Преглед НЦД 26 (2015), 2–9

JenniferEdmond
Trinity College Dublin

TRADITION AND INNOVATION IN THE CENDARI RESEARCH INFRASTRUCTURE

Abstract: The traditions of research infrastructure development have created strong trends and centres of gravity, some of which are useful, and some of which hold back the convergence of analogue research methods and the technologies that could assist them. This is particularly true in the arts and humanities, where the analogue tradition of libraries and archives remains very strong, while new modes of engagement with sources and texts enabled by technological advances remain in their infancy. The Collaborative EuropeaN Digital Archival Research Infrastructure (CENDARI) has established itself as a firm proponent of reevaluating these trends and resisting the gravity where its pull distorts the possibilities for historians to work effectively in the digital age. As such, the project has leveraged its strongly user-centred design processto advance new perspectives on the federation of cultural material an application of knowledge resources within a digital environment. As such, the project represents both technical and social potential to enable new forms of scholarly insightandcommunication.

Keywords: Research Infrastructure, Transnational History, Digital Humanities

Traditions are powerful assets: but they can also be sources of resistance. This is true in particular with regards to new technology, which our traditions can cause us to view as something foreign, dangerous, and needing to be controlled, and also in the world of arts and humanities research, a particularly conservative corner of the research landscape.

Which is not to say that humanistic scholars are luddites, much to the contrary. But we are not programmers, and we don't have much tolerance for tools that either deliver non-intuitive results, or don't allow us to follow their process of argumentation easily.

I say we because I am, at heart, a scholar of humanities, more specifically of German literature. My interest in technology was, in the beginning, opportunistic and superficial: it was the late 1990s, and I wanted access to German language news media in something approaching real time. Over time, however, my interest in the medium began to supersede my interest in the message, and the "cultural interactions" at the core of my research shifted from German writers and their audiences to humanists and computer scientists. My life as a digital humanist had begun, as did my life as an applied humanist, with a focus as much on building things as on simply researching and writing about them.

If you work in the digital humanities, you get to know some fairly established traditions very well:Current historiographical methods can be traced back to around the 1850s, depending on what and how you count. The oldest university press: 1534. The oldest known knowledge organization paradigms go back to 700BC, and the first recorded library to 2600 BC. And, of course, at the start of it all is human nature, which goes back as far as we do.

While all of these long-established habits and patterns shape the work that I do, I want to linger on the first of these just a bit, because the relationship between the historian and their

sources is a key location for observing the shifting ground in the current tension between tradition and innovation.

The historian's concerns with sources usually revolve around their veracity, their completeness, their accessibility and comprehensibility, and it is the calling of the historian to pull them together, verify them and create from them the required or desired record of the fleeting, transitory events of the past.Unfortunately, the categories that the historical researcher uses to categorise his sources don't always map comfortably on to those of the information specialist, the librarian, archivist or museum curator, for whom provenance, completeness and material condition will be of greater concern.

In the analogue world, the need to physically access records eased these differences: historians were required to enter the door of a certain kind of institution, be it a museum, archive or library, and engage with the collections experts they found therein order to access the sources located within. The historian's agnosticism with regards to the location and precise information classification of their sources was easy to manage so long as the records remained in the confines of certain physical spaces. Through face-to-face interaction between experts and collections, competing perspectives could be reconciled, and indeed supportive of each other's end goals. In the digital environment, however, if we want to try and provide a resource bringing together all of the sources historians need to use to formulate and test their conclusions, then we have a much bigger problem.

So progress stalls.

To remedy this lack of progress, we need to recognise that at the heart of this vignette is a simple, but very fundamental truth: that the introduction of technology into a highly refined analogue set of processes presents quite serious challenges to epistomological and cultural norms.

