Humanities, and Beyond: Workshop Activities

20 – 22 June 2016
(Daily 8:30 – 16:30)

This document contains the workshop activities that we will be working our way through in the 'Enhancing Digital Skills' workshops at North-West University. Feel free to read all the way through so you know what's coming, but save the activities for in-class, as we'll be working collaboratively.

We're going to be exploring the new potential of the Internet for scholars, and I'm looking forward to working with you!

Adam Crymble
Lecturer of Digital History,
University of Hertfordshire (England)
(http://www.herts.ac.uk/digital-history)

Editor, *The Programming Historian*
(http://programminghistorian.org)

Email: a.crymble@herts.ac.uk
Twitter: @adam_crymble

Petri Jooste
IT Manager/Lecturer
School of Information Technology
North-West University (South Africa)

Table of Contents

Introduction to the Workshop Series

Welcome to 'Enhancing Digital Skills'. The goal of these workshops is to introduce you to the potential of your computer and the Internet for research.

Workshops are designed to let you experiment with a range of technologies and concepts that are key to digital humanities scholars. Our aim is not to turn you into an immediate expert; instead, by getting a chance to play around with new ideas, I hope you'll have a better idea of which concepts would be most useful for you and therefore where you might focus your future energy. If there is something you find particularly interesting, let me know and I can point you in the direction of resources that can help you continue your learning.

You will be expected to bring with you a genuine curiosity for learning, and a willingness to try and persevere when trying something new. However, **no prior experience or technological aptitude is expected.**

We will use specific examples in our workshops that I hope you find interesting, but what is important here are the skills behind each lesson rather than the specific resources we use. That means the subject matter can be swapped to look at any time or place. By learning those skills, you can then go away and apply them to your own research.

Troubleshooting

You will undoubtedly get stuck from time to time, and you're bound to get frustrated at least once. If you don't, I haven't challenged you enough. If you get stuck, try the following:

1) **Google it lots of different ways**

    The great thing about the Internet is that it contains the answers to just about any technical problem you could have. Chances are someone has had the exact same problem as you, and someone else has written a solution online. You may find some useful tips in Googling for solutions in the following post that I wrote a few years ago:

    Adam Crymble, 'How to Solve Programming Problems if you're learning Programming', *Thoughts on Digital and Public History,* 2014: http://adamcrymble.blogspot.co.uk/2014/04/how-to-solve-programming-problems-if.html

2) **Ask your friends or neighbours**

    If you're sitting next to someone and you can see clearly that they've got further along than you, try the following.

'Hello, my name is X. I see you've made it past this bit of our workshop. Would you mind telling me how you did it?'

We learn together in the digital humanities, so make the most of your colleagues.

3) Ask for help during workshops

There's a reason why we do these workshops in-person. It means the workshop facilitators can be there to help you when things go wrong. We may not always know the answer, but we'll do our best to help.

Day One: The Digital Archive
(Adam Crymble)

Workshop 1 – Adding Value to Digital Records

Learning Outcome:

- To understand what 'digitization' means and the strengths and limits of the computer.
- To think about the needs of end-users when digitizing material
- To push the boundaries of what is possible in a digital realm!

Background:

You have probably come across 'digitised' resources before. Maybe digital images of famous artworks that you've managed to find online. Or digital texts, digital objects, or digital music. In this first workshop we will think about the process of digitization so that we understand the decisions digitisers make along the way and the repercussions of those decisions on what we can do with the finished product. This isn't about the technologies themselves, but about the choices we make. Digitisation is always a matter of compromise. Doing it well can be the difference between a groundbreaking project that changes the way people understand your area of research, or it can be a missed opportunity.

The point of this exercise is to get you to think both about the possibilities of the computer, but also the needs of end-users who are depending on you doing a good job. It's not just the most obvious form of digitization that people will want (eg, not everyone wants a scan of the original from head-on. Someone will want to know what's on the back of an object, even if you don't!)

