doi: http://dx.doi.org/10.23925/1980-7651.2018v21;p63-70

The correspondence of Ferdinand von Mueller: from nineteenth century paper to twenty first century data

Gavan McCarthy[1]

Abstract

This paper presents a case study that demonstrates how a long term research activity, with the intention to create a scholarly edition of scientific correspondence, can be liberated from its print paradigm strictures to join the twenty first century world of interconnected knowledge. The Von Mueller Correspondence Project has produced a corpus of over 15,600 digitally transcribed letters and related materials focused on the period 1840 to 1896. These are complemented by materials in a range of forms that refer to Mueller dating from 1814 to 1931. Mueller was a prolific correspondent and established links with hundreds of fellow botanists and biologists across the globe; most of these, and certainly the most notable, will be registered in the History of Science Society Isis Cumulative Bibliography as Authority Records with links to publications about them and is some cases publications by them. The long-term plan is to systemically interlink the Von Mueller Correspondence Project digital corpus and the Isis Cumulative Bibliography and develop the synergies that will drive digital humanities analysis and future scholarly endeavour. That is the vision but what is the reality? At what stage is the project now? How did it get this far? What steps remain? How does the story of this project help us better understand the imperatives of digital scholarship – its strengths and its challenges?

Keywords

Computational HPS, Digital preservation, History of botany; History of Australian science, Global networks, Information security, Ferdinand von Mueller

[1] Director of the University of Melbourne eScholarship Research Centre, Australia. gavan.mccarthy@unimelb.edu.au. The present paper derives from a presentation at symposium Doing History of Science in a Digital, Global, Networked Community: Tools and Services Linking Scholars and Scholarship, 25th International Congress of History of Science and Technology, Rio de Janeiro 22-29 July 2017.

Introduction

This paper is based on a presentation given at a symposium at the 25th ICHST in Rio de Janeiro 25 July 2017. The topic of the symposium was, 'Doing History of Science in Digital, Global, Networked Community, Tools and Services linking Scholars and Scholarship'. This paper tells the story of how a long term research project of over 30 years has dealt with and has been changed by the emergence of digital tools and related digital services. It examines the journey from the nineteenth century paper world to the twenty first century world of data that is widely available to the 'masses'.

The Von Mueller Correspondence Project, which commenced in 1987, has produced a significant corpus of digitally transcribed documents.[2] The largest set of materials is the letters, to and from Mueller, from the period 1840 to 1896. These are complemented by materials in a range of forms that refer to Mueller dating from 1814 to 1931. In total there are over 15,000 items that have been transcribed. Each item, from the start of the project was created and maintained as a digital document. Mueller was a prolific correspondent and established links with hundreds of fellow botanists and biologists across the globe. Some of these individuals, and certainly the most notable, may already be the focus of scholars seeking to create similar published editions of their archival materials, e.g. Charles Darwin.[3]

Mueller was Victorian Government Botanist (1853-1896) and one of Australia's most well-known nineteenth century scientists, not just in Australia but across the globe.[4] He earned a considerable international reputation for his work in describing the Australian flora and in his pursuit of patronage. Mueller was also Director of the Melbourne Botanic Gardens (1857-1873) and played a major role in the establishment of the Herbarium at the Gardens as a centre of research. At the heart of his work was the dissemination of information about, and specimens of, Australian plants. This focus on the sharing of botanical knowledge remains a core activity of the Royal Botanic Gardens Melbourne and the Herbarium today, over 160 years later.

Mueller's interests lay not just in botany, but also in fields as diverse as geology, exploration and acclimatisation. He was also involved in the trade in stolen Australian Aboriginal and Torres Strait Islander ancestral remains. The Baron may be best understood as working in the field of nineteenth century European natural history with all its attendant motivations, not the least being European colonialism.

