OST des Sciences et des Techniques
Observatoire
Sciences
Techniques

1er Séminaire SciScI - OST
Mardi 11 octobre 2011
Julia Lane (NSF), USA

"Advancing the Science of Science and Innovation Policy: Lessons learned from the US"

TRANSCRIPTION

I think the starting point was the STAR METRICS (Science and Technology for America’s Reinvestment: Measuring the Effect of Research on Innovation, Competitiveness and Science) initiative. The pilot was done at the National Science Foundation, but the project is led by the White House Office of Science and Technology Policy (OSTP).

We developed a data infrastructure that helped us understand, or at least describe, what science investments have been made and the results of those investments. Then Stefano Bertuzzi, who is my colleague at NIH, and the co-lead of the Star Metrics, visited Inca last year and that is when we really started the pilot project.

What I’m going to describe now is the feasibility study, not so much the interpretation of the results. As with everything, a lot of people did the work; I’m just the one who talks about it. These are all the people who did an enormous amount of work on the feasibility study.

Now I’m going to talk about what the goals of the project were, what the approach, the key results and the next steps are. The goal was to document the investments; this is very similar to the projects that we’ve had in the US for basic research agencies as well as applied research agencies, and we had also been doing similar work in Australia, and are starting a project in Norway. What we wanted to do was to say: What can we do with the existing data? What are the gaps, and the next steps? That is what you would expect from a feasibility report.

As Fabien pointed out, the focus on using bibliometrics has its place, but if you are trying to describe the results of science investments, it really is not a very satisfactory approach, for a whole number of reasons. If you’re trying to describe the results of investment, documents don’t do anything. Who does science? Scientists. So what you absolutely have to capture is the activities of scientists. That is at the core of what describing science is about. So the fundamental approach that we have taken, in the Star Metrics program and in other countries, and I think the reason that there has been so much attention paid to the Star Metrics, is that it stands back and says: “If we want to describe the results of science investments, we need to think about what we are measuring and what we are describing.

Essentially, if you think about the scientific enterprise and the way in which research funding works, that it is people who do science — that’s the behavioural unit of analysis —, what grants, dollars and contracts do is they fit the research agenda of individual scientists, groups and clusters of scientists. That is really what the monetary intervention is intended to change. Then what happens is those scientists produce things that the research agencies are interested in. That might be papers, publications, students, conversations, presentations or conferences and so on, what you're really interested in, in what scientists are doing, is creating, transmitting ideas, and having those ideas adopted. That's the core product of interest.

Somewhere along the way, at least this is what our White House interagency group argued, that fundamental concept got distorted. Somehow it came to be a funding grant, and I want to see products attached to the grants, but that is not what science is about. Science is about research networks, about ideas that might take scientific generations to have an impact. If you just try and look into three or five-year slices, you are fundamentally misunderstanding what science is about.

The challenge then becomes: How do you build an empirical framework that captures this scientific framework, where you look at grants affecting networks of people, operating within institutions — that is having access to genome sequencing machines, or large hadron colliders, or scientific assets — as well as a functioning organizational structure, and look at this longitudinally over time. Now the challenge is that the data typically sits in different places. They sit in funding organizations, in research organizations, they sit all over the place. So you can understand why for twenty, thirty years, we have done things using a 1970's reporting framework, where you ask the researcher to fill out a form, that then goes into the reporting system of a research organization, because it's been quite impossible to build a intellectually coherent system. We now have the capacity to do that, because we are in the 21st century, and no longer in the 1970's.
What we are going to talk about is using modern technology to capture scientific achievement. I'm going to let you think a little bit about what scientists do, and in order to do that, I'm going to show you a little video that is going to take three minutes. Instead of thinking about what we have, what I would like you to think about is what science is about. How many of you know these taped videos?... I want you to think about what funding agencies are trying to do.