In particular, it challenges the manner in which we view certain types of edges, edges which are dissolving in the current context. The edges between the libraries and the archive, enshrined in their different collection strategies and description standards, are being challenged. The edges between the scholar and the audience are shifting, and the mediating function of the publisherto both validate and distribute scholarly work is facing intense pressure from all sides. Even our perception of scholarly work in the humanities, which long had been epitomised by the image of the independent 'singleton' scholar, is being forced toward revision in the face of the clear requirement for interdisciplinary and intersectoral collaboration to achieve optimal results from the new digital affordances.

While all of these shifts are proving simultaneously traumatic and liberating, it is the last of these that is the most foundational for the digital humanities. Whether you view it as an approach, discipline, methodology, interdiscipline, or other organisational model, the digital humanities is distinct from traditional humanistic scholarship not just because of technology, but because it must encompass different perspectives: technical, collections, social and domain. To do this is difficult, and to do this well requires a particular commitment to recognize the relevance, importance and expert knowledge represented by each of these contributing perspectives. As one library-based collaborator phrased it: "we're very service oriented, but we don't want that to be confused with servitude" (Siemens, et al, 2011), a telling expression of the fact that long-standing hierarchies within the academic research infrastructure landscape may no longer reflect the nature of contributions or ambitions among the actors in this system.

In fact, there is a growing recognition thatthis collaboration does not happen by itself, leading to the emerging role of a new class of collaborator, alternatively called

"intermediaries," (Edmond 2005) "translators" (Siemens et. al., 2011,) or "hybrid people" (Liu et al 2007, Lutz et al 2008 cited in Siemens et. al., 2011)

But if we are facing down closely held traditions of scholarship, we also face the results of what one can only call accidents of history as well.

For example, the emerging paradigm for historical research is strongly transnational, but our collections landscape is anything but. Cultural Heritage Institutions were largely founded and continue to be funded along national lines, contributing not only to a wider scholarly community, but also to local and national processes of identity formation, transmission and maintenance. Our expectations for the digital availability of data are increasing, shaped by the exemplary digital offerings of institutions such as the Bibliothèque Nationale de France and the Imperial War Museum. But not every country or institution has access to the same resources to deliver this kind of digital presence for its cultural heritage. In the digital world our current collections landscape risks creating perverse incentives for historians that bring to mind the tale of the drunk looking for his lost keys under the lamppost – not because that is where they were lost, but because that is where the light is. A digital footprint takes vision to establish and resources to maintain: a report created for the US Mellon Foundation cited a cost of .22 USD per page (University of Michigan Digital Library Services, 2001), not including the additional significant investment required in hardware, user facing aspects of the system, and long term maintenance. Viewed in the context of the millions of pages that even a modest institution might hold, the scope of the challenge becomes all too clear.

The most visionary project trying to bring change in this space has most certainly been the Europeana Digital Library.[1] As an instigator for the digitization of Europe's cultural heritage and as an aggregator for that content, Europeana has been a spectacular success, currently boasting access to millionsof digital records through its system. Europeana was not designed as a system for research, however, and many of the elements of the system that were optimized for breadth will make it difficult for the system to support the depth required for advanced research. The initial Europeana data model, based on the Europeana Semantic Elements (ESE) established a very low barrier to entry for national libraries and other collection holding institutions, but this low barrier did not guarantee that the interpretation of particular fields, or indeed the data entered into these fields themselves, would be standard or accurate. The ESE also didn't adapt well to the integration of non-library data types, such as archival runs. The 'one record, one digital image' structure, while powerful, made it difficult to represent hierarchies and relationships within document runs, leading to the very low representation of archival data within the current Europeana. The current move to implement a much more robust data model, the Europeana Data Model (EDM) will hopefully address some of these challenges for the future, making the collection far richer and more usable for research purposes. It will be a long time, however, before the Europeana legacy data can be brought to this standard, and, in the meantime, new approaches and sources to aggregation and federation continue to arise.

Is this proliferation a bad thing, or an avoidable one, however?

