Towards a New History Lab for the Digital Past.

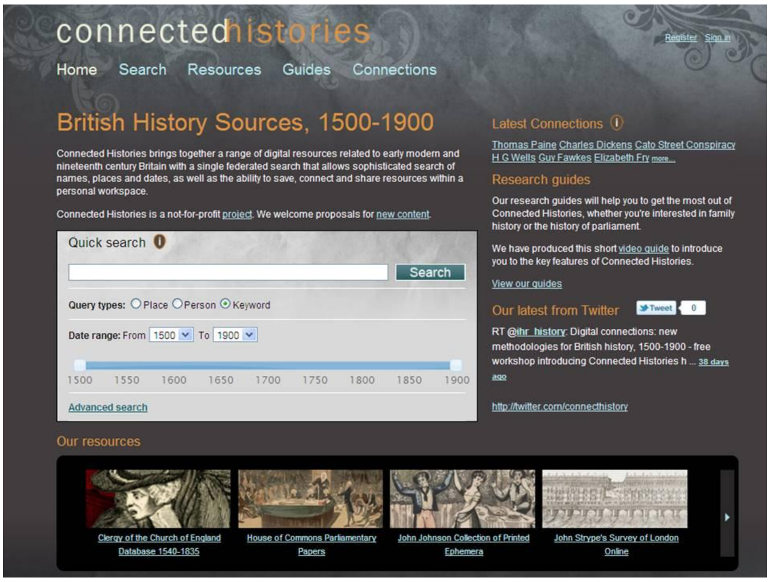
Tim Hitchcock (University of Hertfordshire)

When the Institute of Historical Research was first established in 1921, its purpose and object was described as:

> to become an index to historical knowledge, a focus of historical research, a clearing-house of historical ideas, and a historical laboratory open to students of all universities and all nations.
>
> Institute of Historical Research leaflet, 1921

And those of you who know the IHR, as it has evolved through almost a century of change, will recognise in its seminars, in its unique open shelf library, and in its simple role as the centre of a community of historians, of students, and of the curious and argumentative, the continuing vibrancy of this original spirit and purpose.

In many respects, Connected Histories is a simple attempt to ensure that this spirit and objective continues to thrive online; that immediate access to 2 billion words and 150,000 images - searchable at the click of a mouse, and sharable across time and space – will enhance that community, and the history it creates.



British History Sources, 1500-1900

Connected Histories brings together a range of digital resources related to early modern and nineteenth century Britain with a single federated search that allows sophisticated search of names, places and dates, as well as the ability to save, connect and share resources within a personal workspace.

Connected Histories is a not-for-profit project. We welcome proposals for new content.

Quick search

Query types: ○ Place ○ Person ● Keyword

Date range: From 1500 To 1900

Latest Connections

Thomas Paine Charles Dickens Cato Street Conspiracy H G Wells Guy Fawkes Elizabeth Fry more...

Research guides

Our research guides will help you to get the most out of Connected Histories, whether you're interested in family history or the history of parliament.

We have produced this short video guide to introduce you to the key features of Connected Histories.

View our guides

Our latest from Twitter

RT @ihr_history: Digital connections: new methodologies for British history, 1500-1900 - free workshop introducing Connected Histories h ... 38 days ago

http://twitter.com/connecthistory

Our resources

Clergy of the Church of England Database 1540-1835
House of Commons Parliamentary Papers
John Johnson Collection of Printed Ephemera
John Strype's Survey of London Online

But Connected Histories is also a recognition that the nature of historical research has changed; that we are drowning in an infinite archive – an ever expanding world of information. And that the secure sense of a discipline that knew how to judge quality, how to assess evidence, is challenged by the sheer number of sources we can interrogate for words – at least - if not yet for meaning.

Given the privilege of a few minutes with a powerful audience, I want to do a couple of things this afternoon. First, I want to describe just what Connected Histories does and how it works. And in the process say a bit about why it is designed the way it is, and what issues it is meant to address. And second, I want to talk a bit about how it fits into a trajectory of changing research and publishing practise – to describe where it sits in a process of frighteningly rapid change, and to locate it along with the other resource being introduced today – Mapping Crime.

Connected Histories is what is called a 'federated' search facility, and currently makes some eleven different web resources available – over two billion words of text, and 150,000 images, some free to access, others supported by a JISC license for use in British Higher Education, and others still, commercial sites designed for a wider audience of family and local historians.