(Video playing… not for long)

The point is they are starting a movement, and what happens is, over time, he starts to bring people in. If you think about what a lot of private foundations do is they are trying to indentify the leaders and the followers. Not only leaders, but people who can bring other people in… You can see all the interactions with the scientific community: we have to have a public, we have to have leaders, we have to have followers… The point is: How do we capture that creation of an idea, the way it gets transmitted, and adopted. With this video, I wanted you to get a sense of what science is about. When we talk about impact that is really what we are trying to capture. The ways in which that is done vary, and the ways in which, particularly, young people are communicating, with YouTube videos, Facebook and so on, allows us to capture that information; we know how to do it technically, because the mathematicians have done graph theories, and we now have much more sense on how to do randomized controlled trials, we have new applications… When I was growing up as a social scientist, we looked at individuals, and we looked at firms, now we look at networks, so the social science has caught up with mathematics. There are new ways of communicating knowledge, and it is no longer just publications. There are many, many ways in which scientists communicate.

The database structure has also changed to capture this. Google, Microsoft and Yahoo no longer use relational databases, because they are too structured, they use graph-oriented databases, again building on the mathematics, to capture social networks and the interactions of human beings, and we have new ways of capturing what people are saying through computational linguistics and natural language processing. So the point is that twenty years ago, we didn't have any options, in the way in which we captured the creation, transmission and adoption of knowledge, but now we do, and the point of the White House activity and the activity at the National Science Foundation was how we are going to build a system like this.

One approach that we used was that, when I came to NSF, we had this massive amount of documents that described research that each research team had written in fifteen pages — 200,000 projects, all in PDF form, basically, the 1990's equivalent of paper documents. So the real challenge is: How do you take that information and turn it into something that is useable? We put ten teams of computer scientists together and fifty program officers and directors within NFS, and spent two years figuring how we could analyze this vast text data and make sense of it. We have got the report, if you're interested, but essentially, the consensus we came up with was that using natural language processing and topic modelling was the best way to capture what scientists were doing, using their words. So essentially, the approach here is to take text documents, and then look at the co-occurring words. This is exactly the same approach that Microsoft and Google use, so when you start typing into Google, words are popping up, not because they know what you're thinking, but because they know in your text documents what words co-occur. There are many ways of doing this. We chose one because we had to do so, but there are many others. This happened to be a very good one; it certainly worked for us, and it seems to have worked reasonably well at Inca too.
We are going to use topic modelling to describe the text, and we're going to use that conceptual framework in the HELIOS case, to develop a data schema, which is centred on the person. You've got the funding, you've got the people, and then you've got the results, but the people sit at the centre of those networks of individuals. It's who does the work, and with whom. That's absolutely critical, because that's how ideas get transmitted, and you capture the results of investing in a research stream over time. The "what" is captured through topic modelling, as well as the "where" — because we need to know how people get funded in different parts of the country and where the gaps are, that kind of thing.
The key here is we didn't want to miss a round with creating new structures, so with our experience in the US, and in other countries, there are essentially ten data elements that are needed, and you can pull them from existing systems. This is what the data look like within Inca: the project title, the budget dollar amount, the beginning and end dates, and then what we wanted to have was to get information on the institution and the principal investigator. That is who, where, and then what. This is a description of the research that was being done. We were able to take the full project descriptions, and again the title, and ideally down the road we will be able to use the CV. What was

fantastic about these data is there was also data on the research teams, so it is not just the principal investigators. Again, we are interested in the creation, transmission and adoption of knowledge. A major way in which that happens is because the students get jobs in firms: that is your biggest way in which technology gets transferred from the bench to the private sector, and there is all kinds of case-study research that documents that, and it's why so much regional economic development occurs around universities. Having the research teams was absolutely critical. Now what we have is the core of our data schema, because for one of the first times we have the people and data on the very start of a project, and the project teams that create ideas.