Part 1: Digitisation as Attributes (20 minutes)

You'll be given an object to 'digitise'

In groups of 2-3 create a list of all of the attributes of that object that you could capture and store in a computer's memory, as well as all of the attributes that you cannot capture. How close is your 'digital' object to the original? What compromises do you have to make? Try to be as creative as possible.

Part 2: Digitisation as Adding Value (30 minutes)

As before, you'll be given an 'object' that someone might reasonably want to use in their research. With your group, brainstorm and write down as many ideas as you can for how a scholar might want to use that object in a study. What research questions might you want to ask? You may be coming from different disciplinary backgrounds, so how would people from your discipline want to use that object? Is that different from your neighbours?

Be creative! Think about the materiality (what *is* the object made of?), the function (what is it *for?)*, the form (what does it *look* like?), the context (who *made it*, *where*, *when*, *how*, and *why*?).

Drawing on what you discovered in the previous exercise about what is possible, what would you have to do to 'digitise' your object in a way that would allow the greatest possible flexibility for researchers wanting to conduct the studies you brainstormed? What approach will you take? Photograph it? Describe it? Measure it? A combination? Something else? How can you prevent your decisions from limiting other researchers' options?

If we have time we'll share some of the best examples.

Workshop 2 – What is XML and why should you care?

Learning Outcome:

- To understand what XML is and in what contexts it's useful
- To gain some experience constructing XML

Background:

XML or Extensible Markup Language is used frequently, in particular by scholars who work with texts. XML is used to tell the computer what various parts of your text mean. Computers can store text but do not understand it. Markup makes it clear which elements contain which types of information. We do this naturally as humans with the ability to read. The computer needs a bit of extra help.

You can 'markup' elements that are of interest to you, or, drawing on our previous workshop, you can think about the potential needs of other scholars too. You might, for example, mark up all place names, or names of people. But equally you might think more abstractly, and mark up all of the nouns, or the stage direction in a play.

Markup has a number of scholarly uses. One of which is to build 'advanced search' or 'browse' options on scholarly websites. If you've ever seen a website that lets you search by 'name' or 'date', it might well have used XML to facilitate that feature. In this workshop we will practice marking up some simple texts to get a feel for the process.

If you decide to pursue this skill further you will want to look into standards such as the Text Encoding Initiative (TEI), which has built a set of rules so that we all mark up our texts in the same way and our results are therefore interoperable.

(TEI CHEAT SHEET)

All XML tags are in angle brackets. Usually they surround the word or words they pertain to, following a pattern like this:

My words I want to <tag>highlight</tag>.

In the above instance 'highlight' has been tagged. Notice the closing tag includes a '/' in it. This tells us it is a closing tag rather than an opening tag. You can tag more than one word at a time.

There are thousands of combinations of TEI tags and attributes. Sometimes it helps to get started with a few simple ones that are particularly relevant to humanities research. Spend a few minutes familiarizing yourself with the

following tags, then move on to Activity 1. Please note the tags are case-insensitive (capital letters don't matter).

<DATE>
Description: contains a date in any format.

<EVENT> (I.E. EVENT)
Description: any phenomenon or occurrence, not necessarily vocalized or communicative, for example incidental noises or other events affecting communication.

<GEOGNAME> (I.E. GEOGRAPHICAL NAME)
Description: a name associated with some geographical feature such as "Windrush Valley" or "Mount Sinai".

<GEOG> (I.E. GEOGRAPHICAL FEATURE NAME)
Description: contains a common noun identifying some geographical feature contained within a geographic name, such as "valley", "mount" etc.

<JOIN>
Description: identifies a possibly fragmented segment of text, by pointing at the possibly discontiguous elements which compose it.

<OCCUPATION>
Description: contains an informal description of a person's trade, profession or occupation.

<PERSNAME> (I.E. PERSONAL NAME)
Description: contains a proper noun or proper-noun phrase referring to a person, possibly including any or all of the person's forename, surname, honorofic, added names, etc.

<PLACENAME> (I.E. PLACE NAME)
Description: contains an absolute or relative place name.

