The History of Science in the age of networked digital humanities

This session reports on a significant information infrastructure project in the history of science as the discipline prepares itself to deal with the major scholarly and societal challenges of the twenty-first century. History of science scholars are seeking new analytical tools and research methodologies and are looking to computer aided tools that utilise well-structured data. The session is drawn from participants in an international collaborative project between the History of Science Society, the University of Oklahoma, the University of Melbourne and the Division of History of Science of the International Union of the History and Philosophy of Science. Simply stated, the IsisCB project is taking 100 years of history of science bibliographic data and transforming it into a web or graph of information objects that represent not just the cited works but more importantly, the people, organisations, events, subjects, themes, and time periods of the world captured in the work of the historians. The project involves, digitisation of print only materials, optical character recognition, computer aided parsing, semantic transforms, large-scale data transfer, industrial-scale search and faceted browse facilities, visual analytics, application programming interfaces and user-facing design. It involves taking a community and a well-established print-focused way of working on a revolutionary change process to create an open, persistent citable high quality data source that will be amenable to utilisation using linked data protocols and technologies.

The history of science is a global, multidisciplinary endeavour that seeks to help us understand one of the most potent forces that has shaped contemporary society, namely science. Although often represented as a discipline in its own right, practitioners are frequently located in other disciplinary settings such as history, philosophy, population health, medicine, social science, media and communication studies. Its themes and studies, while often starting with a local story, inevitably, like the science itself, document networks and interconnections that cover all continents. In the early twentieth century a community of practice began to emerge in Europe as historians of science started to formalise their own networks and communication channels. In 1913 they created *the Isis Current Bibliography of the History of Science* and it has been maintained as an annual print-based bibliography ever since. The latest edition was published by the History of Science Society in 2014 and included references to over 4000 scholarly works. The total corpus will comprise hundreds of thousands of records and millions of relationships and will continue to grow. The three papers will be presented by members of the IsisCB project working group and will cover a range of the technical and informatic challenges, and explore the opportunities for the history of science discipline emerging as a result of the development of the IsisCB web resource.

Mapping the History of a Discipline through Curated Bibliographic Data: A Case Study of Social Networks as Found in the Isis Bibliography of the History of Science, 1913-2013

Stephen P. Weldon (University of Oklahoma, Norman, Oklahoma, United States),
Sylwester Ratowt (University of Oklahoma, Norman, Oklahoma, United States),
Birute Railiene (Wroblewski Library of the Lithuanian Academy of Sciences, Vilnius, Lithuania)
John Stewart (University of Oklahoma, Norman, Oklahoma, United States)

The 100-year-old Isis Bibliography is a valuable data resource for exploring the development of the field of history of science over the course of the twentieth century. The recent collaborative project between the Isis Bibliography of the History of Science and the eScholarship Research Centre (ESRC) at the University of Melbourne has created a robust digital research dataset consisting of over 300,000 bibliographic citation records going back to 1913. The new tool, called the IsisCB Platform, is more than a bibliographic database; it makes use of bibliographic data, but supplements that data with links to other content using Linked Open Data protocols.

To create this tool, we have had to ingest new data and rethink how bibliographic data can be used in a big-data, computational environment. Part of the IsisCB project involved digitizing sixty-years of data in several volumes of cumulative print bibliographies. We are now parsing and merging this data with forty years of born-digital data. The combined bibliographic dataset covers one hundred years and spans the period in which history of science developed as an academic discipline. By reconceptualizing this dataset as a resource linking publications, institutions, and people, it becomes an information-rich resource that can be used to tell the story of scholarship in this area over the twentieth century.

This paper serves two functions. First, it introduces the IsisCB Platform and explains how the bibliographic dataset has been reconceptualized so as to create a scholarly tool with multiple research capabilities. Second, it provides a brief example of how a subset of this data can be used to explore the history of the discipline itself. By using computational tools, we can address historical questions about the social, institutional, and intellectual history of the scholars themselves. It will be shown how the IsisCB data can serve as the foundation for digital humanities research when it has been thoughtfully restructured in the right way.