---

[2] "Correspondence of Ferdinand von Mueller project," Herbarium of the Royal Botanic Gardens Melbourne website, accessed 1 July 2017, http://www.rbg.vic.gov.au/science/herbarium-and-resources/library/mueller-correspondence- project

[3] "Darwin Correspondence Project," University of Cambridge, England, accessed 19 April 2018 https://www.darwinproject.ac.uk/

[4] "Mueller, Ferdinand Jakob Heinrich von (1825 - 1896) – Biographical Entry," *Encyclopedia of Australian Science*, eScholarship Research Centre, University of Melbourne, accessed 1 July 2017, : http://www.eoas.info/biogs/P000663b.htm

The Von Mueller Correspondence Project

In 1987, Professor Rod Home of the Department of History and Philosophy of Science at the University of Melbourne embarked on a project, with an international team,[5] to locate, transcribe, edit, annotate and publish, as a scholarly edition, 'the correspondence of Baron Ferdinand von Mueller.' An Advisory Group was established at the commencement of the project.[6] The objective of the project was to locate as much of the distributed extant outgoing and incoming correspondence as could be found in archives, universities, museums and herbaria across the world. The project was necessary as Mueller's original letter-copy books and much of his inwards correspondence, as a personal collection or archive, had been destroyed. This is believed to have happened in the 1930s but records that document this are scarce. Why these records were destroyed is still shrouded in mystery.

During the course of the project contact has been made with 100s of institutions within Australia and across the globe, seeking copies of letters and other relevant documents. Over the subsequent decades, since 1987, over 15,000 items have been located and transcribed. New items continue to emerge. It is believed that Mueller may have written well over 100,000 letters. However, the project did not attempt systematically to collect facsimile copies of the documents that were of publishable quality. Photocopies of the materials were taken using the readily available and affordable technologies of the day. Their purpose was to enable transcription and checking. More recently digital images of the materials have been captured. The photocopies were gathered purely to support transcription, quality assurance and reference, with no thought, due to perceived costs, that a facsimile edition would ever be a possibility. If the project started today, in a very different technological setting, this policy most certainly would have been different.

On the positive side, transcription does make the materials far easier to read, and the annotations provide essential context, background and explanation. Mueller was noted for his challenging handwriting and many documents used gothic German script limiting the utility of materials in original form. The value of the painstaking scholarly transcription and annotation cannot be overestimated.

Of the larger emerging corpus (the 15,000), some 750 selected documents (about 5%), were published in print form in chronological groupings. Some were made available on CD ROM.[7] However, it became increasingly clear, as the project continued, that it was not viable

---

[5] The project website lists the team as: Rod Home (The University of Melbourne and Royal Botanic Gardens Victoria); Sara Maroske (The University of Melbourne and Royal Botanic Gardens Victoria); Helen Cohn (Royal Botanic Gardens Victoria); Arthur Lucas (King's College London); Thomas Darragh (Museum Victoria and Royal Botanic Gardens Victoria); Monika Wells (Royal Botanic Gardens Victoria); and Johannes Voigt (Universität Stuttgart)

[6] The author was an original member of the Von Mueller Correspondence Project Advisory Group and remained in touch with the project although the Advisory Group only met once at the commencement of the project.

[7] Roderick Home, Arthur Lucas, Sara Maroske, Doris Sinkora and Johannes Voigt, eds, *Regardfully yours: selected correspondence of Ferdinand von Mueller, vol. 1: 1840–1859* (Peter Lang, Bern, 1988).
Roderick Home, Arthur Lucas, Sara Maroske, Doris Sinkora and Johannes Voigt, eds, *Regardfully yours: selected correspondence of Ferdinand von Mueller, vol. 2: 1860–1875* (Peter Lang, Bern, 2002)
Roderick Home, Arthur Lucas, Sara Maroske, Doris Sinkora, Johannes Voigt, and Monica Wells, eds, *Regardfully yours: selected correspondence of Ferdinand von Mueller, vol. 3: 1876–1896* (Peter Lang, Bern, 2006)

to publish in print form or on CD-ROM all of the found materials. As it emerged the fact that fewer and fewer computers had CD or DVD drives has validated this decision.

The work of locating, transcribing and editing continues to this day. The ongoing commitment of the editorial team, despite the initial funding being exhausted many years since, is testament to their dedication to the project. The editors set high scholarly standards and documented their processes in a detailed protocol document. This document has continued to evolve with the project. New letters and related documents continue to appear and new knowledge that helps in the understanding and annotation of the existing materials continues to be unearthed. This process of steady ongoing accumulation and explication of knowledge sits at odds with the vision of the project at its commencement in the late 1980s. It had been assumed that there would be an end-point and the product would be a 'definitive record'.