<ROLENAME> (I.E. ROLENAME)
Description: contains a name component which indicates that the referent has a particular role or position in society, such as an official title or rank.

<TIME>
Description: contains a phrase defining a time of day in any format.

Adam Crymble – 2016.

ACTIVITY 1 - TEI MARKUP CHALLENGE

Using the tags on the cheat sheet, do your best to tag the following passages (type them out on your computer. Use a plain text editor such as Komodo Edit). Think about whether or not there is more than one tag you could apply in a given situation:

1. The park has a lovely duck pond.

2. Our butcher sold us some rotten meat yesterday afternoon.

3. Dr. Havingsbury goes to church every Sunday.

4. Mrs. Wellington, who was wearing her Wellington boots, accompanied me on a trip, where we saw a statue of the Duke of Wellington in Wellington, New Zealand near the Wellington's Boot public house - where I hear they serve excellent beef Wellington. I do adore Mrs. Wellington.

Adam Crymble – 2016.

ACTIVITY 2 – TEI IDS

Most TEI tags let you add IDs. This means that if something appears twice in your text, you can use the unique ID to say: this is the same as the other tag with the same ID.

What you use for the ID is up to you (but you should devise a system that you use consistently within a single project or the ID becomes meaningless)

You could add an ID like this:

<persName id='MrsWellington'>Mrs. Wellington</persName>, who was wearing…I do adore <persName id='MrsWellington'>Mrs. Wellington</persName>.

Tag the following passage using some of the tags you have already learned, and add IDs where you think it's appropriate.

(Available online at: http://shakespeare.mit.edu/lear/full.html)

ACT I
SCENE I. King Lear's palace.
Enter KENT, GLOUCESTER, and EDMUND

KENT
        I thought the king had more affected the Duke of
        Albany than Cornwall.

GLOUCESTER
        It did always seem so to us: but now, in the
        division of the kingdom, it appears not which of
        the dukes he values most; for equalities are so
        weighed, that curiosity in neither can make choice
        of either's moiety.

KENT
        Is not this your son, my lord?

GLOUCESTER
        His breeding, sir, hath been at my charge: I have
        so often blushed to acknowledge him, that now I am
        brazed to it.

KENT
        I cannot conceive you.

Adam Crymble – 2016.

ACTIVITY 3 – TEI ATTRIBUTES

In addition to IDs, many tags also let you add 'attributes'. These attributes let you specify using a limited vocabulary (of your own creation), exactly what it was the tag captured.

For example, you might want to use the <occupation> tag with a limited vocabulary of attributes (merchant, soldier, other) so that you can isolate merchants and soldiers in your document(s). You can do this by adding a 'type' attribute to your occupation tag:

2. Our <occupation type='merchant'>butcher</occupation> sold us some rotten meat yesterday afternoon.

Try adding TEI tags from the list above, and include both IDs and attributes. For the sake of simplicity, assume all tags can have a 'type' attribute (if you were working on a TEI compliant project, you would have to check the online manual to be sure).

(Available online at:
http://www.oldbaileyonline.org/browse.jsp?ref=t17360908-8)

Thomas Hornbrook , was indicted for stealing a Gelding of a wan Colour, value £3 the Goods of Thomas Merry , August the 10th.

Thomas Merry. My Horse was in the Church-yard of Yalling in the County of Huntingdon , on Friday Night, August the 6th, on the 7th in the Morning he was gone. I had him cry'd at St. Ives, and 2 or 3 other Places, and at last this next Witness Mr. Unwin gave me Information of him.

Mr. Unwin. I live in Well-street, near Well-Close-Square. The Prisoner came to me the Sunday after the Horse was lost, about 3 or 4 o'Clock in the Afternoon, and told me he had got 2 Horses, but as one was enough for his Business, he would dispose of the other. One of them was sit for a Saddle or a Chaise: I said, I wanted such a one, and he brought this wan Gelding to me on the Tuesday following, and told me he was worth 5 or 6 Guineas; but if I would buy him, I should have him at any Price. The Horse was in very bad Condition, when I saw him: and I suspected he was not honestly come by. He call'd for Bread and Cheese, and sent me out for some Onions; instead of going for Onions, I got a Constable and had them before Justice Farmer: the Prisoner had a Boy with him, whom I took up too, and he gave me the Information.

