

An (increasingly) visible college: Mapping and strengthening research and innovation networks with open data

Juan Mateos-Garcia*, Konstantinos Stathoulopoulos and Sahra Bashir Mohamed

Nesta, 58 Victoria Embankment, London
Juan.mateos-garcia@nesta.org.uk

17 August 2017

Abstract

Innovation policymakers need timely, detailed data about scientific research trends and networks to monitor their evolution and put in place suitable strategies to support them. We have analysed the Gateway to Research, an open dataset about research funding and university-industry collaborations in the UK in a project to map innovation in Wales. We use supervised learning and Natural Language Processing to improve data coverage and measure activity in research topics, build a recommendation engine to identify new opportunities for collaboration in the Welsh innovation system, and present the results through interactive visualisations. Our results suggest that Wales is becoming more competitive in areas identified as strategic targets by Welsh Government, that its research ecosystem is geographically diversified, and that research collaborations tend to take place between organisations that are geographically close. The data sources and methods we have used in the project can help understand this system better, and support it more effectively.

Keywords – Innovation systems; research and innovation policy; topic modelling; open science

1 Introduction

1.1 The policy problem

The economics literature has long emphasised the importance of basic research as an input into innovation processes that enhance productivity and economic growth, providing an important rationale for public investment in science (Griliches, 1991). Unfortunately there is no guarantee that these investments will produce economic impacts. Other things are needed too. Public investments in

science have to be complemented with incentives that encourage researchers to consider the practical application of their work. Networks matter too: Strong collaboration networks between academic researchers and private and public-sector organisation spread information about opportunities and needs, enhancing the practical and policy relevance of publicly funded research, and its potential for impact (Gustaffson and Autio, 2011).

This ‘systems failure’ public rationale for science and innovation policy has informed many interventions and programmes to encourage commercialisation of research, and more interactivity between academia and industry, including through mission-oriented research (Schot and Steinmueller, 2016). The challenge for policy researchers and analysts is to measure these activities and their impacts to inform policy. Traditional datasets used to analyse scientific activity, such as publications, citations and patents are limited for this because they focus on the academic dimensions of research activity and cover only the minority of innovation activities resulting on patents. Measuring collaboration networks and their evolution through innovation surveys is unwieldy, and their high level of aggregation renders them less useful for policy targeting and evaluation (Bakhshi and Mateos-Garcia, 2016).

1.2 The data opportunity

A recent wave of open datasets with information about publicly funded research projects promises to address existing gaps in our understanding of the research and innovation ecosystem, and inform better science and innovation policy (Fealing, 2011).¹

¹ For more information on these datasets, see <https://data.europa.eu/euodp/en/data/dataset/cordisH2020projects> (Cordis), <https://www.starmetrics.nih.gov/> (STARMETRICS) and <http://gtr.rcuk.ac.uk/> (Gateway to Research).

These open datasets include CORDIS (with information about European Commission funded R&D programmes), STARMETRICS (with NSF funding data) and Gateway to Research (focused on UK research funding). There are several reasons why they can help overcome the issues we mentioned above: they contain information about all participants in publicly funded research collaborations, including academic institutions and partners in the private, public and NGO sector; they capture research collaborations regardless of the type of knowledge outputs being produced, providing a more comprehensive view of university-industry engagement in different disciplines and sectors; they are more timely than other data sources such as publications or patents, and contain detailed information about research organisations and businesses involved in collaborations, which opens up new opportunities for data merging and enrichment, and for policy targeting at a high level of detail.

1.3 About this paper

In this paper, we illustrate some of these opportunities through the findings of our exploration of one of these datasets, the Gateway to Research (GtR), during a collaboration with Welsh Government to develop a data platform about Wales' economy and innovation system.² More specifically, our analysis of GtR sought to:

1. Map the landscape of research activity in Wales by discipline and geographically in order to help policymakers understand what are the areas of research strength for Wales, and how this links with strategic policy priorities.
2. Map research networks and identify 'gaps' and opportunities for collaboration that might be addressed through targeted interventions, or by improved networking strategies by their actors in the Welsh innovation system.