It includes

British History Online

British Museum Images

British Newspapers, 1600-1800

Charles Booth Archive

Clergy of the Church of England Database 1540-1835

House of Commons Parliamentary Papers

John Johnson Collection of Printed Ephemera

John Strype's Survey of London Online

London Lives 1690-1800

Origins Network

The Proceedings of the Old Bailey Online, 1674-1913

And we are in the process of adding several more.

Underpinning its searches are indexes of every word in those eleven 'distributed' websites – each of which were chosen to represent large bodies of academically credible and relevant material. As a part of this index, each word is associated with a web address, a URL, that allows you to click through to the original. This creates a basic facility that in response to a word or phrase search will return tens of thousands of results, each associated with a snippet of text, and each linked to the full resource held elsewhere.

In other words, at its fundament and in its water, Connected Histories is simply a comprehensive index of words. But in the process of creating that index, we also sought to assign meaning to some of them. Using a methodology called natural language processing, we identified names and dates and places (to an accuracy of around 75%). So, in addition to an index of all the words in these 11 resources, we have also created indexes of all the names and places and dates mentioned: all the names in the Burney Collection, and all the dates in the Parliamentary Papers (however they are expressed).

In other words, what we have is not just one, but four indexes, and you are searching each of these 2 billion words, or the millions of names or dates or places, each time you enter a query – allowing you to combine keyword and name, date and place searches to find just what you want.

That it works is a testimony to the hard work of the technical staff at the HRI and IHR, to Kathy Rogers and Bruce Tate in particular. But also to Sharon Howard, who managed the project, and to the large team of people involved.

The starting point for this project was always an attempt to address what Digital Humanists tend to label the 'silo effect' – the idea that one of the problems with small scale websites and resources of the sort so many of us have worked to create over the last fifteen years, is that you tend to go along to one site – do a bit of research – before heading to another. That just like traditional forms of research most of us forget what we knew in the British Library, during the short walk to the London Metropolitan Archives.

And in its most basic formulation the silos Connected Histories seeks to blow apart, are the boundaries between web sites. You can now cross search the British Museum image collection, against Strype's history of London, and associate images of specific locations, with descriptions and commentary on them. You can search the Parliamentary Papers, in combination with the records of all the sessions papers of the county of Middlesex – bringing onto a single screen precept and practise. There are sixty thousand settlement examinations that can now be cross referenced against apprenticeship documents, and trial records.

But this blithe image of easy cross searching, fundamentally understates the complexity of the issue, and the precise reasons Connected Histories is designed in the way that it is.

One overwhelming, and very real, aspect of the 'silo effect', is that while many of the primary sources we need are freely available, many others are not. The walls of some silos are much more difficult to breach than others. While frustrating, this is not necessarily a bad thing. Unless we can convince the state and the taxpayer to pay for universal digitisation, we can't really complain if the cost of digital resources is being borne by the end users. Early modern and modern Britain is quite simply the most digitised where and when in existence, because of the combined efforts of the academy, of the great cultural institutions of Britain, of individual scholars, and private publishing companies motivated by profit. But, at the same time, as scholars and teachers, we need access to all those resources. Or at the least we need to know what is in them, in order to make an informed decision about what we want out of them.

The model of a series of indexes to the original material is precisely designed to address this issue. We don't need to have direct access to all the products of ProQuest and Gale, of the Origins.net, or JISC funded projects like the John Johnson collection, if we have an index that tells us what they contain. These materials can sit behind their paywalls, the intellectual property they contain safe from harm; while we can now interrogate them from a distance.

In other words, Connected Histories is designed to build a bridge between the academy and commercial publishers; it is designed to mess up the models of delivery, and the walls of division that keep us apart. These 'silos' are almost philosophical in character, but even more than the technical ones that divide one website from another, they need to be breeched; and that is part of what this project has been about.

But, the silo effect goes beyond even this. It exists between our own ears as well. At its best and most compelling, history is a community of scholars, sharing knowledge and effort in pursuit of a real and usable understanding of the past – it is a collective project. At its worst, it is a collection of egomaniacs, desperate to be lauded as the great authority on this or that – however specialised and narrow that might be. The much lamented lone scholar is as frequently a Casaubon, forever seeking and failing to find the key to all knowledge; as they are a Dorothea, driven by enthusiasm and a desire to share with others. At some level, the 'silo effect' is inherent in the idea of 'authorship' (and the 'authority' it implies). It is there when we decide one persons' work is literature, and another is art history; when we label by period or methodology, when we decide who to exclude from the conversation, and who to include.