Cox, the Boy. The Prisoner and I had swept the Prosecutor's Chimney, and when we had done, he came to me and fell a Laughing; I asked what he laughed at; why says he, here is a Saddle and Bridle, and there's a Horse in the Church-yard, we'll come at Night and take them away, and go to London. We came about 12 o'Clock, he took the Bridle and Pannel out of the Stable; I held open the Gate; he rode the Horse out, and we both came away together. Guilty, Death .

ACTIVITY 4 – HOW DID THE OLD BAILEY ONLINE DO IT?

The Old Bailey Online, where the trial you just marked up lives, has marked up all of their trial transcripts using TEI. Go back to the website you just marked up and at the bottom of the page, click on the 'View as XML' link.

Take a look at the tags the Old Bailey Online team has used to mark up their text. How is their markup different from yours?

Write down all of the tags you have not seen before and search the TEI manual (online) to find out what they are used for.

Are there tagging decisions here you disagree with? Anything else you think they should have tagged?

Have you figured out how they are using <join> tags?

Workshop 3 – Building a Virtual Archive and Exhibit with Omeka

Learning Outcome:

- To understand what goes into building a virtual archive and exhibit
- To gain some experience constructing a digital archive/exhibit
- To learn to consider the needs of audience

Background:

We increasingly turn to the Internet to find material for our humanities-based research. Huge digital archives provide access to the world's art, artifacts, textual treasures, and more. These database-driven websites may look intimidating, but there are options available that let you build them without any prior technical knowledge.

In this workshop we will use a free online platform known as Omeka (omeka.net) to build a simple but powerful digital archive and exhibit to showcase digital material of interest to us. This will involve thinking critically about how to tell a story with cultural objects, and you will need to consider your audience as well as your own interests.

Part 1: Getting Started with Omeka (30 minutes)

We are going to use the free online version of Omeka – a platform (or 'content management system') for building websites of artifacts. It was originally developed for the museum and archive community. There are limits to what you can achieve with the free hosted version of Omeka. If you wanted more flexibility you could download the software and host it on your own server (instructions on the site), but for now, we will focus on content generation.

Work through the introductory Omeka lesson by Miriam Posner, available from the *Programming Historian*:

> Miriam Posner, 'Up and Running with Omeka.net', *The Programming Historian* (2013), http://programminghistorian.org/lessons/up-and-running-with-omeka

You can either work with materials from your own research / area of interest, or you can find some objects from existing repositories / museums / galleries / archives. If you can't find anything, ask and we'll find a collection that interests you.

Part 2: Creating an Exhibit (20 minutes)

Once you have a few objects in your repository, move to the next step and build an exhibit that brings everything together for an audience. It pays to spend a few

moments asking yourself who might be interested in this material, and therefore what type of presentation would interest them. Do you think they would be interested in reading a lot of in-depth research? Do they just want a few facts? Will they want an interesting narrative that ties together the objects in your collection? Thinking about audience is key to a good digital exhibit.

When you're ready, move onto the second lesson in the series:

> Miriam Posner and Megan R. Brett, 'Creating an Omeka Exhibit', *The Programming Historian* (2013), http://programminghistorian.org/lessons/creating-an-omeka-exhibit

Spend the rest of our workshop experimenting with the features on Omeka.net and improving your site with users in mind. You might want to consider what aspect of your research you could share in an online exhibit and how that might help you to understand your material.

Adam Crymble – 2016.

Day Two: Intro to Programming
(Petri Jooste)

Workshop 4 – Programming in Python

Learning Outcome:

- To become comfortable with the basics of computer programming
- To get some experience with a particular programming language (Python)

Background:

Computer Programming may sound daunting, but it's easier than you think. Learning a little bit about programming means you can build your own tools to do new kinds of research. It's worth trying out the basics to get a sense of what it's like and what it can achieve.