The answer is complicated. Indeed the landscape is crowded with digital projects, many of which are conceived in isolation from each other, and may be unaware of their own reinvention of proverbial wheels. The reasons for this are many: the conception of projects often is not as informed as it should be. Visibility for projects in the inception or development phase is either difficult or disincentivised, as projects need to distinguish

[1] www.europeana.eu

themselves from others in a competitive funding landscape. In addition, digital humanities projects have traditionally been conceived of to address very specific use cases, making the reuse of either their data or infrastructure difficult for successor projects. But not all of the reasons for non-reuse are avoidable or bad. Just as the technical development is becoming more and more focused on user-centered or indeed participatory design, so also must we recognize that the advanced uses to which digital humanists want to put their digital data sometimes command a differentiated approach.

To give an example of this from the CENDARI project (also known as the Collaborative EuropeaN Digital Archival Research Infrastructure,[2] which I will discuss in more detail later on), the charge to the project was to create a digital archival research infrastructure for medieval and modern history. There seemed to be an underlying assumption within this charge that modern and medieval historians, as representatives of one and the same discipline, would have roughly similar requirements. Yet the progress of the CENDARI design process proved again and again how different both the material records and the methodologies of these two communities would turn out to be. Some of this is a matter of the longer time step between the researcher and his or her object of study: medievalists, for example, have far more access to digital source material, at least in part because there has been a longer tradition of work on it. But many of the differences we had to accommodate were present in the analogue practices of the communities as well: our cohort of medievalists emphasized the importance of the item level and the library holdings, while the modernists privileged taking a view from the collection level over primarily archival holdings. Even the permeability of the disciplinary methodologies by those from other fields was subject to different attitudes within the two cohorts. Some of this difference is based on the state of methods, some on the state of collections: but even for the CENDARI historians it became quickly clear that 'one size' would not 'fit all.'

Given this need to balance the desire for technical alignment with recognition of different requirements, and against this backdrop of traditions and accidents, the CENDARIproject has been looking to introduce some sensitive and targeted innovations: to enable the federation of heterogenous data types (the making of what we call 'data soup'); the avoidance of high up-front investment in digitisation and metadata creation typical of digital libraries and archives and other pitfalls of library paradigms; the provision of strong incentives to modify the habits of scholarship and the overcoming of a 'consumerist' attitude toward technology among humanist researchers, privileging instead the makingof system 'decisions' transparent and interchangeable, rather than something to be accepted as a 'black box.'

In this way, we seek to create a fit-for-purpose, flexible cyber-infrastructure for historians. But this term of 'cyber-infrastructure for research' probably needs somewhat more explanation.

At its most basic level, a research infrastructure should allow finite individuals to achieve beyond their individual capacity to know, to do, to see, and/or allow valuable resources to be leveraged widely and publicly. I have written elsewhere about models of infrastructure and how they are applied in CENDARI (see Edmond 2013), so I will not rehearse those definitions and arguments again here. The concept we come back to again and again, however, it the desire to implement something 'below the level of the work' (see Edwards et. al.): that is, to support work as an outgrowth of current practices rather than forcing users to adapt their practices to fit the new environment. In this spirit, we developed

[2] www.cendari.eu

the following as a mission statement for the CENDARI project, emphasizing a three-fold emphasis in the project's underlying vision:
The CENDARI project's mission is to integrate digital archival resources for medieval and modern history, leveraging extant networks and projects to enhance the discoverability and usability of the resources.

With this strategic vision in mind, we can say with confidence that CENDARI is not a digitization project, a digital library/archive project or indeed a search and browse 'portal' project. What CENDARI is, however, exists at the intersection of a number of traditional boundaries, including:between content holders and content users; between technology and humanities; and between the digital and the analogue worlds of scholarship. These liminal spaces reflect the three communities that come together in our mission statement, and the imperatives that they place upon us in terms of supporting their scholarship.