More recently, a digital online edition capable of being systematically updated became a feasible option. The hope is that a digital scholarly data-centric form would provide both the project team, and interested researchers from across the globe, with an updateable reference resource geared to the analytical expectations of twenty-first century digital humanities and scholarly practice.

Risks and challenges

It was in this context, in 2011, Professor Rod Home approached the eScholarship Research Centre (ESRC) at the University of Melbourne for 'advice on databases' in relation to the Von Mueller Correspondence Project. Although it was not clear what he had in mind, the back story was both more revealing and ultimately more worrying. The Mueller project team had been using Apple Macintosh Computers (Macs) since 1987 and had commenced by transcribing the photocopied source materials into Microsoft Word for Mac Version 1, which had originally been released in 1985.[8] The plan was to keep on migrating the files forward to current versions of Word for Mac and to replace the laptops and Mac operating systems as the project progressed. The request in 2011 was prompted by two events: the Mac used to hold the corpus was showing signs of not booting up correctly; and, the upgrade to the latest edition of Word for Mac revealed that a few hundred items were still in the earliest forms of Word. Disturbingly, it was shown that these early forms could not be brought forward to more recent Word versions without the loss of the styling and annotation used by the editors to mark-up the documents.

In other words it was possible that, in the worst case scenario, the whole corpus might be lost or that some of the editing and mark-up of the un-migrated files from beginning the project three decades earlier would have to be done again.

Although the editors, in particular Arthur Lucas, were interested in and aware of mark- up languages, such as SGML, at the outset they had not conceived the project as a

[8] "Microsoft Word," *Wikipedia,* accessed 19 April 2018, https://en.wikipedia.org/wiki/Microsoft_Word

digital humanities project but as a traditional print-based edited scholarly correspondence project with a series of books forming the final output. The choice of tools was incidental. The editors used digital technology that was readily available (Apple Macs and Microsoft Word) and with which they felt comfortable. It provided acceptable productivity, enabled them to work in different locations and helped them progress their print publication goals.

In 1987, they had not anticipated the time frame of the project, now entering its fourth decade, nor the dramatic changes in technology that would occur over that period. To deal with the information security 'crisis' facing the project, that is the incomprehensible but possible risk of the loss of all the 15,000 items due to the boot-up issue, a full copy of the digital corpus and related materials was copied to the ESRC project 'file share' located in the University of Melbourne Data Centre (a structured and managed digital object storage platform). Protocols were devised and enacted for batch updates as required. The original folder structure seemed fine but the file-naming protocols were identified as not ideal for cross-platform management of the documents. These are illustrated later. Fixing this was straightforward and dealt with programmatically.

The next issue, the upgrading of all old Word format files into a common and recent format, proved more challenging. For files in Word for Mac version 4 and later there was a migration path that could bring forward the text and the scholarly mark-up. However, this did not apply to files in Word for Mac versions 1 and 3.[9] Although the text could be brought forward the scholarly mark-up, the styles and footnotes, would be lost. Research by Chris Kirk, an Honorary Research Fellow at the ESRC, brought considerable computer skills to the task of assisting in the file migration process. He discovered, to our surprise and distress, that Microsoft had no records of the internal proprietary formats and coding schema used in the original Word for Mac releases which meant that programmatic conversion was not technically feasible without significant expense. In the end it was decided that it would be cheaper and quicker for the editors to re- style and re-annotated the few hundred items by hand. This proved to be the case.

Once a uniform corpus had been established in a recent form of Word for Mac, the next goal, in our quest for information security,[10] was to see if it was possible to transform the Word .doc files into an XML form to future proof the corpus against the archival negligence of Microsoft. That is, we needed a non-proprietary form suitable for long-term digital preservation. The TEI P5 (Test Encoding Initiative) XML schema, a mature standard well tested and used in the digital humanities was the first choice.[11] The expectation was that in this form the files could be held in digital repositories and ingested into search, retrieval and analytical engines. Experience by others showed that the files in this form could be readily transformed to HTML for web presentation as well as into digital print forms such as pdf or indeed back into Word. This process of enabling multi-model dissemination forms helped meet the key objective of future-proofing the decades of scholarly endeavour.