There are lots of different programming languages, each with their own strengths and weaknesses. Each has its own books that you can buy to teach you the basics. But there's also now a great set of online tutorials that will take you through fundamental principles step-by-step.

Code Academy, offers free self-directed python tutorials:

- 'Python', *Code Academy* https://www.codecademy.com/learn/python

We'll be working through some online material that is published under the Creative Commons license.

Please remember to install Python on your laptop. It can be installed on all operating systems (Windows, Linux, and Mac OS). Please install version Python version 2.7.11 – we will not be working in Python version 3.X. Tomorrow's activites will need a working installation of Python version 2.X.

https://www.python.org/downloads/

The more you do the easier you will find our workshop tomorrow.

Day Three: From Texts to Maps
(Adam Crymble)

Preparatory Work and Background

Our case study today will consider ways that we can map a series of historical mini-biographies of Oxford University students who studied during the 16th and 17th centuries. The full dataset includes mini-biographies of 60,149 students, including which college they attended, when they arrived, and in many cases, where they came from or what they went on to become. This is typical of the type of source a scholar might want to map. It has lots of places listed, but at the moment it isn't structured in a way that makes it easy for us to see the big picture – the trends of where students came from. Our workshops today will help us overcome those barriers and will leave you with the skills to build your own digital maps using records of interest to your own research.

Figure 1: The Alumni Oxonienses Dataset

The dataset has gone through a number of processes before we got our hands on it. It was originally compiled by Joseph Foster in the 19th century, who used the admissions books of the colleges themselves along with his own research, to generate this information. These two volumes were titled *Alumni Oxonienses* (the alumni of Oxford). Similar works have been created for Cambridge (*Alumni Cambrigienses*) and Trinity College Dublin (*Alumni Dublienses*).

The first volume was then digitised by British History Online (http://www.british-history.ac.uk/alumni-oxon/1500-1714) and put online as fully reusable text in a format that mimicked the alphabetical nature of the original volumes.

Finally, I downloaded the whole set in 2014 and converted it from a series of web pages into a spreadsheet so that it can be further structured and sorted to our heart's content. During this process, I added a column that makes it possible to re-sort the dataset back to its original order, if ever it gets out of order during the analysis stage. You should have been given a copy of this dataset by the workshop organisers.

- *Alumni Oxonienses* (1500-1714).

Adam Crymble – 2016.

Admittedly, the people who appear in these records are not very diverse or representative of wider society. They are all men, probably almost all white, and nearly all of the upper classes, and probably include a large number of second sons who would not have inherited the family fortune as would their older brother. University at the time was largely a place for would-be clerics, but also lawyers, scientists, and merchants. It's where you got your religious or legal training before heading off for a life in the monasteries, or preaching the word of God, or as a lawyer.

Our goal is to find out where they came from originally and to map the results as a 'heatmap'.

Before you come to class:

- Download a copy of the dataset (Alumni Oxonienses).
- Install Python v. 2.7 (NOT VERSION 3) on your computer.
- You may find it helpful to read the short 'Preface' to the original volume so you understand the records we're working with: http://www.british-history.ac.uk/alumni-oxon/1500-1714/v-vi

Workshop 5 – Building Digital Maps with Google Fusion Tables

Intended Learning Outcomes:

- Understanding how to geo-encode records, semi-automatically
- Understanding how to interrogate geographic evidence using Google Fusion Tables

Optional Reading:

Digital mapping provides us with a heads-up way to understand patterns in our datasets. Many scholars have done this to great effect. If you want to see how digital maps can help us understand data, read through how Ben Schmidt has done so with historical data about whaling vessels.

- Ben Schmidt, 'Data narratives and structural histories: Melville, Maury, and American whaling', *Sapping Attention* (2012).

Background

Sometimes it can be difficult, if not impossible, to see the forest for the trees when working with a spreadsheet full of data. This week we will use some online tools to map the place of origins of a subset of the students in the *Alumni Oxonienses* dataset to see what we can learn.