Section 2 describes data collection and processing, Section 3 describes our outputs, and section 4 concludes.

2 Data

2.1 Data collection

The Gateway to Research data are available through an open Application Programming Interface (API) with various endpoints for *projects*, *organisations*, *funds*, *people* and *project outcomes*. In May 2017, we downloaded the first three datasets, including information about the research projects that had been funded and their topics (in the *project*

dataset), the organisations that had participated in projects and their location (in the *organisation* dataset) and the funding awarded to projects (in the *funder* dataset).

2.1 Data processing

We started with a dataset of 72,592 projects. One of our main interests was to monitor levels of activity in different research disciplines in Wales. This would allow us to map Wales' research specialisations against the sectors identified in Welsh Government's Science strategy, and to identify the research capabilities in different locations and organisations.

This led us to exclude from the analysis those projects that did not have any research subject information or abstract, such as Studentships, Knowledge Transfer Partnerships or projects supported by Innovate UK. This left us with 33,373 projects (90% of which are research grants).

2.1.1 Classifying projects into high level disciplines

Classifying projects into research areas and research topics was not a simple task. Initially, we focused on the tags (e.g. 'microeconomics', 'robotics', 'materials') assigned to projects by funders, using them to draw a network of research activity where the tags that tended to appear in the same projects were linked to each other, see figure 1, and we then used community detection algorithms to identify tightly knit 'tag communities' in that network (Blondel et al, 2008).

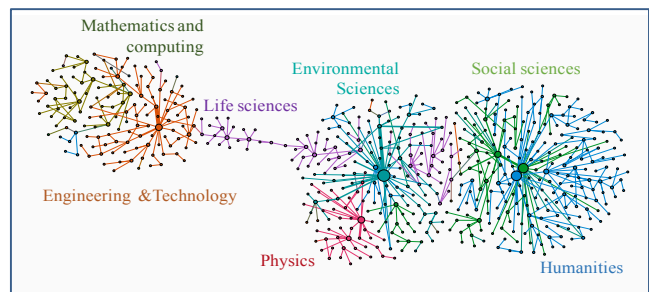


Figure 1: Research tag network. Source: Gateway to Research (2017). Here we visualise a maximum spanning tree of the network to maintain visibility. The community detection was performed on the full network.

This analysis identified 7 quite intuitive research areas (which we labelled as Arts and Humanities, Engineering and Technology, Environmental Sciences, Life Sciences, Mathematics and Computing, Physics, and Social Sciences) that mapped well against the research funding councils (AHRC, EPSRC - primarily funding projects in both

² For more information about the project, see here: <http://www.nesta.org.uk/project/arloesiadur-innovation-dashboard-wales>

Engineering and Technology and Mathematics and Computing - NERC, BBSRC, STFC and ESRC). We classified projects in the research area where it had more tags.³

Almost all (99.8%) projects in the data had a start date of 2006 or later, consistent with the idea that GtR primarily covers research funded in the last decade. We also found that the research tags we had been relying on to classify projects into communities are adopted inconsistently over time and research fields: for example, the Biotechnology and Biological Sciences Research Council (BBSRC) only started tagging its projects in 2011 (note the bump in Life Sciences activity after 2011 in top panel in Figure 2). Meanwhile, the Medical Sciences Research Council does not use research tags (relying instead on ‘health categories’ to classify its projects). In total, 5,962 projects funded by the MSRC lacked tags, and the same was true for 4,040 projects funded by the BBSRC. As many as 1,046 EPSRC project were untagged too.