---

[9] There was no Word for Mac Version 2.

[10] "Information security," *Wikipedia*, accessed 19 April 2018, https://en.wikipedia.org/wiki/Information_security. I use this term in a broader sense than defined in Wikipedia to include the security of information through time ensuring that it is as immune as possible from the ravages of information entropy.

[11] "TEI: P5 Guidelines," *<Text Encoding Initiative>*, accessed 19 April 2018, http://www.tei-c.org/Guidelines/P5/

Although the Word .doc form was not amenable to these types of transform it was found that the styles and annotations were sufficiently explicit in the Word .docx form. The corpus was transformed by a batch process from .doc. to .docx. It then became the standard for the editorial team.

Testing, using the version of OxGarage Conversion available at the time, showed that these docx versions of the files should be able to be successfully transformed into TEI P5.[12] Sadly, things were not quite that simple. The OxGarage tools and services were not geared to deal with the type and complexity of mark-up employed by the Mueller editors. However, we had established a practical foundation whereby the editors, some now in their seventies, were still able to keep working on the corpus in Word, the technology of their choice, knowing that it looked feasible to take the next step. The idea was to build a batch-processing pipeline that could convert the whole corpus to TEI P5 as required. At this point, realising that I had neither the time nor the skills, we engaged the services of an XML and TEI expert, Conal Tuohy (based in Brisbane), to join the team.

The first task was the development of a Word-TEI transform processing pipeline – or as Conal called it the 'VMCP upconversion'.[13] This task was managed in an agile manner using a sequence of iterative small steps. First, it was established that the basic process worked for the most common styles. Conal, working closely with Arthur Lucas (based in the UK), then extended the transform to cover the increasingly uncommon and peculiar variables, instances and exceptions that became evident. By the time we got to embedded tables we finally hit the point determined by 'the law of diminishing returns' where we had to accept that the expectations of the editors, the capabilities of the technology and the virtually non-existent budget would result in compromise.

The second task was the selection of a tool or service that would enable the presentation, exploration and examination of the entire corpus. After some discussion, Conal recommend the open source eXtensible Text Framework (XTF) produced and supported by the California Digital Library.[14] What was surprising was that there were very few options available from which to choose. An instance was established for the Mueller TEI corpus utilising the non-styled or 'out of the box' presentation interface. Our goal, at this stage, was to explore the functionality of XTF and its use in expressing the TEI mark-up in a way that was useful for the editors with a particular emphasis on data quality assurance. So despite the underlying XTF code being based on an earlier version of HTML, not HTML 5 as we would have hoped, it still seemed the best choice. As a mature product it was easy to 'run up' and has proved to be stable and transferrable to virtual machines on different servers.

As mentioned, a standardised folder structure, in conjunction with standardised file-naming, was established early in the life of the project by the editors. e.g. Folder = '1840-9'; File name = '40-05-06-final.docx'. It evolved slightly through the life of the project into a stable and familiar form that was readily usable by the editors and sufficiently explicit to

---

[12] "OxGarage Conversion," *<Text Encoding Initiative>*, accessed 21 July 2017, http://www.tei-c.org/oxgarage/. See also: GitHub: https://github.com/TEIC/Oxgarage, accessed 21 July 2017.

[13] "VMCP-upconversion," *Github*, accessed 19 April 2018, https://github.com/Conal-Tuohy/VMCP-upconversion

[14] "eXstensible Text Framework," California Digital Library, accessed 19 April 2018, https://www.cdlib.org/services/access_publishing/publishing/tools/xtf/

those outside of the project to see what was going on. It was proposed to use this structure for the TEI files, e.g. '40-05-06-final.xml'. In that way, there is a 1:1 mapping of the .docx folders and files to the TEI folders and files. It was argued that the use of identical folder and filing structures for different representations of the same information was pragmatic, and sensible. It certainly kept things simple and therefore understandable and easily processed. To facilitate the work of the editors the VMCP-upconversion was set to run each evening which meant that each day's work could be checked in XTF.