This builds upon work done by the *Digital History Workshop* students at Hertfordshire in 2014. Those students were challenged to 'add value' to the original dataset, which involved extracting the place of origins of the students, along with the colleges they attended. We will use their dataset to map those students.

> Corey Albone, Jack Dunne, Namiluko Indie, Bethany Reid, 'Oxford Knights', *The Oxford Knights Archive* (2014-15). [revised slightly by Adam Crymble, February 2015, to make this exercise easier]

Part 1: Getting to Know the Dataset (20 minutes)

As has been the case throughout this module, you will need to spend a few minutes getting to know the dataset they created. In this case, the group decided to focus on entries of individuals who would be knighted later in life (eg, those who were 'sir' soandso). This includes the mini-biographies of 891 individuals. From the original records, the group extracted the following information and organised it cleanly into new columns:

- Surname
- Forename
- Relations (relatives mentioned)

Adam Crymble – 2016.

- Matriculation date (start of studies)
- Matriculation year
- Origin
- Modern county/state
- Country
- Distance from Oxford (miles)
- College
- Place of Origin

This has given us a wide range of new options for how we can analyse and understand these individuals. Spend ten minutes getting to know the dataset and think about how it is different from the original available on *British History Online* (http://www.british-history.ac.uk/alumni-oxon/1500-1714/pp1-28). Once you are comfortable, move on to the next step.

Part 2: Geocoding the data using Google Fusion Tables (60 minutes)

Before we can map the whole set, we need to 'geocode' it. That means associating the various place names in the 'place of origin' column with the geographic coordinates on the globe that corresponds to the same place. There are a number of ways you can do this, ranging from getting out your atlas and doing it by hand, to using various online tools that will try to match entries automatically.

We're going to use the latter option, and will be trying a tool known as Google Fusion Tables (https://support.google.com/fusiontables/answer/2571232).

Visit the address above and create a new Fusion Table. Upload the *Oxford Knights* dataset. This will take a few minutes.

Next, we want to tell the tool which columns in our spreadsheet we want to geo-encode. It's automatically guessed some it thinks look like place names and highlighted them yellow. We could geo-code those, but we're interested in the 'Place of Origin' column, which includes the most comprehensive information.

Geocoding takes time, so we want to reduce the amount of work that the program does so we get to our map sooner. To do this, set all columns with yellow in them to 'text' by clicking on the column header and then 'Change'. Change to text. This means the tool will ignore those columns for geo-encoding purposes.

We do, however, want the 'Place of Origin' column to be set to 'Location'.

Then click on 'File' -> 'Geocode'. This will take a few minutes. It will take WAY longer if you don't do the step above. Watch as the program does its processing. What do you think 'Quality: X% ambiguous' means? What does this tell us about this as a scholarly tool? Which entries were ambiguous? Can we know?

Adam Crymble – 2016.

When the geocoding is done, you should see a map with lots of points on it, showing where these knights came from originally. If you don't see the map, try clicking on the tab that says 'Map of Origin'.

Fixing the Obvious Errors

Your map might have an entry in Vancouver, Canada for someone named Richard Breame. If so, you'll note by clicking on that entry that it's misinterpreted 'Surrey' which is a County in England just south of London, for a town on Canada's west coast by the same name. That's not right, because our data refers to people who died long before that town was founded. So we want to change it.

Go back to the data 'Rows 1' tab.

To find him we need to filter the records by something unique about Richard Breame. His name should make it easy enough. You can either sort the 'Surname' column 'Sort A to Z' by clicking on the column header. Or you can add a filter and isolate only people named 'Breame'. Let's do that so that you know how to do it.

Click on the 'Filter' button and select 'Surname'. Now you can either type in his name or scroll through the options. Type it in and click 'Find'. That should bring you to the right entry.