None of this actually speaks to the functionality of the system, however, and to the manner in which these many imperatives are embodied in actual technology to support actual work.We imagine this functional system as a pipeline, running from data through to knowledge. The historian moves through an iterative cycle of finding data and interrogating, assimilating and otherwise coming to grips with that data, until she or he decides that new knowledge, worthy of sharing with the larger community of peers, has been created. At various junctures, the process may interface with other aspects of the world outside of the researcher's own processes, but this is not to say that the process is strictly linear or that it always starts and stops in the same place. The nature of knowledge creation is nomadic and iterative, and the CENDARI system must support this.

What this model represents is the manner in which the building blocks of the scholarly research process are conceived for the purpose of building technology to suit them, resulting in the following areas of emphasis:

- **Embedding the analogue processes** trusted by scholars and archivists at a deep level in the project
- Focusing on **exposing 'hidden collections'.**
- Instigating **adaptations** in the research ecosystem
- Creating a robust **'enquiry environment'** reflecting the journey from data to knowledge

The following concluding sections will discuss each of these emphases and how it has been delivered, in turn.

1. Embedding analogue processes

CENDARI cannot be successful unless it is trusted by scholars and archivists at a deep level. This trust will not come without mutual understanding a common denominator level of dialogue and activity, and must be grounded in the understanding that the digital environment can only supplement, rather than supplant, the analogue. As such, the analogue processes need to reappear transparently in the digital.

In order to understand what practices could be leveraged and what areas of frustration among historians could be addressed as a basis for further development, CENDARI's design was grounded upon an extensive set of participatory design techniques. Designed and implemented by project partners in INRIA, Paris and at the University of Göttingen, this process began with ideation meetings, in which three different groups of potential users were introduced to the kinds of processes they might expect to engage in within a digital

environment, and asked to create short video mock-ups of how they would foresee technology working for them.

Researcher scenarios were described at some length, and mapped according to both the set of tasks they described and a set of technical processes that might be deployed to support these tasks. Finally, two full prototype developments were undertaken to flesh out aspects of the project, resulting in the project annotation environment and intelligent meta-search functionalities. In the final phase of the project, the final implementation of the full technical environment, including collection search and browse, individual workspace and note-taking environment, and the production of archival research guides, are all being aligned and connected through both an internal and external facing process, using a 'Trusted User Group' as our first set of semi-external users. This development process has not been without its difficulties -- in particular, it is hard to set baselines and versions in the course of such rich dialogue with your users – but the utility of the final product will have been worth the extra effort.

2. Focusing on exposing 'hidden collections'

From the outset, CENDARI did not want to be a project that enhanced existing digital data without regard for the vast amount of historical record that was not yet available in digital form. As such, the commitment to exposing 'hidden collections' was very much within our mind from the outset.

To honour this commitment, however, took a lot of management. In the end, we developed four different basic pathways for ingesting data, and four different documents for use with different kinds of institutions at different points in our engagement with them. We also changed our collection strategy twice during the course of the project, each time to reflect our developing understanding of how much data would be missing from any 'comprehensive' digital infrastructure, and each time to better reflect the nature of the resources available and the model of transnational historical enquiry we felt a system like CENDARI could support. Much data that we expected to be easy to harvest either wasn't available, or wasn't available to us, but in the end we found that a thematic approach allowed us to provide the richest and most transparent possible overview of this state of affairs, highlighting undigitised or otherwise less known collections in a contextualised format, without losing sight of the scale of resource we had been tasked to create.

3. Instigating adaptations in the research ecosystem

Digital methods are not only changing the way in which we access and use resources, they are changing the affordances for how we communicate knowledge as well. Our initial work raised our awareness of this, and we began to view CENDARI not only as a new environment through which to view library and archival data, but also as a potential new form of scholarly commons.