This bias in the data limited our ability to monitor research trends. We decided to address it by training a supervised machine learning model on a dataset of (generally more recent) projects with discipline labels, using the text in their abstracts and the identity of their funders as predictors. We then used this model to predict the disciplines for (generally older), unlabelled projects.⁴

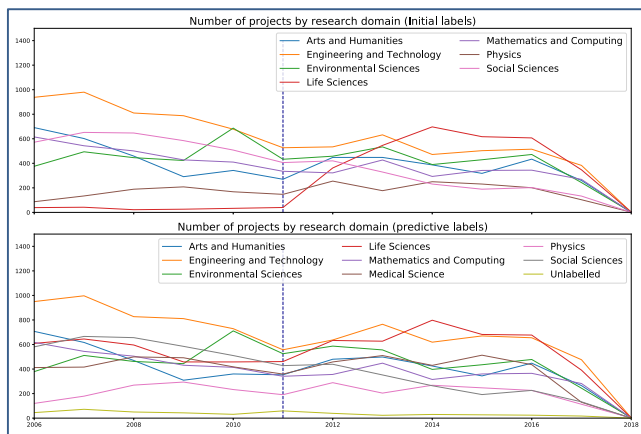


Figure 2: Number of projects by domain (before and after predictive analysis). Source: Gateway to Research (2017)

We classified projects into the discipline with the highest estimated probability, except in those cases where this probability was below 0.3 (we kept those unlabelled). By the

³ If there was a draw at the top, we allocated the project to one of its top areas randomly.

⁴ MSRC funded projects were given a new research discipline label, ‘Medical Science’. We trained logistic and random forests models on the data using a multi-label ‘One versus Rest’ classification framework and three-fold cross-validation. Random forests performed best in the analysis.

end of this process we had reduced our list of unlabelled projects from 6,721 to 565 (see panel 2 in Figure 2), resulting in a final, labelled dataset of 45,491 projects.

This analysis provided, for each project in the corpus, a vector indicating the probability that it belonged to each of our 8 research disciplines. Figure 3 represents, for the projects classified in one discipline (vertical axis), the average probabilities (weights) of other disciplines (horizontal axis). It shows stronger overlaps between technology focused disciplines like Engineering and Technology, Mathematics and Computing, and Life Sciences on the one hand, and the Arts and Humanities and Social Sciences on the other.⁵ We use this by-product of our supervised learning further down the line.

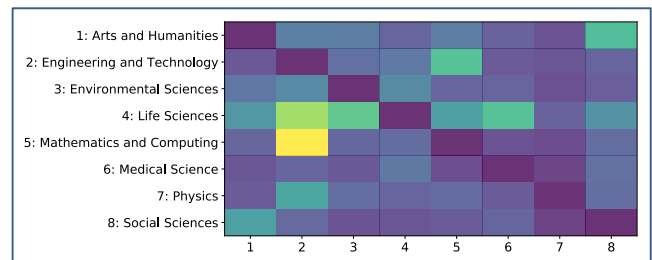


Figure 3: Discipline overlap in the predictive model

2.2 Data analysis

2.1.1 Topic modelling

Science and Innovation policymakers want to measure levels of activity in detailed research topics and domains - as an illustration, Welsh Government has identified specific research areas as ‘strategic’ in previous policy documents (Welsh Government, 2012).

In order to dig below the 8 highly-aggregated research disciplines we had identified so far, we used Latent Dirichlet Allocation (LDA), a topic modelling algorithm which estimates a model for a corpus of documents where ‘topics’ generate clusters of related (co-occurring) words in that corpus with different probabilities. This results in words probability distributions over topics, and topic probability distributions over documents (see Blei (2003) for a canonical reference, and Yau et al (2014) and Sugimoto et al (2011) for recent applications of LDA to the analysis of scientific corpora).

⁵ Note that we converted the diagonal into zeroes because otherwise they dominate the heatmap. On average (and unsurprisingly), the probability estimated for projects actually classified in a discipline was 0.83.

Our initial topic modelling of the research project corpus generated very noisy results. A visual inspection suggested that the algorithm was failing because of the heterogeneity of languages being used in different research disciplines. To address this, we split our corpus into sub-corpora by research disciplines, and trained a LDA model inside each of them, with much easier to interpret results. We extracted 200 topics for each discipline.