The problem is that 'Surrey' is ambiguous. We need to specify that we mean the Surrey in England. We can do this by editing the 'Place of Origin' column entry. Click on the cell in the table that you want to change. The whole row should go yellow. Then click on the 'Edit' button, which looks like a pencil. This should pop up an edit row window. In 'Place of Origin' change what's there to 'Surrey, England' and then click on the little 'Geocode' link. It will try to put this near a town called Guildford, which is in Surrey – close enough because we don't have any finer-grained detail about where this person comes from. You can click 'Done' and then 'Save'. Your entry should now be correct. If you go back to the 'Map of Origin' tab, your entry on the Canadian west coast should be gone.

There's another obviously wrong entry in the Congo in Africa. If you can figure out where that is supposed to be, you can change it (I couldn't figure out where Wenge was!).

Making a 'Heatmap'.

Back to the map, the way it's formatted at the moment doesn't tell us a lot about the distribution of these people because so any of the dots overlap one another. We can see where these people are clustered if we use a 'heat map' instead.

Click on the Heatmap tab (or click on the 'Map of Origin' tab at the top and click 'Change Map', if you can't see that option). You can increase the radius and opacity of the heat map, which should make it clear where these people are clustered.

Take a look at the patterns. Note any areas that have lots or not very many students. Take a look in particular at the city of Cambridge (the home of another famous university). What do you see?

Filtering by Year

On the 'Map of Origin' tab, click on 'Filter' and select 'Matriculation Year' – which is the year the person started university. Filter from 1000 to 1600 and click 'Find'. The map should update.

Does it look noticeably different?

Filter again, this time from 1600-1714. What changes?

Remove the first filter. Filter this type by 'College'. Select 'Brasenose College' and make a note of what you find.

Now remove 'Brasenose College' and select 'Christ Church College'. How is it different?

Again, this time select 'Exeter College'.

All three colleges have roughly the same number of graduates. How is the spatial distribution of them different? What does this tell us about Oxford Colleges? Did you know that before? Could you tell from looking at the original records?

How is this process fundamentally different as an approach to understanding the past to the process you typically undergo when writing essays?

Workshop 6 – Extracting Keywords from Texts

Intended Learning Outcomes:

- Basic programming exposure with a scholarly-need in mind
- Ability to work independently through a written technical tutorial
- Ability to extract keywords from texts

Background

It's one thing to use someone else's structured data, but we won't always be able to turn to someone else's pre-cleaned dataset. It's important to know how to do the process yourself so that you can apply your new skills to any material you come across.

In this workshop you will learn how to extract sets of geographic keywords from these mini-biographies of Oxford students, and then put them neatly into the spreadsheet next to the appropriate entry – just like in the dataset we mapped last session. This involves learning how to use gazetteers of keywords, along with some simple programming (don't worry, we'll go step-by-step).

Part 1: Using Gazetteers to Extract Sets of Keyword (120 minutes)

I've published a peer-reviewed tutorial on just this task in the *Programming Historian*, so we'll work through that lesson.

Work through:

> Adam Crymble, 'Using Gazetteers to Extract Sets of Keywords from Free-Flowing Texts', *The Programming Historian* (2015), http://programminghistorian.org/lessons/extracting-keywords

You can stop once you have finished the section on refining your gazetteer – about two-thirds of the way through.

Using a sample of the data from the reign of King James I, you should now have the skills to extract the place of origin of each student. You should now have the ability to map these students using Google Fusion Tables as in the last workshop, and could generalize what you've learned to apply to any set of records that contain place names.

Resources for Improving your Skills

I hope you've found something useful for your research in our workshops. If you would like to follow up on your skills development there are lots of free resources available online.

I am involved with the *Programming Historian* (http://programminghistorian.org), which we used in a number of our workshops. On the site you can find a number of additional tutorials designed to help you learn new digital skills to improve your research capability. The *Programming Historian* also accepts submissions, which are peer reviewed before being incorporated into the final project, so if you have a skill you'd like to share, please get in touch.

Good luck with your learning, and I look forward to seeing all of the great research you produce with these and other new skills.

Thanks for taking the workshop series!

Adam Crymble
London, 2016.