Online Tools for DH - Workshop

Out(side) the Box - Online tools for DH/HSS

What do you do when your personal computer is no longer sufficient for your research needs? Do you need to run that web scrape for months at a time? Some DH tools are only available as web based services, and many run well when self hosted in cloud resources. In this workshop we will look at some web only tools, as well as how to host your own in the cloud. After a quick introduction to some tools and resources, we will work through a couple of examples, one web scrape and visualization, and one text analysis.

Resources

Let's start by looking at how to find DH tools. Well, just do a web search with your favourite search engine.

There are lots of lists of tools, many associated with university libraries or DH/HSS programs, and they often contain lists of lists of tools. It won't take long to find a handful of quite large lists of tools:

- TAPoR
    - This is now in version 3, rebuilt in 2018, and incorporates a previous large list, DiRT. Originally targeted at textual analysis tools, with the incorporation of DiRT it now includes many other areas, including GIS, photo/video/audio, etc.

- Digital Humanities Tools
    - Maintained and curated by Alan Liu, this is targeted at free or mostly free tools, with a "[b]ias toward tools that can be run online or installed on a personal computer without needing an institutional server." ...in other words, right up our alley!

On the social sciences side of things, I found the following (from a similar random search for "social sciences tools"):

- SAGE Ocean Research Tools Directory
    - This is provided by the SAGE Publishing company as part of their SAGE Ocean research portal.

- Data Analysis tools & training
    - This is maintained by the Bodleian Library at Oxford.

There seem to be fewer such resource lists for social science than for humanities, though that's anecdotal. At any rate, I was introduced to the following by a researcher in geography who asked me to install one of these tools in the Compute Canada cloud for him:

- SciencesPo médialab Tools
    - This group, founded by Bruno Latour, does a lot of code development of open tools, which is awesome! And this list led me to this:

- DMI Tools
    - This set of web based tools, produced by the Digital Methods Initiative, is ready to use right there. Some require authentication (i.e. membership in the Initiative) to use, but others are open to use by anyone.

On the social sciences side of things, instead of "textual analysis" you will find social scientists performing "influence mapping" or "controversy mapping", and so on. But many of the 'tools of the trade' are the same: Gephi, Voyant, OpenRefine, RAWGraphs, etc.

Examples

Research has basically three phases, gathering, analysis, and publication/presentation. Yes, that's an oversimplification. But it gives us a narrative arc with which to look at tools. I'm going to skim through two examples, one from social science and one from more traditional DH, textual analysis. We will look at one real example and one fake.

Transboundary E-Waste

This is the real example, a multi-year research project that used web-based tools to produce a physical book and an online book. Here's the online end result:

- Reassembling Rubbish

The physical book was published by MIT Press in 2018 (Reassembling Rubbish), and this is the associated electronic publication, hosted on Scalar. Josh Lepawsky is a professor of Geography at Memorial University of Newfoundland.

Scalar is a well-known digital scholarly publishing platform. The results, as you can see, can be very "book-like", insofar as they can include chapters, indices, footnotes, etc. This is an example of a "book companion". The platform is capable of less book-like use too, though, e.g. digital exhibits. But they all have a definite 'scalar' look and feel.

In 2016, Josh came to Compute Canada looking for help with hosting some web based tools he needed for the preparation of material for this project. He had already received some help from Memorial University, specifically with hosting media files for the publication, which he had already started building on Scalar. But he had needs they were unable to provide, namely cloud-based application hosting. Specifically, he wanted to use this tool as a data gathering and analysis engine:

- Hyphe

This is a 'web scraper', but targeted at social scientists, with a focus on curation of web 'corpora', sets of web pages curated and organized by the researcher in a guided fashion. At the time, it needed Linux to run, which Josh did not have, and he needed to be able to run jobs independently of his desktop computer. So we got him some Compute Canada Cloud resources, and I built a VM for him and installed the tool. These days, it's easier to do, as it is now provided as a 'container' which can be run on any OS that can run Docker. But the benefit of desktop independence is still a strong incentive to run this in the cloud.

Let's look at one of the results of Josh's work, before taking a closer look at the tool:

- Transboundary Movement of Electronic Waste: mapping a controversy
- the source file for the graphic

The 'corpus' extracted with Hyphe was refined and examined using Gephi. And to produce a presentable 'network view' they used this tool (a Gephi plugin to allow interactive display in a web browser):

- Sigma Exporter

Some guides to how to use Hyphe:

- A research paper by the developers
- A guided use case of the tool

Grab just a few of the 'seeds' from the second of these and look at Hyphe in action.

The results take a lot of work to refine, and extract useful information, even after just a few minutes of crawling.

Text analysis

This is the fake example. I'm going to grab an e-text from ibiblio (Project Gutenberg), push it through a couple of Voyant tools, and display the result in a project on CWRC. As a companion to the TAPoR tool list, the team that produced TAPoR also produced this site, a collection of "research methods and techniques for analyzing text":

- Methodica Commons

One of their resources is a large collection of 'recipes' for doing things like sentiment analysis, authorship analysis, etc. I did a search for 'Voyant' and got a list of recipes that use Voyant, and then picked this one to work through. The text I chose is John Stuart Mill's On Liberty.

- Voyant

Tools to look at: 'Terms', 'Cirrus', 'Links', 'Trends', 'Contexts', 'Collocates', 'Documents'

Things to note: 'Options => Stop words', 'Export => View/Visualization' (graphical tools), 'Export => View/Current Data' (tabular tools)

What do you do when you have some digital material from a research project or an archive and you want to present it online? Scalar is one option, if your project is of that type (book like). You might also consider a presentation/collection platform like Omeka, or a more repository-like platform such as Islandora. Discussing these platforms, what they can do, how to use them, etc., is the topic of another workshop and more! But to give you a flavour of what that might be like, I'm going to show you how to get started with one.

The Canadian Writing Research Collaboratory is "an online infrastructure for literary research in and about Canada designed to meet the challenges and embrace the opportunities of the digital turn." (See 'About'.) You can see from some of the active projects that this has a quite broad sense of 'literary research' that includes born digital material as well as more traditional sources. Explore the 'Collaboratory' => 'CWRC Commons' => 'CWRC Videos' for some short introductory material.

The underlying technology is Islandora, which is widely used for library digital collections. The 'repository' back-end is Fedora and the 'presentation' front-end is Drupal. The 'Islandora' part is a collection of Drupal plugins that provide 'CRUD' functionality on the back-end, including management of metadata schema, object types, objects, collections, etc., as well as Solr and triplestore indexing. 'CRUD' is 'create, read, update, delete', the full set of functions for life cycle management of objects in the repository. Don't worry if the word salad doesn't mean much to you. That's a topic for a separate workshop/course. But two important things need to be said about CWRC in particular:

- It is hosted, which means if you work with them, they take care of all the underlying OS and application management, including backups and so on. That is a big advantage if you need the data integrity and metadata flexibility that this sort of platform offers and you don't have or have access to the technical skills to manage the software stack yourself.
- The power of Drupal, in significant (though asterisked) measure, can be used for presentation development. You are not restricted completely to the sort of "modern repository" look and feel of Islandora. The caveats have to do with how the front-end accesses and displays the back-end objects/data. But it is quite possible to develop lovely looking online exhibits.

Another 'value add' of this platform includes integration with tools such as Voyant and CWRC-Writer, as we will see.

Other Resources

- UofT Collections - examples of use of Islandora for exhibiting collections.
- NYAM Collections - more Islandora collections.