Then the next step was, could we link that to their activities? We built on work that had already been done at OST. The question that started out was: Can you take data from Inca and repurpose it to answer the types of questions we wanted to be answered? And the answer is yes. You can take information on individuals, and research teams, and thus describe collaborations at the beginning of a project. For example — this is the kind of thing that was in the report — we could identify unique researchers, and the unique institutions. That is the framework in which we are operating. The second question is: Can we describe what research is being done, and can we put it in an international context? Now, it turns out in cancer research people spend a lot of time identifying what cancer there is, and what kind of analysis is being done. But can we go into a more fundamental label that doesn't require people to write down what's being done? Can we do this by taking the text information? The answer, again, is yes. We can use the general topic areas of research. Let me give you an example. Here is a list of the topic areas that are being funded by Inca research. You could see the string of words that describes the topic areas, and this is ranked in order of the most commonly occurring topics in the documents. What we are able to do is to put that in a context of NIH research. So because we used the NIH topic-modelling algorithm on these documents, we were able to say: Here is where our funding sits within the NIH corpus.

Here, for example, is a study that is funded by Inca, and that is the abstract that describes that project. When we look at that project, and do the topic modelling, it says these are the topics that best describe the research that is being done in that project. You see that it is a bag-of-words approach, strings of words that you want to put in that context. Now what we can do is put it into the context… It is going to pull up the topic areas associated with that topic at NIH. You are now looking at how that proposal fits in to NIH research. There are the topic words at the top, there are the phrases using the words that the scientists use in their documentation, these are the tags that are associated with it, and then here are the grants at NIH that are most closely related to that research. Here are also the main institutes; not surprisingly, because of Inca, NCI (National Cancer Institute) is the major funder, you can see the major NIH grant mechanisms, and then the review panels to which that topic has gone. That provides lots of information in a fairly textual manner, but you can also show it visually. What you can see, if you wanted to drill down into a particular area, how that fits in, and these are the topic areas that are the closest, and then, clearly, this is the NCI grants… This is putting them in a non-textual framework, in a visual framework. Not a single principal investigator had to fill out a single form to do this; this is automatically generated. Now, it may not be perfect, but because it is a Bayesian algorithm, you have the potential to set up feedback loops so that if it is wrong, if a program officer thinks it's wrong, or if the researcher thinks it's wrong, you could connect that in, and re-update.

Then we wanted to look at the bibliometrics and the patents, and so we used the databases that word Inca… We should be careful to say there is no causality here, because of the overlap of times, it is just describing what Inca research is being done, relative to the research being done by Inca-funded researchers. You are then able to show what are the top probabilities; you'll have the top topic areas that Inca grants are in blue, and the red reflects the topic areas of publications in which Inca researchers were involved. I said at the beginning I wasn't going to do any analysis, but just one thing that hits you is, obviously, in some areas, Inca is funding much more than the publication corpus of document over that time period exist, and then in other areas there is a lot of research that is going on, and Inca is not investing that much in that. That is one interpretation that could be drawn.

You can also describe the link between Inca-funded researchers, and this was one thing that Valérie and Fabien were very interested in: How can we describe with whom and in what place the research is being done, so the scientific networks of collaboration. The sense is not just what research are we doing, but also can we say something about the private sector with which Inca researchers are working, both in terms of their initial collaborations, and in terms of subsequent patents and publications. These are the publications that showed up. We are actually more interested in the collaborations represented by those publications, and what you can see is the number of Inca-funded researchers who are cited on that, and many more total researchers, so that gives you a sense of the richness of the Inca-funded collaborations, and then you are able to find out how many of those are associated with private-sector firms, and what are the unique firms.

Then you can show where the Inca-funded research is being done, both in terms of the research institutions and collaborating institutions. These are clickable wireframes; they are all driven from the collaboration data, the patent and the firm data. The pink is an Inca-funded firm, and the blue is a private sector firm that is related to, that is where people have been working with Inca-funded researchers.

Essentially, here, what we want to do is to show the interactions between this Inca-funded research organization and people within the firms. This is fake data for now; this is an example of how this might work. Here are the individuals, here are the patents, here are the publications, here are the awards, and that is the icon for the businesses. If you want to zoom in on a particular area, you can show the networks that are associated with that. Here is the information on the actual awards, and then here are the topic areas in which that award was made, and then you can show who are the firms (blue) who were cooperating on that award. Does that make sense? Then you can show something about the people in those firms, and who they are collaborating with. If you want to go back to a geographic view, you can then zoom into the geography and say: OK, I'm interested in, for example for a particular firm, how they connected with Inca-funded researchers. There again, you pop that up and you can see the links, not just in terms of the abstract networks, but the very detailed information on the individuals. And if I want to drill down into one particular person, and see how they are related, I can pick that up and see how that network has changed.