Bibliographic records contain extensive information about the institutional and publication information of people. Information scholars working on social network analysis have developed various ways of comparing and understanding multiple kinds of networks that can be produced from bibliographic data (Batagelj, Cerinšek, Batagelj, and Cerinšek 2013). We will be employing some of these methods. Recent work in social network analysis has shown the utility of bibliographic datasets for exploring history of disciplinary change in the humanities (So and Long 2013). Such an analysis has not yet been done to any great extent in the field of history of science, but the field is ripe for this kind of work (Laubichler, Maienschein, and Renn 2013).

The composition of the Isis Bibliography dataset over the years makes this analysis both possible and exciting. The Isis Bibliography contains full author, title, and citation information for each of its entries. Unlike some bibliographic databases, the Isis Bibliography also has classifications or subject index terms for most of the items, along with descriptive material for many of them. Moreover, because there have been relatively few bibliographers over the years, there has been a great deal of consistency in the categorization and tagging the records (Weldon 2013a).

Finally, the Isis Bibliography contains the best set of information on the academic network of historians of science. The first installment of the Isis Bibliography was one of the critical moments in the founding of the discipline. George Sarton published his bibliography in 1913 in the first issue of his journal *Isis*. Over the next couple of decades Sarton and several other scholars in Europe and America began producing bibliographies, publishing journals, and holding conferences. These would ultimately lead to the creation of a small but dedicated community of scholars focused on the study of the history of science. As the discipline grew, so did the Isis Bibliography.

We will conclude our paper by looking at how data in the IsisCB can illuminate historical questions. We will build a few network graphs to help us analyze the scholarly genealogy of the discipline. We will try to tease out some of the stories that lie buried within it: Which universities and research institutes have been instrumental to the growth of the discipline? How so? In what ways has gender played a role in the development of the discipline? Do the data help us identify national styles? In the end, the paper will showcase preliminary results of our analysis to show how bibliographic data can be applied in new and unique ways when the data are made available in accessible ways.

References:
Alfonso-Goldfarb, A. M., Waisse, Silvia, & Ferraz, M. H. M. (2013). From Shelves to Cyberspace: Organization of Knowledge and the Complex Identity of History of Science. *Isis*, *104*(3), 551–560. doi:10.1086/673274

Allen, C., & Group, the I. (2013). Cross-Cutting Categorization Schemes in the Digital Humanities. *Isis*, *104*(3), 573–583. doi:10.1086/673276

Alsukhni, M., & Zhu, Y. (2012). Interactive visualization of the social network of research collaborations. *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*. doi:10.1109/IRI.2012.6303017

Anderson, R. J. (2013). The Organization and Description of Science Archives in America. *Isis*, *104*(3), 561–572. doi:10.1086/673275

Batagelj, V., & Cerinšek, M. (2013). On bibliographic networks. *An International Journal for All Quantitative Aspects of the Science of Science, Communication in Science and Science Policy*, *96*(3), 845–864. doi:10.1007/s11192-012-0940-1

Coscia, M., Giannotti, F., & Pensa, R. (2009). Social Network Analysis as Knowledge Discovery Process: A Case Study on Digital Bibliography. *2009 International Conference on Advances in Social Network Analysis and Mining*. doi:10.1109/ASONAM.2009.65

Jie Tang, J. T., Duo Zhang, D. Z., & Limin Yao, L. Y. (2007). Social Network Extraction of Academic Researchers. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. doi:10.1109/ICDM.2007.30

Laubichler, M. D., Maienschein, Jane, & Renn, J. (2013). Computational Perspectives in the History of Science: To the Memory of Peter Damerow. *Isis*, *104*(1), 119–130. doi:10.1086/669891

Mattison, D. (2010). The twittering of the search world.(LiveLinks). *Searcher*, *18*(7), 24.

Nicholas C Mullins; Carolyn J Mullins. (1973). *Theories and theory groups in contemporary American sociology*. Harper & Row.