As with all of our user centered work, this was not based upon a purely digital development: even without looking overly much at the affordances of the digital, John Guillory wrote of the habits of scholarship in terms of "the scholarly monograph" being "pulled apart and read like the Sunday paper" (Guillory, 2008).The moment in which CENDARI is being built is an exciting one, enabling us to envision a form of scholarly communication that not only teasesout the nuances of a scholarly argument, like the traditional monograph, but which also allows us to highlight new methodological approaches to personal memories, proxy records, incorporation of multimedia, etc; that harnesses linked

open data to better interact with the scholarship of others, and which can embed complex objects into scholarly outputs to enable a different sort of dialogue to emerge between the scholar and the reader of scholarship.

From this excitement, the concept of the CENDARI archival research guide (ARG) was born. The ARG was designed to inhabit a communications space between the finding aid and the monograph – but as an enhanced publication, we also wanted it to align to the CENDARI ontologies and harness the power of central disambiguation of persons, places and things, through linked open data sources like DbPedia and VIAF.

The technical format of the ARG has been relatively easy to design –indeed, in its final format, it should look very much like the CENDARI user's working environment at the close of an extended project. But imagining and paving the way for works like these to become accepted as scholarly currency has been far more difficult. The work will go on well past the CENDARI project to enable this, and to address the wider ecosystemic needs to reflect and protect humanistic knowledge creation processes, validate outcomes; and encourage sharing for both the faster and more efficient advancement of research as well as to underpin the development of better research tools.

4. Creating a robust 'enquiry environment' reflecting the journey from data to knowledge

Finally, we needed a technical environment that would reflect the desire to create an infrastructure model distinct from a digital library or archive. This technical model is complex, and has been described elsewhere in more detail (see Edmond, Bulatovic and O'Connor, 2015). But our service orientation and focus on leveraging linked data, rather than metadata, has led us to a configuration we feel will suit historians into the future, enabling as much technical assistance for knowledge creation as can reasonably be made transparent, while also allowing data sources, knowledge resources and investigation tools to be called and deprecated as needed, all around the model of a central data API gating the flow of information within and among the components of the system.

We need this flexibility in order to maintain the central role of the historian, the arbiter of interpretation and relevance in a world of data that is messy, incomplete, contradictoryand very open to interpretation. We hope that rather than trying to simply recreate the historian's current data environment in a technical box, CENDARI can instead reflect all of this complexity and multivalence, building a new tradition of digital scholarship true to its roots in the humanities.

Acknowledgements

The CENDARI project is the work of a large and distributed team across 8 countries and 14 institutions, and the work of many collaborators underpins the activity profiled above. The CENDARI project has been funded by the European Commission's Seventh Framework Programme, under Grant Agreement 284432.

References

[1] Edmond, Jennifer *The Role of the Professional Intermediary in Expanding the Humanities Computing Base,*Literary and Linguistic Computing, 20:3 (2005), 367–380

[2] Edmond, Jennifer, *CENDARI's Grand Challenges: Building, Contextualising and Sustaining a New Knowledge Infrastructure*,International Journal of Humanities and Arts Computing,7:1-2 (2013), 58–69

[3] Edmond, Jennifer, NatasaBulatovic, Alex O'Connor *The Taste of 'Data Soup' and the Creation of a Pipeline for Transnational Historical Research*, Journal of the Japanese Association of Digital Humanities, Forthcoming 2015
[4] Edwards, P, S. Jackson, G Bowker ,C Knobel*Understanding Infrastructure: Dynamics, Tensions and Design*http://hdl.handle.net/2027.42/49353 (Accessed November 2012)
[5] Guillory, John *How Scholars Read*, ADE Bulletin,146 (2008)
[6] Siemens, Lynne, Richard Cunningham, Wendy Duff, Claire Warwick *A tale of two cities: implications of the similarities and differences in collaborative approaches within the digital libraries and digital humanities communities*, Literary and Linguistic Computing, 26:3 (2011), 335–348
[7] University of Michigan Digital Library Services*Assessing the Costs of Conversion, A Handbook for the Andrew W. Mellon Foundation*http://www.lib.umich.edu/files/services/dlps/moa4_costs.pdf (Accessed March 2015)

edmondj@tcd.ie