SCIENCE, TECHNOLOGY & INNOVATION INDICATORS

Open indicators: innovation, participation and actor-based STI indicators
PARIS 2017

Introduction to the Conference

Philippe Larédo

Dear colleagues,

With the organisers of this conference, we have considered it useful that I introduce to you the developments we have made over the last 4 years around RISIS.
They belong to the series of debates we had had these last years about access to data. The previous debates (in particular in Leiden, Lugano and Valencia) have put forward three main aspects: the need for new indicators, the growing access to new public sources, and the conditions to take advantage of the very different types of sources.
This helps to better locate what we have been doing within RISIS. So let me address them in turn.

The need for new indicators

This may look evident, but what strikes me, is, since the beginning of the 2000s the growing need for public 'evidence' nurturing policy and strategy debates. This has a number of implications that I have gathered under four aspects.
They need to address the issue under discussion (we all know the image of the lamp post), and this is what has been pushing in the last two decades for going far beyond the long established OECD financial approaches and datasets.
To nurture debates, policymakers cannot be satisfied by a picture at a given time, indicators (and that is the deep meaning of the term) focus on trends and transformations, enabling comparisons in time and space.
But these indicators nurture, and this is the third critical aspect for me, only if they are robust. Too often instead of nurturing discussions about the transformations they describe, they nurture a discussion about the representativeness and validity of the indicators themselves.
Barré et al. have added a fourth dimension dealing with strategic behaviours. Most indicators witness a strong asymmetric distribution with a few entities (large firms in particular) dominating the indicator. It is thus a myth to consider all actors in one type to be equivalent, and the same occurs for spaces with metropolitan areas. The explosion of rankings of all types translates this need. It becomes thus critical for policy or strategy to be aware of these situations. This is why Rémi and his colleagues have proposed a new category of indicators, positioning indicators.

This implies that new indicators should be theory-based, that they entail time series, and that they reflect on the way they consider actors and spaces. This has strong consequences for database building, especially when this is no longer delegated to professional bodies, statistical offices. The number of ‘technical sessions’ in this conference is a testimony of the importance of the investment to make.

The growing access to new public sources

This movement has been rendered possible by the explosion of new public sources. We all think of big data and open access, out of which new datasets should be built. I looked to the papers in this conference and it clearly reflects the fact that we are far from such a world. Yes there is a multiplication of public sources, but the greatest share, whatever Governments wish or say, remains private. They are private because they are owned by private companies that need to live on their use; they are private because they have been built by a few research individuals or groups that wish to gain an advantage from this investment. This is why the last conferences have debated on the hybrid model proposed by Paul Wouters. In RISIS we try to implement it, organising access to all researchers for one use only: ‘publishable research’. RISIS presently offers access to 11 datasets, for which there has been an important effort to insure quality and robustness. All these datasets have undergone a shared process for quality control, for documentation (all the metadata is online) and for harmonisation. Harmonisation is crucial as it enables problem-based integration of the different datasets as is shown in quite a few papers presented in this conference.

Let me say few words about these datasets and our views on the future.

When we proposed RISIS to the EC, we told that we would in this first round focus on 6 themes we thought of major importance for knowledge dynamics and policy.
* Innovation dynamics is the first (and this is mirrored by the efforts done in STI these last two years. We have helped to upgrade and develop two datasets linked to large firms and venture capital backed start ups (**CIB** and **VICO**) and have pushed the creation of a specific dataset on fast growing mid size firms in Europe. It has just opened and there is a presentation in the conference doing a first exploration of this new dataset (**Cheetah**). We also are opening access to an enriched Patstat database on patents (IFRIS Patstat). We clearly miss a dataset on social innovation, and we hope to build an encompassing one in a companion EC project that is presented as a poster.
* European integration comes second with two datasets in the process of being integrated: **EUPRO** for European projects and **JOREP** for transborder programmes (and projects). Again we hope to have future extensions covering in Europe, national and regional funding agencies and funding NGOs.
* Public research dynamics is being built along the two core lines: opening access to **Leidenrank**, an enriched version of the WoS, and enlarging and enriching the ETER dataset on universities (now integrated into **ORGREG** covering PROs and university hospitals).
* We are using the work done on the **nanoS&T** dataset as a demonstrator for building new datasets on emerging technologies. A presentation in the GEO session linking publications, patents and EC projects shows how dynamics deploy at the metropolitan level in Europe.
* Our fifth focus is on PhD careers giving access to a national dataset (**Profile**, there is a presentation of its key results in the career session) and to the EC-built **MORE** dataset on mobility in Europe. Clearly we still have important efforts to do to build a European-level dataset on careers, probably through completely new approaches away from surveys or panels.
* Finally we have developed in relation with OECD, the World Bank and JRC, a specific repository of policy evaluation, **SIPER**. This is a very different dataset but long awaited by

international institutions. This is the only dataset that will be freely available to researchers and policymakers alike. It is still in development but already accessible in its prototype form.

Taking advantage of different types of data sources

Clearly this is not the end of it, and RISIS is fully open for integrating other datasets of interest for the field. We might also organise differentiated levels of integration within RISIS. But whatever the effort, many questions and projects will have to consider other datasets outside of RISIS, or require that new complementary databases are built from the fast growing raw data available on the web.

This conference mirrors the importance of such developments, in particular with the sessions on altmetrics. But identifying, selecting and transforming raw and unstructured data into usable datasets is a long and often problematic process.

This is why RISIS has dedicated important resources to help us do this by developing platforms and tools to support the construction, enrichment, treatment and visualisation of new problem-based datasets. Let me underline three specific developments supported by RISIS.

* I have spoken of positioning indicators. Whatever the dataset, we face at the community level important problems dealing with organisational and spatial aspects. We mobilise here the efforts done in the ICT community to develop approaches to automatically match actors, but we consider that it is not enough and we have developed important efforts to build a 'register' of important public and private actors at European level (see the ORGREG register for public actors in the new datasets, and we have a firmreg register in the making by combining the 3 firm datasets). We also are developing two services for geocoding and clustering addresses (so that analyses can be made at the relevant level, in particular metropolitan areas).

* Our colleagues in Amsterdam have made important efforts in creating a new platform – SMS – to help access external datasets, develop, structure and enrich new datasets and organise a RDF based integration of datasets. This is still in development but already accessible through on-site visits.

*And we, in Paris, have developed CORTEXT as a freely available platform for semantic treatment and visualisation of textual dimensions of datasets. It has now been running for 3 years with around 200 different projects per month.

I shall not say much more on these two platforms since we have organised two special sessions this afternoon that will help those interested to know more about them and their use.

Dear colleagues, that was the initial presentation of RISIS I wished to make.

We are at the end of a first round of 4 years and expect to be one of the themes selected for a second round in the next H2020 call for research infrastructures.

Of course we are going to evolve, and this requires that we have a clearer look at the challenges that face us.

This explains why we have asked Ismael to reflect on these, and have high hopes both in his keynote and in the discussion that will follow.