Documents as data

A straightforward introduction to the new processes was put together by Conal.[15] This website provides a link to the XTF repository, log files, the raw xml files, the VMCP-upconversion code as well as information about the styles and how they can be utilised as facets in XTF. A facet, in this context, is the systematic and consistent encoding of data that enables the data set (or corpus) to be filtered to only reveal those items that meet the criteria of the selected facet or facets. The Von Mueller Correspondence Project top level facet groups are: Filename; Validity; Status; Author; Features; Styles; Addressee; Language; Plant Names; and Date. Each of these high level facets is represented by varying range of instances. For example: for Validity, a document is either valid or invalid; for Author, the range or number of possible authors is not prescribed; for Date the documents are further faceted by decades. The calculation of the documents that meet the conditions of the facets is calculated each time the corpus is updated (that is daily).

The instances of each top level facet are presented in a navigational pane that enumerates the number of occurrences (that is the number of documents that have that particular instance of a facet). This proved to be a very important quality assurance feature as it enabled me, Conal and the editors to start to see and comprehend the corpus as a whole. For Rod, Arthur and the team, this knowledge of the corpus as a whole was burned into their souls, but for outsiders, like me, it was a revelation. Personally, when I saw this for the first time I was genuinely excited – at last we had a tool that would enable us to manage this large complex and potentially unwieldy collection of stuff. It marked the start of the next generation of the project, in a future-proof form with a product that could be handed onto the next generation of Mueller scholars to maintain and further develop. It was also a product that would be useful for others to use for their own research.

In addition to filtering by facets, once the TEI P5 XML forms of the documents are loaded into the XTF repository, XTF provided a 'free-text' search function which has the effect of creating a filtered subset which can be revelatory but never comprehensive in the same manner as the facets based on the editorial styles and annotations. Shown here is the top part of the results based on a search on the term 'Brazil'. The search returned a sub-set of 57 files of which 34 have been designated 'final' by the scholars.

As a further example, a narrower search based on 'Rio de Janeiro' was done revealing

[15] "Ferdinand von Mueller correspondence." Conal Tuohy, accessed 19 April 2018, http://vmcp.conaltuohy.com/

a sub-set of 5 files. The effectiveness of 'search' is directly related to the quality and consistency of the data, which in turn is a product of the diligence and care of the scholars in their transcriptions. The 'search' of the TEI in XTF exposes the quality of data – which to me seems very good overall, engendering confidence in corpus and its use for further research purposes. Mueller was in Rio de Janeiro in September 1847 for two weeks while en route to Adelaide. He used the time to collect botanical specimens. Sadly, these were destroyed in a storeroom fine in Adelaide so they never made it to Melbourne. All this I learned in a matter of minutes, from a quick study of the annotated letters mentioned above. However, without the annotation the overall story would have remained obtuse and hidden.

A TEI – XTF presentation format was devised by Conal, in conjunction with Arthur, specifically to aid in the editing, checking and quality assurance of the data. Technically, this is the TEI XML expressed in HTML within XTF. The thing to note here is that the footnotes follow the paragraph to which they refer, as the concept of a 'page' does not make as much sense in this form. This file is a good example of the structural (Sartrean) unpredictability that the original documents may take.

My observation here is that I find it a lot easier to read and understand the XTF– TEI form than the print version represented in the traditional scholarly book-style typeset representation.

Although not all would agree, the point is that we are not restricted to just one representation. Interestingly, the immediate and spontaneous feedback was that the mark-up was clear (obvious), the data explicit, and very usable.

Conclusion

The representation of scholarly knowledge as human readable and comprehensible information was the goal of the print edition to which has been added its representation as computer analysable data. To enable this transformation additional tools and expertise were required. The transformation was dramatic and radically changed the work of the original team of scholars. It has opened up the reach of the Muller Correspondence and created data ready for the next generation of digitally-oriented Mueller enthusiasts, including those involved in: computational history and philosophy of science; digital preservation; history of botany; history of Australian science; and local and global networks.

It is in relation to this last point it is argued that IsisCB Explore could become the focal point for connecting Mueller data into the wider world of the history of science and this was a topic of conversation in Rio at the Congress.

It is hoped that we have created a future-proof version of the content that will be accessible and processable by generations of scholars and scientists from here on. And that we have found a way that will allow the editorial team to hand on their work with confidence. So they can finally retire – or keep working if they want to!