# MODS: Mapping knowledge with data science

2017 marks the centenary of the enrolment of the UK's first doctoral student at Oxford (Simpson 1983), making it an opportune moment to take stock of the impact that the PhD has had on the British academic landscape. This proposal builds on the availability of EThOS, a unique data set of more than 450,000 PhD completions held by the British Library, to evaluate the ways in which a doctorate both traces and influences the flow of ideas within and between institutions. At a time of great uncertainty for UK Higher Education (HE), understanding the geography of innovation embodied in these outputs and their cross-cutting linkages offers a rare opportunity for a student to apply cutting-edge data science techniques to a question with fundamental policy implications: how does ground-breaking work (whether theoretical or empirical) in the natural and social sciences, as well as the humanities emerge from supervision within particular institutional contexts, and how is it disseminated over time and space to other institutions?

## CONTEXT

The project takes as its starting point the idea that the PhD constitutes a uniquely formative period in the 'life of the mind': the dissertation will typically provide the foundation for the early publications and technical expertise upon which subsequent career progression is founded. In short, the PhD positions the candidate within a kind of academic lineage, or genealogy, that they will in turn transmit to their own students. There is thus a complex relationship between doctoral research and the flow of ideas between individuals and institutions (in all forms). We believe that, by examining how institutions and their geographical placement and interaction affects thesis production, it becomes possible to analyse both the role of institutions in promoting and sustaining innovation, and the contributions to knowledge transfer made by individuals who move between institutions.

In so doing, the project joins up—and creates additional opportunities to deepen the links between—our understanding of the evolution of the PhD thesis (see: Simpson 1983), work on 'scientometrics' and our ability to trace knowledge flows, and our appreciation of when, where, and how innovation spreads. The project therefore addresses a major gap in the history of the PhD in Britain, which essentially ends in 1957 with the principal researcher expressing her "greatest wish that others will now take it on from [t]here" (Simpson 2009, p.xxxiv). But it also speaks directly to contemporary policy debates since, although HE statistics are now abundant (see: Higher Education Statistics Agency and the Society for Research in Higher Education), they tend to focus on administrative needs—such as completion rate, student background, and measures of employability and publication—and fail to capture the impact on knowledge production and transfer.

This last aspect of the research positions the research with respect to economic geography where the contexts within which 'knowledge flows' occur remains hotly debated: although it is generally presumed in economic models that proximity is important (see review in McCann 2007) because it supports the creation of 'tacit' understanding (Polanyi 1966, developed in Storper and Venables 2004 and elsewhere), others have noted that alternative forms of proximity exist, including institutional, professional/disciplinary, and temporary (*e.g.* Boschma 2005, Torre and Rallet 2005, Torre 2008, Massard and Mehier 2008). However, all such work has typically taken as its object of study a single industry and, generally, a single locale such as the Marche region of Italy (*e.g.* Capello and Faggian 2005; Quatraro 2009); what is unusual about the EThOS data is that it enables us to interrogate—using the same data format and approach—such widely varying disciplines as history and particle physics, and to better-understand how such transfers work on *their own* disciplinary terms.

## METHODOLOGY

The British Library has a record of approximately 90% of recently published theses from more than 137 participating institutions across the UK (British Library 2016a), but the data set is continuously expanding and new attributes are being added. In principle, each EThOS record contains at least 14 fields: Author, Title, Supervisor, Awarding Body, Institution, Date, Qualification, Abstract, Keywords, Dewey Decimal Classification, Embargo Date, Restrictions, Institutional Repository Link, EThOS Link. Recent efforts by the Library have seen the metadata enhanced to include details such as grant numbers and sponsors (British Library 2016b), although this will not be available for earlier dissertations. More seriously, some of the critical fields for the proposed research are, at best, patchily completed: roughly 6% of records contain supervisor details, 50% have Abstracts, and about 55% include Keywords. A second, richer but less extensive source overlaps with EThOS and could be employed for filling in gaps and some types of validation, and more than 150,000 dissertations are available in full electronically.

Filling in this number of missing attributes cannot be done by hand, but modern text-mining approaches should help us to complete this process in an automated manner and to extract additional value from the data (*e.g.* Finlay *et al.*, 2012). The BL has scanned and converted to text using OCR technology the 'front matter' of more than 150,000 dissertations and estimates that 60% explicitly acknowledge the supervisors; these scans will serve as both training and testing data sets for machine learning algorithms, allowing the researcher to 'teach' them to extract missing information. The researcher would then take the 'enhanced' dataset resulting

from the above work and produce a labelled property graph in which theses, authors and institutions represent nodes, and the supervisory relationships between authors and theses, and the institutional affiliations between authors, institutions and theses represent edges.