Schatten, M. (2013). What do Croatian Scientist Write about? A Social and Conceptual Network Analysis of the Croatian Scientific Bibliography. *Interdisciplinary Description of Complex Systems*, *11*(2), 190.

Sidiropoulos, A. (2012). Finding Communities in Site Web-Graphs and Citation Graphs.

So, R. J., & Long, H. (2013). Network Analysis and the Sociology of Modernism. *Boundary 2*, *40*(2), 147–182. doi:10.1215/01903659-2151839

Weldon, S. P. (2013a). Bibliography Is Social: Organizing Knowledge in the Isis Bibliography from Sarton to the Early Twenty-First Century. *Isis*, *104*(3), 540–550. doi:10.1086/673273

Weldon, S. P. (2013b). Introduction. *Isis*, *104*(3), 537–539. doi:10.1086/673272

Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326. doi:10.1002/asi.22680

Zhixing Huang, Z. H., Yan Yan, Y. Y., Yuhui Qiu, Y. Q., & Shuqiong Qiao, S. Q. (2009). Exploring Emergent Semantic Communities from DBLP Bibliography Database. *2009 International Conference on Advances in Social Network Analysis and Mining*. doi:10.1109/ASONAM.2009.6

Why standards are critical: Graphing knowledge by building on well-established entity-relationship standards can look well into the future

Ailie Smith (The University of Melbourne, Australia)
Marco La Rosa (The University of Melbourne, Australia)

Through its history as the Australian Science Archives Project (1985 to 1999), and the Australian Science and Technology Heritage Centre (1999 to 2007) (eScholarship Research Centre, n.d.), the eScholarship Research Centre (ESRC) at the University of Melbourne has well-established links with the international history of science community. Since 2014 the ESRC has been collaborating with Dr. Stephen Weldon from the University of Oklahoma on the Alfred P. Sloan Foundation-funded project – the Isis Document Indexing Platform: A Curated and Community-Based Resource for History of Science.

The ESRC has been using the World Wide Web to publish and disseminate history of science and other humanities datasets including bibliographic data, archival descriptive data and contextual information since the mid-1990s. Examples of these are web resources such as the *Frank Macfarlane Burnet Guide to Records* (McCarthy et al., 2001), and the *Encyclopedia of Australian Science* (eScholarship Research Centre, 2012), which includes bibliographic records from the 'Bibliography of the History of Australian Science', published annually in *Historical Records of Australian Science* (Cohn, 2014).

These resources are created from data captured in two systems developed by the ESRC – the Online Heritage Resource Manager (OHRM) and the Heritage Documentation Management System (HDMS). Both of these systems are based on international archival descriptive standards (ISAD(G) and ISAAR(CPF)) (International Council on Archives, 2000 & 2004) and have long been used for the generation of HTML outputs for dissemination of information as web resources. However these outputs are not readily sharable with other organisations or other resources operating in a similar sphere, resulting in limited reuse of this data and increased potential for duplication of work.

The increasing use of Extensible Markup Language (XML) schemas by information professionals to enable system interoperability, data sharing and data reuse has provided the ESRC with greater opportunities to collaborate, to rethink the way we present information online, to share data with and harvest data from other systems and to represent and reinterpret data through the use of a variety of search and analytical tools.

By generating Metadata Object Description Schema (MODS), Encoded Archival Description (EAD) and Encoded Archival Context – Corporate Bodies, Persons and Families (EAC-CPF) (Library of Congress, 2013a & 2013b, Staatsbibliothek Zu Berlin, 2012) XML outputs for these resources in parallel with the web presentation HTML outputs, the ESRC can make better use of the data underpinning these resources. This includes not only the applications and services developed to display and navigate individual datasets, but also the capacity to build services that run across a range of datasets, and to harvest and share data with external sources at national and international levels to improve collaboration and extend the use of the data beyond individual research projects.