We don't all need to work collaboratively, or to abandon our notions of intellectual property, but in the spirit of a 'history lab', we do need to share our work, and remember the common purpose of historical research. And Connected Histories is again an attempt to address these particular silos.

By creating individual workspaces that build into a new body of ‘connections’ , by allowing users to link documents, and names, and stuff, across billions of words, and then pooling those links and allowing them to be explored by a wider public; Connected Histories, is designed to build a new shared body of knowledge grounded in everyday practical scholarship. It is designed to nudge the lone scholar to become a more sociable animal.

In many respects all these ‘silos’ are part of our inheritance from the Enlightenment. They are inherent in every library catalogue, and in the practise of individual scholarship leading to named authorship. They reflect the co-evolution of the academic community, in a symbiotic death grip, with commercial publishing; and they were imported without fanfare or thought, into what one might want to describe as Web 1.0 – that first iteration of the internet created in the image of older forms of scholarship and communication – with e-mail, e-spreadsheets, e-footnotes, e-everything – all mimicking an older intellectual technology.

In other words, there is a bigger ‘silo’ out there – a division that is more fundamental to the internet and the cultures of scholarship than the mere distance between the technical implementation of British History Online, on one hand, and The Burney Newspapers or the Old Bailey on the other.

Yes, we want to consult this material in one go; and yes we need to overcome the boundaries, created by pay walls and subscription; and yes we want historians to work together in a common laboratory of ideas and connections. But, what really needs to break down is the silo that suggests that information itself is something to be consulted and collected; that it is an unchanging object of study, rather than a pool of constantly changing stuff that can be interrogated from any angle, and pursued along any trajectory.

The most fundamental silo Connected Histories is intended to address is between traditional forms of criticism and scholarship that assume we can contain data in an internally structured and divided, ‘library’; and the emerging world of text and data mining, that sees data as a process – something to be played with and analysed on a massive scale, across boundaries of genre and type.

The innovation at the heart of Connected Histories, the one I think is most interesting, is the methodology used to allow us to sit in London this afternoon, and locate the site and its gubbins in the IHR; while the indexes it interrogates sit on a server in Sheffield – which distributes pointers to eleven different servers around the country.

What has been created by the Institute and the Humanities Research Institute in Sheffield, is a model that uses an ‘API’ as its core. An API is an Application Programming Interface (the most widely used version of which is Google Maps), and it is designed to allow you to create a simple query that can address a dataset from a distance (in this case four indexes). It is not a website, or a 'front end', it doesn't need to exist as a visual or physical thing. It is essentially a series of agreed conventions that allow anyone to address a web resource and ask it for a bit of the data it contains. What Connected Histories does, is locate the 'front end' in London, with information about the sources, with the workspace and connections, etc., but that front end's main job is to address the API in Sheffield, to gather the data required from the indexes, to bundle it up into an xml file, and to present it in an attractive way to the end user who can then navigate to the original sites, create links, and share searches.

In other words, the indexes in Sheffield have been created as a standardised and generic resource, which is then addressed by a specialised and bespoke search and save environment.

For most of us, this is a seamless process of little interest; but what it does is create a space between search and data, that can now be occupied by anyone. In other words, and unlike most free-standing websites - it is designed to be mash-upable.

With a bit of technical nous you can now generate a bit of code that will automatically select and download the contents of all the indexes, reflecting all the words in Connected Histories. The text miners who do this will not be gaining access to the original resources – there is no intellectual property issue here (beyond that of this project). There is no question of them being able to recreate the sites so laboriously constructed by whatever business or academic model was employed to create them in the first instance; but they will have access to what amounts to a detailed description of the contents of it all – the index of every word, and name and place and date.

Or to put it another way, the API architecture breaks down the structures of online resources into their component parts – separates out data from processing, from delivery - allowing each to be re-used and re-purposed. At the moment it looks like a traditional website with a single front end and datastore, but that front end can address more than one data store; and the datastore can be addressed by more than one front end.

The API architecture addresses that final wall, that silo that means that providers are on one side and data consumers, forced to query the data through an ever narrowing front end, are on the other. Suddenly, we are all mixed up in the infinite archive.

To my way of thinking, this comes under the heading of an unalloyed good thing. An outcome that liberates data, while protecting it; that makes for better history (whoever is writing it), and contributes to the democratisation of scholarship.

But it is only one step in a longer journey; and I want to spend the next few minutes pointing up three or four directions, that I think Connected Histories helps make possible; or which seem to grow naturally from it.

And the first is to do with those academic text miners, suddenly empowered to access ridiculously large bodies of data. What do you do with a 2 billion word index?