We then predicted the topic distribution for each project. Acknowledging the possibility that a project might contain topics from several disciplines (as figure 3 for example suggests) we fit models trained on all disciplines in all projects but we weighted the probability of a discipline’s topic in a project by the probability that the project was in that discipline in the first place (based on the supervised models we trained during pre-processing).

This gave us, for each project, a vector with around 1,600 values representing its weights in 200 topics for 8 disciplines. Although this data had high resolution - just as an example, it included topics such as “bee, colony, pollinator, landscape, crop, specie, honeybee, bumblebee”, “theory, string, quantum particle, physic, black hole, gravity”, “graphene, plastic, flexible sheet, tube, printed, substrate, layer” or “manufacturing process, fabrication, printing, additive, technique, precision, material”, which capture highly specific research topics of potential interest to policymakers, the sheer number of topics made them hard to report. We were also concerned about potential noise in the data for smaller topics. We addressed this by aggregating these research topics into a smaller number of research domain using, once again, community detection inside a topic network where edge weights were based on the jaccard distance (size of the overlap of topics above a weight of 0.01). This resulted in a final set of 88 research topics that we labelled by hand, and used in the rest of the analysis.

3 Outputs

The primary outputs of Arloesiadur are a collection of interactive data visualisations and open datasets that policy users and other stakeholders can explore to understand different innovation trends, geographies and networks in Wales in a way that makes for better informed policy. We have worked closely with an external data visualisation agency to produce three visualisations of the Gateway to Research data whose processing and analysis we have described so far, including:

- A visualisation of research trends at the national and local level (see screenshot in figure 4)
- A visualisation of local specialism (see screenshot in figure 5)

- A visualisation of research collaboration networks and new opportunities for collaboration (see screenshot in figure 6).

In the rest of the section, we describe additional analyses to generate these visualisations, and provide illustrative policy-relevant findings.

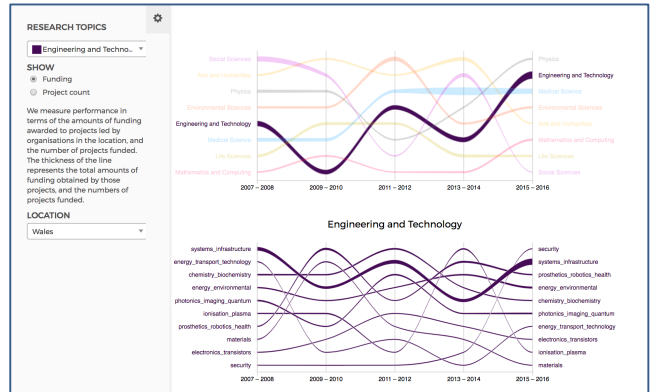


Figure 5: Screenshot of interactive research trends visualisation



Figure 4: Screenshot of interactive research geography visualisation

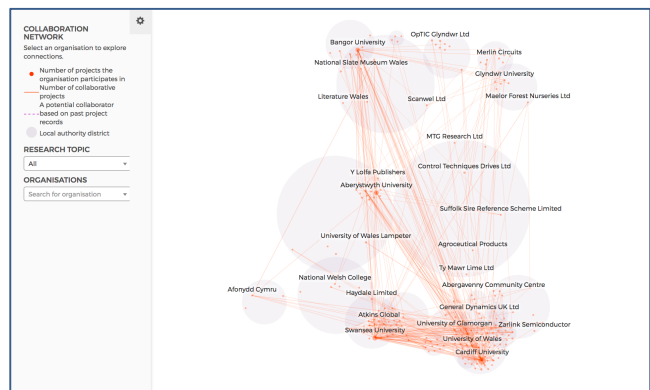


Figure 6: Screenshot of interactive research collaboration visualisation

2.2.1 Reporting local trends

All our visualisations required classifying projects into locations and into research topics. To do the first, we geocoded all organisations in the Gateway to Research data using Google's Places API, and classified them into nations (e.g. Wales) and its principal areas (administrative boundaries roughly capturing local economies).⁶

When it came to classifying projects into research areas, we used a different approach for each of the visualisations.