Complex network models can be easily stored and queried using graph databases to provide enormous flexibility and potential leverage when working with highly interconnected data (Robinson, Webber, and Eifrem 2015). Converting the data to a graph representation allows us to draw on an extensive literature on information dissemination in networks (*e.g*. Egghe and Rousseau 1991, Newman 2004), and several measures have been developed to use such links as a proxy for community affiliation: popular approaches include viral infection (*e.g.* Cheng *et al*., *Pei et al*., 2015) and genetic transmission (Myers *et al*., 2012). Other relevant recent work includes: Dietz *et al*. (2007), which deals with the impact and influence of papers on each other, and Sugimoto *et al*. (2011), which employed keyword data and analysis of thesis titles to identify the sub-disciplines to which theses belonged, and to identify the emergence of interdisciplinary research based on theses spanning more than one acknowledged discipline. Gargiulo *et al*. (2016) develop these ideas further by incorporating aspects of abstracts and supervisory relationships.

The student will thus also draw upon work on community detection in networks (*e.g.* Fortunato 2010; Blondel *et al.* 2010). The evidence from patent data—heavily used by economic geographers as the only available proxy for 'innovation' (*e.g.* Sonn 2007)—is that authors 'subscribe' to communities with high levels of 'local' citation (Henderson *et al*., 1993). Comparing known community memberships (academic affiliations, disciplinary affiliations, etc.) to communities extracted from abstracts and keywords should provide new insights into the overlapping social and intellectual currents affecting knowledge flows between organisation (see, for example, the multi-dimensional concept of proximity elaborated by Boschma 2005). A quantitative approach to meso- and macro-scale understanding has been substantially explored by, amongst others, Börner (Light *et al.* 2014, Skupin *et al*. 2013, Börner 2012, etc.) and we would draw on her insights for this project.

**TIMESCALES**

Since the format of the front matter required to fill in missing attributes in the EThOS data is highly structured, depending on the starting competencies of the researcher it should be possible to complete this enrichment process within 7 months of the beginning of the project. Since the enrichment process is expected to continue beyond the life of the studentship and would therefore, ultimately, revert to the BL, the project will prioritise open source tools and text-mining libraries to avoid recurring licensing costs. In parallel with this extraction effort, the student will also be familiarising themselves neo4j, the open source graph database which will be the primary data store and analytical tool for the subsequent analysis of the data set. The Department of Geography already possesses high-performance desktop systems with 32GB of RAM that should support rapid exploration of the data, and one of the principal supervisor's other doctoral students (currently in their 2nd year) has already tested neo4j against an un-enhanced EThOS data set. Once an appropriate representation is agreed, the process of loading the data into the graph database takes hours.

Allowing for 2 months to conduct additional data linkage and validation, the primary research on knowledge transfer mapping at the disciplinary and institutional levels is expected to begin approximately 9 months after the start of the project. Since the approach is algorithmic it should be possible to continuously incorporate new data on a rolling basis and although it is, in principle, possible to examine all disciplines simultaneously we propose that the project focus on 2–3 different domains in order to better understand the differences between them. This has the benefit of narrowing the research in a way that helps the student towards completion while enabling them to draw upon specific relevant literatures. The exact choice of domains will be left to the student, but interesting areas include: Geography, since it stands at the intersection between very different traditions; Artificial Intelligence, since there is anecdotal evidence of a 'brain drain' of researchers into lucrative corporate posts but no model of what impact on academia this 'loss' might have; and history, since this project is fundamentally historical too.

**DISSEMINATION & KNOWLEDGE EXCHANGE**

The project has particular salience to our understanding of the shifting dynamics of HE in the context of 'Brexit' and the high demand for researchers in particular domains. The A.I. 'drain' implies, on one level, that funding for such research by U.K. research councils and universities has had a significant impact on the discipline and might therefore serve to validate particular approaches to disciplinary development; however, the destination of those leavers (Silicon Valley, for the most part) and the extent of the drain (the majority of senior staff and many early career researchers) may well ultimately undermine the ability of the U.K. to continue leading in this highly competitive field. There is also a clear opportunity to challenge conventional narratives of 'impact' measured solely through citation counts or impact factors by tracing disciplinary development across generations of scholars. By prioritising the use of open data and open methods, the proposed research will continue to generate impact via ongoing dissemination and adoption by researchers.

# Bibliography

British Library (2016a). About EThOS. url: http://ethostoolkit.cranfield.ac.uk/tiki-index.php?page=About%5C+EThOS (visited on 10/26/2016).

— (2016b). The EThOS UKETD DC application profile. url: http://ethostoolkit.cranfield.ac.uk/tiki-index.php?page=the%20ethos%20uketd%5C_dc% 20application%20profile (visited on 10/26/2016).

Boschma, R. (2005), 'Proximity and innovation: a critical assessment', *Regional Studies*, 39(1):61–74.