What I want to do, is to begin the process of modelling what recorded language since Gutenberg, looks like; how does vocabulary change; how do genre evolve; how are ideas passed from medical literature to political science, to novels; how is changing technology and a changing environment reflected in changing texts. In a sense, half of the last century was taken up in worrying about whether text, words, reflected a knowable universe, or were themselves controlling discourses, leaving humans powerless to imagine something new or describe something real – held captive in words.

I like to describe what we can now access as 'massive text objects' – too large read, too complex to be contained in traditional taxonomies. But, if we can begin to model them – if we can know both the absolute amount of language recorded; and how it changes from source to source, and decade to decade, we can use it in a more sophisticated way to trace first, the controlling forms of language, but also to more securely tie description to an underlying and knowable historical past. If you know the shape, and texture of what has survived, you can begin to think through how it might relate to Herbert Butterfield's, '...genuine relationship with the actual…'.[1]

It is in text mining massive text objects that the hope of a new empiricism in historical analysis lies.

But for myself, I suspect there also something else also going on. The urge to create new connections, to escape our inherited taxonomies, can already be seen in projects such as Mapping Crime – being demonstrated later today. By tying material related to crime available through the John Johnson Collection of printed Ephemera to other repositories and other genre, a reconstructed set of links begins to emerge, that confound the structures

[1] Herbert Butterfield, *The Whig Interpretation of History*, (George Bell and Sons: 1950), p. 73. See also Michael Eamon, 'A "Genuine Relationship with the Actual": New Perspectives on Primary Sources, History and the Internet in the Classroom', *The History Teacher* 39.3 (2006): 32 pars. 6 Sep. 2006 (http://www.historycooperative.org/journals/ht/39.3/eamon.html).

created by librarianship. The data itself, its newly digital form, seems to suggest the need for new connections.

And the API model at the heart of Connected Histories is itself an attempt to embed this idea and aspiration at the core of the design process. It assumes that new connections are there to be made, and that they will inevitably cross boundaries of form and origin to encompass an ever expanding body of inherited artefact.

To take just a small example, we will soon be able to geo-reference at least a portion of the place names in Connected Histories, tying all that text to space in new ways. By modelling maps in the way that we are beginning to model massive text objects, we can relate historical geography to present geography, to secure a further line between representation and a knowable past; and by using an API methodology we are ensuring that it is all mash-upable, with resources from wherever they come. By September (if we keep to schedule), we will have the ability to mash-up eighteenth century London as found in the Old Bailey, and in London Lives, in Google Maps, with a rectified version of Rocque's 1746 map of London, in combination with around 3 million artefacts dating from the period, dug up by the Museum of London.

This JISC funded project is in hand, and is in many ways the natural outcome of what has been described as the spatial turn in historical studies. But by putting an API at the heart of the system, it will again facilitate the re-use and re-imagination of what we can do with a few billion lines of data.

And, of course, we can take the same approach to that other great inherited body of evidence: objects. The historians and the museums will work together eventually (the logic is too ridiculously obvious to need re-enforcing), and at that point, the ability to cross reference maps and texts and objects, again will begin to change how we can evidence the past.

And if we could add to the museum collections, that other massive online record of surviving historical artefacts; that other massive resource digitised by accident – the auction catalogues – we would have created an entirely new resource, available in a new way. Auction catalogues have been created as online, digital resources for over a decade, and already contain detailed descriptions and images of millions of objects: the record of what individuals have valued and preserved on their own behalf, from the past. And thousands more images and descriptions are added each month.

Again, these represent a massive lens through which we can observe the past, and a silo dividing related and cognate materials. Connecting them to texts and maps and stuff, will help us better understand the whole.

It is intended that Connected Histories will grow over time. In its first update in September the National Archives ‘documents online’ will be added, as well as two key nineteenth-century resources: 65,000 digitised British Library books from the JISC Historic Books Platform and the JSTOR collection of pamphlets on social and political issues. Suggestions for additional content are welcome.

But beyond more text, we are confronted with the challenge of integrating more different things. And with each new variety of stuff, we move to a different kind of understanding, more sophisticated, better articulated, more firmly rooted in a clear model of what it is we are looking at; what we can securely see, and what we can’t.

All in all, I think it is kind of cool.

But I also think it remains part of that bigger project:

> “to become an index to historical knowledge, a focus of historical research, a clearing-house of historical ideas, and a historical laboratory open to students of all universities and all nations."

Connected Histories.