It was a natural progression for the ESRC to begin working with Dr. Stephen Weldon from the University of Oklahoma on the Isis Document Indexing Platform: A Curated and Community-Based Resource for History of Science project. The IsisCB web resource brings together one hundred years of bibliographic records from the *Isis Current Bibliography of the History of Science* ('About Isis CB') with the technology base of the ESRC to create a rich, searchable, contextualised online resource for exploring publications relating to the history of science. Working with the researchers from the University of Oklahoma, the ESRC is helping to create XML records, develop ingests of XML and free-

text content and build on existing technology platforms to create search and browse interfaces for interacting with the bibliographic data. This work will take resources that have previously been available in static print formats and allow them to be explored and recompiled in new ways, providing a valuable tool for historians of science.

The work on the IsisCB Platform builds on a range of technology developments that the ESRC has undertaken and collaborated on over recent years, making use of XML data. Internally, the ESRC has been developing indexing, search and faceted browse and filtering capabilities using XML data and Solr/Lucene technology, supported by high availability large scale server infrastructure that has been developed and continues to be supported in-house (La Rosa et al., 2014). In addition XML data are being used by the ESRC to develop flexible online presentation of humanities datasets that are targeted at the needs of specific user communities.

Externally, two of the ESRC's web resources began being harvested in 2010 by the National Library of Australia's *Trove* service (Dewhurst, 2010), which had adopted the EAC-CPF schema for harvesting information about people and organisations and their related records and publications (Dewhurst, 2008). XML data from a number of ESRC resources have also been harvested by the Humanities Networked Infrastructure (HuNI) project (Deakin University, 2014), and the underlying data structure has been mapped to the Registry Interchange Format - Collections and Services (RIF-CS) schema used by Australian National Data Service (ANDS) for their Research Data Australia registry (Australian National Data Service, n.d.). Through collaborations with other professionals – both within and beyond information professions – the eScholarship Research Centre is contributing to and developing research infrastructure, information resources and public knowledge spaces that widely communicate humanities research data.

With a focus on the development of the Isis Document Indexing Platform, this paper will investigate the ESRC's experience in the development and utilization of MODS, EAD and EAC-CPF XML data exported from its systems, and the collaboration with other researchers to develop XML outputs from their own systems. It will also look at a range of tools and services that leverage the standards-based XML outputs to allow researchers to explore, interrogate, and present the data in order to reinterpret and derive new understandings of it, and the technologies that underpin these services. It will also discuss the potential to better engage with public search interfaces such as Google by including schema.org markup metadata in web templates in order for resources to reach a wider audience (Starr, 2014). The paper will explore the way the standards-based approach will help to create the Isis Document Indexing Platform as a sustainable history of science resource well into the future.

References:
'About Isis CB', *ISIS Current Bibliography*. http://isisbibliography.org/about/ (accessed 3 November 2014).

Australian National Data Service (no date). *Research Data Australia*. http://researchdata.ands.org.au/ (accessed 3 November 2014).

Cohn, H. M. (ed), (2014), 'Bibliography of the History of Australian Science, No.34, 2013'. *Historical Records of Australian Science*, volume 25, number 1, pp.123?141.

Deakin University (2014). *HuNI Virtual Laboratory*. https://huni.net.au (accessed 3 November 2014).

Dewhurst, B. (2008) 'People Australia: a Topic-based Approach to Resource Discovery'. *VALA2008 14th Biennial Conference and Exhibition*. http://www.valaconf.org.au/vala2008/papers2008/116_Dewhurst_Final.pdf (accessed 3 November 2014).

Dewhurst, B. (2010). 'Encyclopaedia of Australian Science available via Trove'. *ARDC Party Infrastructure Project Blog*, 7 May 2010. National Library of Australia.

https://wiki.nla.gov.au/display/ARDCPIP/2010/05/07/Encyclopaedia+of+Australian+Science+available+via+Trove (accessed 3 November 2014).

eScholarship Research Centre (2012). *Encyclopedia of Australian Science*. Retrieved from: http://www.eoas.info (accessed 3 November 2014).

eScholarship Research Centre (no date). *Our History*. http://www.esrc.unimelb.edu.au/about?us/our?history/ (accessed 3 November 2014).