Now, obviously, a theory of change is very much that we are interested in seeing how funding is going to fix those networks, so developing a sliding bar that can go and show how those networks expand and grow is one of the directions we'd like to go in.

What are the next steps? One of the things that you think about when you are developing a database like this, is you want it to be able to show visually what is going on, so how can you turn information about a theory of change in networks and network collaborations into some measures that describe to yourself and the rest of the world what you are doing? This is a tentative, these are just suggestions, and we would like to hear discussions from the group as well… Within the project label — remember we are starting from the creation of ideas — how many internal and external collaborators are there? Since we are talking about funding a research area, how many Inca-funded researchers are related to other researchers in the same area? Are they dominating, or are they starting to contribute? And so on…

One of the other things that we should be interested in is to take from the business world some sense of measure of the result of your investments, in way that makes sense to everyone else. One of the things we found is Congressmen don't care about counts of papers and publications. What they really want to know about is: Tell me in words that people would understand what the impact of your research is. So one thing that has proven quite effective is to talk about the same way that they do in business, about the process of ideas moving from bench to practice. One way in which businesses do that is they stage-gate things. They have stage 1, stage 2, stage 3, stage 4, and you can write that down. In every discipline, we have an understanding of what the process by which ideas get created, transmitted and adopted is. Then what you ca do is you can say: Mr Congressman, it used to take five years to go from here to here, and with the research funding we have sped that process up to be only four years, and that is a discipline invariable measure, so we sped it up by 20%, or we have increased the probability of the ideas being adopted by 2%. So those are sensible behaviourally-based measures that could be used. Or you could say something like: There weren't many people is this particular research area, let's say fifty researchers, and we have added twenty new researchers to the field. That number has increased over time, so that is the sliding scale, and you have the math to go with it. You can also look at the breakthrough coverage. If I had been successful in showing you the video, one of the things you want to look at is who is joining. As a program director at NSF, every program director can tell you when they think a particular field is vibrant and growing. One of the things they look at is if it is the same old people who come and talk to the same old people all the time. So how do you capture what is going on? Are there new people? And in particular the future is the graduate students and post-docs, because they are the ones who are going to bed their career on a particular field, and if they think it is winning, they will come in. They may not stay in the field, directly in academia, but as I said before, they go out into the business world and they carry those ideas with them; that is the technology transfer that is of fundamental interest, much more than publications or patents. The modal number of publication citations is zero, or one, depending on the field, and 75% of patents never get applied. The people make a big difference. Valerie convinced us of this, the links to research organizations and to businesses. We can measure, we don't just have to count; one of the things that could be interesting to do in the future is to measure the quality of those links, but that's another story. Right now we can just start by thinking about the number of links, and

what industries are involved. Are they industries that the particular government in charge cares about, or not? And then capture the international framework as well.
So those are some ideas as to indicators that could be generated out of this framework that make sense; they make sense in terms of the central framework, they can be built given our existing data infrastructure, and they make sense to legislators. The concern that the legislators certainly have in the US is: We give money to a university or a research organization, and then you never see it again. So what you want to be able to do is to convey the notion that there is an active international and national research agenda.

Conclusion, en français :
Avec ce travail, nous avons démontré la faisabilité, puisque nous avons utilisé nos données telles qu'elles sont construites, les outils tels qu'ils existent, et donc non seulement nous pouvons exploiter les informations — la méthode fonctionne avec nos données — mais nous avons pu intégrer à la fois les outils et les données. Ce que nous voulons maintenant, c'est continuer le développement de cet outil et mener une étude pilote, en définir les possibilités d'extension (investissements d'infrastructure, soutiens aux étudiants…), définir les contours de la phase suivante, discuter avec les parties prenantes, puisqu'il existe d'autres interlocuteurs que l'Inca dans le domaine de la recherche sur le cancer.