In the case of the trends analysis (figure 4), we classified each project in the research area with the biggest probability (weight), and calculated the number of funded projects and total amounts of funding raised by projects led by organisations in each research area and location. This eliminated the risk of double counting in projects or funding. We controlled for wider changes in research trends by calculating revealed comparative advantage indices that considered the relative specialisation of a location in a research area compared to UK averages.

In terms of findings, our analysis of research trends suggests that Wales has become more competitive in its ability to attract funding in Engineering and Technology, Medical Sciences and Mathematics and Computing (although starting from a low base in the last area). Physics has also seen relative growth in funding. By contrast, Arts and Humanities and (especially) Social Sciences have declined in the most recent period. Interestingly, there is strong overlap between high performing areas and research disciplines that were identified by Welsh Government as 'grand challenge areas' in its Science Strategy (these were "Life sciences and health", "Low carbon, energy and environment", "Advanced engineering and materials").

Our analysis also allows us to drill down into more detailed research areas and combinations of research projects, and track interesting trends. For example, the data suggests that Wales has growing strengths in research areas related to the "data revolution" such as robotics and cybernetics, prosthetics, robotics and health, bioinformatics, statistics and data analysis, and security. In 2015 and 2016, projects led by Welsh organisations in these research topics were awarded almost £5m by UK Research Councils. Some examples we find in the data include uses of deep learning for cell imaging led by Swansea University, development of robots that learn through play in Aberystwyth University, or a network to enhance big data analyses for plant research in Cardiff University.

2.2.1 Reporting local specialisms

In the case of the local specialisation analysis (figure 5), we were more interested in measuring the 'research capabilities' in a location, so we classified projects on their three most important projects, and counted any funded projects with participation from organisations in the location (regardless of whether they had led the project or not). This means that there will be double counting in the number of projects and levels of funding of obtained because projects can contain more than one topic, and involve more than one organisation.

The findings of our analysis suggest that the Welsh research ecosystem is geographically diversified, with research capabilities present in different locations. For example, Medical Sciences, Social Sciences and Arts and Humanities are more important in Cardiff, an emerging UK 'creative cluster' (Mateos-Garcia and Bakhshi, 2016), while Engineering and Technology are more important in Swansea, where we also see strong activity in Mathematics and Computing. Meanwhile, Ceredigion and Gwynedd are highly competitive in Environmental Sciences.

2.2.2 A prototype recommendation engine

An important question for Wales' innovation performance is to what degree are these geographically dispersed research capabilities combined in collaborative projects – we explored these questions in our third visualisation

First, we created a network showing previous research collaborations between Wales-based organisations (left graph in Figure 7). Organisations that have previously collaborated in research projects (as captured in the Gateway to Research data) are connected in the network. We have arranged the nodes based on their geographical location.

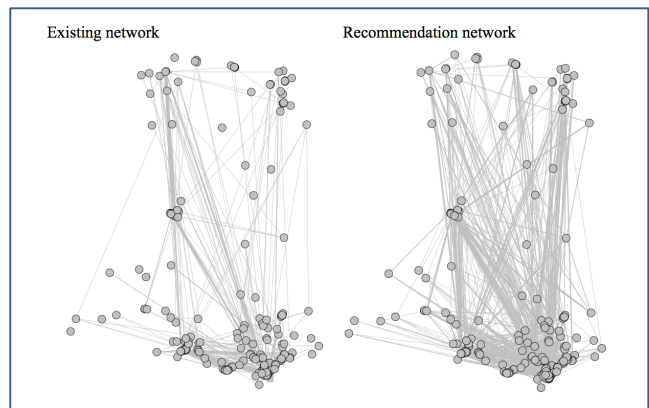


Figure 7: Existing and potential research collaboration networks between Welsh organisations

⁶ We used the shapefiles available from the Office for National Statistics Open Geography Portal: <http://geoportals.statistics.gov.uk/>

Although the analysis of existing collaboration networks indicates that there is connectivity between different research institutions and tech communities in Wales, with 283 research organisations in Wales involved in projects with other Welsh organisations in the last 3 years (45% more than in the 3 years before), organisations still tend to look for collaborators close-by: a third of the research collaborations we identified were inside the same principal area.