International Council on Archives (2000). *ISAD(G): General International Standard Archival Description, Second Edition*. http://www.icacds.org.uk/eng/ISAD(G).pdf (accessed 3 November 2014).

International Council on Archives (2004). *ISAAR (CPF): International Standard Archival Authority Record For Corporate Bodies, Persons and Families, Second Edition*. http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf (accessed 3 November 2014).

La Rosa, M., McCarthy, G. & Smith, A. (2014). Nothing can ever be lost: Infrastructure for Archival Research and Dissemination. *eResearch Australasia 2014*. http://eresearchau.files.wordpress.com/2014/07/eresau2014_submission_29.pdf (accessed 3 November 2014).

Library of Congress (2013a). *Encoded Archival Description: Version 2002 Official Site*. http://www.loc.gov/ead/ (accessed 3 November 2014).

Library of Congress (2013b). *Metadata Object Description Schema: Official Web Site*. http://www.loc.gov/standards/mods/ (accessed 3 November 2014).

McCarthy, G., Manhal, O., O'Sullivan, L., Tropea, R., & Sherratt, T. (2001). *Frank Macfarlane Burnet Guide to Records*, Australian Science and Technology Heritage Centre, Melbourne. http://www.austehc.unimelb.edu.au/guides/burn/burn.htm (accessed November 2014).

Staatsbibliothek Zu Berlin (2012). *Encoded Archival Context Corporate Bodies, Persons and Families*. http://eac.staatsbibliothek?berlin.de/ (accessed 3 November 2014).

Starr, B. (2014). 'Demystifying The Google Knowledge Graph'. *Search Engine Land*, 2 September 2014. http://searchengineland.com/demystifying?knowledge?graph?201976 (accessed 3 November 2014).

The case of the missing records: The long journey of the Correspondence of Ferdinand Von Mueller into the world of digital history of science.

Gavan McCarthy (The University of Melbourne, Australia)

Within the context of the development of the Isis Cumulative Bibliography Online (IsisCB) as a global information infrastructure service for the history of science community, this paper presents a case study that demonstrates how a long term research activity, through the creation of 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 15,000 digitally transcribed letters 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. 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 IsisCB as Authority Records with links to publications about them and is some cases publications by them. The plan is to systemically interlink the Von Mueller Correspondence digital corpus and the IsisCB and develop the synergies that will drive digital humanities analysis and future scholarly endeavour.

In 1987 Professor Rod Home of the Department of History and Philosophy of Science at the University of Melbourne embarked on a project to locate, edit and publish the correspondence of Baron Ferdinand von Mueller. 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. 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 and the dissemination of Australian plants and botanical knowledge. These roles remain the foundations of the Gardens and the Herbarium today, over 160 years later.

The major challenge of this project, that included collaborators from the United Kingdom and Germany, was to locate as much of the outgoing correspondence as could be found in archives, universities, museums and herbaria across the world. This need to reconstruct the Mueller 'archive' resulted from the destruction of the locally held letter-copy books, inward correspondence and associated files about three decades after Mueller's death in 1896. Why these records were destroyed is still shrouded in mystery. Subsequent research revealed that Mueller may well have sent out well over 100,000 letters during his career. The research team, over the last 25 years, have managed to locate, copy, transcribe and prepare for scholarly publication over 15,000 items. Over 750 selected documents have been published in print form and some made available on CD ROM. (Home et al 1998, Home et al 2002 and Home et al 2006). This work still continues today due to the continuing commitment of the scholars despite the initial funding being exhausted many years since. 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 accumulation and explication of knowledge sits at odds with the original print-based paradigm that configured the project at its commencement in the late 1980s. It would appear that an online edition capable of batch-processed updating could provide both the research team and scholars from all nations with a reference and analytical resource geared to the expectations of twenty-first century digital humanities.