Capello, R. and Faggian, A. (2005), 'Collective learning and relational capital in local innovation processes', *Regional Studies*, 39(1):75–87.

Cheng, J. and Adamic, L. and Dow, P.A. and Kleinberg, J.M. and Leskovec, J. (2014), 'Can cascades be predicted?', *Proceedings of the 23rd international conference on World wide web*, ACM, pp.925–936.

Dietz, L. and Bickel, S, and Scheffer, T. (2007), 'Unsupervised prediction of citation influences', *Proceedings of the 24th international conference on Machine learning*, ACM, pp.233–240.

Egghe, L. and Rousseau, R. (1990) *Introduction to infometrics: Quantitative methods in library, documentation and information science*.

Finlay, S.C. and Sugimoto, C.R. and Li, D. and Russell T.G. (2012). "LIS Dissertation Titles and Abstracts (1930–2009): Where Have All the Librar* Gone?" *The Library Quarterly: Information, Community, Policy* 82(1):29–46.

Fortunato, S. (2010), 'Community detection in graphs', *Physics Reports*, 486(3):75–174.

Gargiulo, Floriana et al. (2016). "The classical origin of modern mathematics", arXiv preprint: arXiv:1603.06371.

Henderson, R. and Jaffe, A.B., and Trajtenberg, M. (1993), 'Geographic localization of knowledge spillovers as evidenced by patent citations', *The Quarterly Journal of Economics*, 108(3):577–598.

Kelley, E.A. and Sussman, R.W. (2007). "An academic genealogy on the history of American field primatologists". In: American journal of physical anthropology 132.3, pp. 406–425.

Lazer, D. and Pentland, A. and Adamic, L. and Aral, S. and Barabàsi, A.L. and Brewer, D. and Christakis, N. and Contractor, N. and Fowler, J. and Gutmann, M. and Jebara, T. and King, G. and Macy, M. and Roy, D. and Van Alstyne, M. (2009), 'Life in the Network: the coming age of Computational Social Science', *Science*, 323(5915):721–723.

Light, R, and Polley, T. and Börner, K. (2014), 'Open Data and Open Code for Big Science of Science Study', *Scientometrics*.

Massard, N. and Mehier, C. (2008), 'Proximity and Innovation through an 'Accessibility to Knowledge' Lens', *Regional Studies,* 43(1):77–88.

McCann, P. (2007), 'Sketching Out of Model of Innovation, Face-to-face Interaction and Economic Geography', *Spatial Economic Analysis*, 2(2):117–134.

Myers, S.A. and Zhu, C. and Leskovec, J. (2012), 'Information diffusion and external influence in networks', *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.33–41.

Newman, M.E.J. (2004) "Coauthorship networks and patterns of scientific collaboration" *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5200–5205.

Polanyi, M. (1966 [2009]), *The tacit dimension*, U. Chicago Press, ISBN13: 978-0226-67298-4.

Quatraro, F. (2009), 'Diffusion of regional innovation capabilities: evidence from Italian patent data', *Regional Studies* 43(1):1333–1348.

Robinson, I. and Webber, J. and Eifrem, E. (2015). Graph Databases: New Opportunities for Connected Data," O'Reilly Media, Inc.".

Sen Pei, S. and Muchnik, L. and Tang, S. and Zheng, Z. and Hernán A Makse, H.A. (2015), 'Exploring the complex pattern of information spreading in online blog communities', *PLoS ONE*, 10(5):e0126894.

Simpson, R. (1983). 'How the PhD Came to Britain: A Century of Struggle for Postgraduate Education', *Research into Higher Education Monographs*, Society for Research into Higher Education. isbn: 9780900868955.

Skupin, A. and Biberstine, J.R. and Börner, K. (2013), 'Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach', *PLoS ONE*, 8(3): e58779.

Sonn, J.W. and Storper, M. (2007), 'The increasing importance of geographical proximity in knowledge product: an analysis of U.S. patent citations, 1975–1997', *Environment & Planning A*, 40(5):1020–1039.

Simpson, R. (2009). The development of the PhD degree in Britain, 1917- 1959 and since: An evolutionary and statistical history in higher education. Edwin Mellen Press Lampeter.

Storper, M. and Venables, A.J. (2004), 'Buzz: face-to-face contact and the urban economy', *Journal of Economic Geography*, 4(4):351–370.

Sugimoto, C.R. and Ni, C. and Russell, T.G. and Bychowski, B. (2011), 'Academic genealogy as an indicator of interdisciplinarity: An examination of dissertation networks in library and information science', *Journal of the American Society for Information Science and Technology*, 62(9):1808–1828.
Torre, A. and Rallet, A. (2005), 'Proximity and Localization', *Regional Studies*, 39(1):47–59.
Torre, A. (2008), 'On the Role Played by Temporary Geographical Proximity in Knowledge Transmission', *Regional Studies*, 42(6):869–889.