In order to identify new opportunities for collaboration, we built a recommendation engine that would identify potential collaborators for an organisation by looking for the strongest collaborators with the 10 organisations most similar to it.⁷ We sought to ensure recommendation relevance by filtering from the recommendation set organisations that had no overlap in the research areas where they work. We used all this information to create an alternative ‘recommended network (right panel, figure 7).

This second network is denser than the existing one, and also displays a much higher propensity for collaboration *between* principal areas as well as inside them. In fact, when we estimate the assortativity coefficient for this second network based on principal district areas (that is, the propensity for organisations to be connected with those in their same location), we find that it is only 1% of the value of the assortativity coefficient in the ‘real’ network. This suggests that there are substantial opportunities for research collaboration between organisations based in different parts of Wales. One important goal for our visualisation is to, increase the visibility of research activities already taking place in Wales, potentially encouraging even more and better collaborations in the future.

4 Conclusions and next steps

The analysis we have presented in this paper supports the idea that open datasets about research can be harnessed to generate relevant information for innovation policymakers, and create tools and resources that empower actors in the innovation system to make better decisions. We will be able to confirm this hypothesis when the platform goes live in Autumn 2017.

We have three main next steps: first, to further fine tune and adapt our NLP analysis. We are particularly interested in adopting modelling frameworks that combine LDA with word embedding that capture word semantics and generate more intuitive topics. Second, organisation data in Gateway to Research is only available at a high level of aggregation

(i.e. ‘University of Cardiff’ instead of specific departments of teams), which constrains our ability to make targeted recommendations. We will explore options to address this drawing on external data sources like ORCID, Google Scholar or Microsoft’s Academic Knowledge API.

Finally, we want to build on our current descriptive (if policy relevant) analyses to start modelling the complex dynamics of industrial clustering and collaboration and how they are shaped by variation in the types of knowledge produced in different industries, helping us to move from better measurements of scientific communities and fields to a stronger understanding of their emergence and their evolution. Open datasets like Gateway to Research will enable substantial strides in that direction in years to come.

Acknowledgements

James Gardiner participated in the early stages of data collection, and Cathy Atkinson, Katharine Breeze and Steve Dempsey provided helpful feedback. Arloesiadur is supported by Welsh Government.

References

- Bakhshi, H., and Mateos-Garcia, J. (2016). *New Data for Innovation Policy*. Paper presented at the OECD Blue Sky Conference, Ghent September 2016.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.
- Fealing, K. (2011). *The science of science policy: a handbook*. Stanford University Press.
- Griliches, Z. (1991). *The search for R&D spillovers* (No. w3768). National Bureau of Economic Research.
- Gustafsson, R., & Autio, E. (2011). A failure trichotomy in knowledge exploration and exploitation. *Research Policy*, 40(6), 819-831.
- Mateos-Garcia, J., & Bakhshi, H. (2016). The Geography of Creativity in the UK: Creative Cluser, Creative People and Creative Networks.
- Schot, J., & Steinmueller, E. (2016). Framing Innovation Policy for Transformative Change: Innovation Policy 3.0. *SPRU Science Policy Research Unit, University of Sussex: Brighton, UK*.
- Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations

⁷ We measured similarities through the cosine distance between organisations’ research profiles based on the research topics of the projects they participate on.

using latent Dirichlet allocation. *Journal of the Association for Information Science and Technology*, 62(1), 185-204.

Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, 100(3), 767-786.

Welsh Government (2012). *Science for Wales*. Cardiff: Welsh Government.