In 2011, Professor Rod Home approached the eScholarship Research Centre (ESRC) at the University of Melbourne for 'advice on databases' in relation to the Mueller Correspondence project. However, the subtext was both more revealing and ultimately more worrying. The Mueller project team had been using Apple Macintosh Computers since 1987 and had been transcribing the photocopied source materials into Word for Mac Version 1 which had originally been released in 1985. 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 in the earliest forms of Word and that they could not be brought forward without the loss of the styling and annotation used by the scholars to mark-up of the letters. Although the scholars were interested in and aware of mark-up languages, such as SGML, at the outset of the project they had not conceived it as a digital humanities project but as a traditional edited scholarly correspondence project with a print form being the final output. The scholars used digital technology that was readily available, improved productivity and helped them achieve their print publications. They had not anticipated time frame of the project, now entering its fourth decade, nor the changes in technology that would occur over that period.

To deal with the 'crisis' facing the project (that is the potential loss of the 15,000 items), a full copy of the corpus and related materials was copied to the ESRC project 'file share' (a structured and managed digital object storage platform) and protocols enacted for batch updates as required. The original folder structure and file-naming protocols were problematic in a generalised digital world so changes were instigated. However, the next key issue was the upgrading of all old format files into a recent format that would enable batch processing of the corpus in the future. Research by Chris Kirk at the ESRC revealed that Microsoft had no records of the internal proprietary formats use in the original Word for Mac releases which meant that programmatic conversion was not technically feasible without significant expense. In the end it was cheaper and quicker for the scholars to re-style the few hundred items by hand.

Once the uniform corpus had been established a new goal was set to transform the .doc files into TEI P5 (Test Encoding Initiative) XML schema form, a non-proprietary form suitable for long-term digital preservation; a form that could be ingested into search, retrieval and analytical engines such as Solr/Lucene; and, a form that could be readily transformed to HTML for web presentation. Although the .doc form was not amenable to these types of transform it was found that the styles and annotations employed by the scholars were adequately preserved in the .docx form and that using Oxgarage these versions of the items could be successfully transformed into TEI P5. At the time of writing the pipeline processes for the full transforms are being prepared. In the meantime the lead scholars, now in their seventies, are still able to keep working on the corpus, adding and refining their knowledge in technologies they are comfortable using.

Despite what appeared to be two cases of the destruction of records that could have resulted in considerable loss of historical knowledge, the dedication of a few scholars will result in a digital corpus of significance to the history of science in Australia. In addition it will also reflect on the social life of the Australian colonies in the nineteenth century; contribute to research into Australian flora and the study of our natural environment; and, contribute to the characterisation of Australian colonial administration and the relationship of Australian scientists to the national and international scientific communities. Moreover, as a suite of over 15,000 uniquely and globally identified and citable XML objects the next generation of scholars using annotation and mark-up tools will be able to continue to add value to the collection. The transform of the Isis bibliographies into an informatically similar form will enable the two systems to share data and create many intersections. The new world of interconnected data sets and computable social networks is not something that the Mueller scholars could have envisaged in 1987. For Mueller himself, it was his world and those networks and connections were held in his mind and externalised for posterity in his records. With a little care and attention from digital humanities curators and the emergence of robust digital preservation and publication platforms, the knowledge conceived in the critical scholarly print form will have its value multiplied many times over.

References:

Home, R.W., Lucas, A.M., Maroske, S., Sinkora, D.M. and Voigt, J.H. (eds) (1998*). Regardfully yours: selected correspondence of Ferdinand von Mueller, vol. 1: 1840–1859*. Peter Lang, Bern.

Home, R.W., Lucas, A.M., Maroske, S., Sinkora, D.M. and Voigt, J.H. (eds) (2002*). Regardfully yours: selected correspondence of Ferdinand von Mueller, vol. 2: 1860–1875*. Peter Lang, Bern.

Home, R.W., Lucas, A.M., Maroske, S., Sinkora, D.M., Voigt, J.H. and Wells, M. (eds) (2006). *Regardfully yours: selected correspondence of Ferdinand von Mueller, vol. 3: 1876–1896*. Peter Lang, Bern.