RISIS

Research infrastructure for research and innovation policy studies

SEVENTH FRAMEWORK PROGRAMME

A new data infrastructure for science and innovation studies – RISIS

http://risis.eu

Philippe Larédo

UNIVERSITÉ PARIS-EST

École des Ponts ParisTech

IFRIS

MANCHESTER 1824

The University of Manchester

MIoIR

RISIS in brief

- EC infrastructure programme project
- Ambition: build a data infrastructure complementary to present statistical approaches, supporting the development of "positioning indicators" (Lepori et al., 2008)
- 2 main goals:
  (i) integrate existing fragmented datasets and open on a free-of-cost basis to European researchers (100 projects anticipated) → **http://datasets.risis.eu/**
  (ii) develop new software platforms to support dataset building and treatments (from heterogeneous & unformatted textual corpuses)

RISIS four main features

- Dataset building platform from web sources, providing 'services' (SMS - under test)
- Reference datasets to support enrichment (actor based) & suite of tools (e.g. geolocation and clustering)
- Access to robust focused datasets (9 presently, 12 by end of year) → **http://datasets.risis.eu**
- Semantic treatment platform: on line, already more than 1000 users → www.**cortext**.net/
- And a central access which enables to follow on-going developments: www.risis.eu

What is RISIS for researchers

- The possibility to use one or more datasets to ‘position’ your own topic – access is free of charge, takes place mostly ‘on site’, but costs for visits paid for by RISIS (once project is accepted)
- The possibility to use platforms to clean & enrich your own dataset, and to treat it (especially for semantic treatments via CORTEXT)
- Important documentary effort and use cases on line for learning about the datasets
- Many training courses (free of charge) to know about datasets, learn to use platforms, learn about tools & methods

Why RISIS

- Two central lessons from innovation studies
  - a strong asymmetry in knowledge production & innovation
  - strong agglomeration phenomena
- They highlight the limitations of classical input/output indicators based upon statistics → thus the ambition to offer new resources: a first series of datasets open for research
- New possibilities offered by open data → development of platforms, reference databases & tools

2 lessons & their consequences

- Cope with strong asymmetry in knowledge & innovation dynamics
- e.g. 200 groups represent half of world industrial R&D (IPTS scoreboard) / 200 universities in Europe produce 80% of scientific articles /….
- 4 implications:
  1- keep the identity of actors → central use of public data (whether public or private, free of charge or paying)
  2- choice of the organisation (and not the individual) as the central reference unit
  3- extensive use of textual data
  4- need for a number of ‘reference databases’ or ‘registers’ on actors/organisations (to enable comparison & integration)

2 lessons & their consequences

- Take into account the other face of globalisation, the importance of ‘place’ e.g. ‘glocalisation’
- See work on agglomeration dynamics – cf. example of world production of nanoscience concentrated in 200 urban areas
- 3 implications:
  1- extensive need and tools for geolocalisation at the lowest level of aggregation (e.g. addresses of authors in papers, inventors in patents, participants in projects…)
  2- need for clustering tools enabling to follow effective forms of agglomeration…
  3- but also need for a type of standardisation of ‘urban areas’ / ‘metropolitan areas’ in order to foster comparisons.

A first series of problem-based datasets

- 6 themes focused during this creation process, and for each of them, selecting/building focused datasets.
- Innovation dynamics of firms: 3 datasets
  - globalisation of R&D activities of large firms (CIB, see Laurens at al., 2015 for first unexpected results),
  - long-term dynamics of small high tech firms (VICO dataset)
  - innovation characteristics of fast growing mid-size firms (whatever industry, under construction, first access expected at end of 2016)
- European integration with enlarged EUPRO (a longitudinal dataset of all European projects) and with JOREP (focused on trans-border funding programmes in Europe)

A first series of problem-based datasets

RISIS

- Public research dynamics with the ETER register, and enlarged Leiden ranking
- New ‘dominant sciences’: A demonstration dataset to develop tools for characterising emerging technologies, based on nanosciences and technologies (IFRIS Nano)
- PhD careers datasets combining a panel-based longitudinal approach (PROFILE) and transversal approaches on mobility (MORE), plus the development of a novel approach enabling integration of a number of national datasets
- R&I policy support tools: SIPER repository of “science and innovation policy evaluations”, articulated with the OECD-World Bank Innovation Policy Platform

Take advantage of open data

- A source for complementary, more flexible, more ad-hoc indicators …
- But there are generic conditions that are underestimated
  - data is not information → need for conceptual frameworks as a precondition to select, identify & access relevant data
  - relevant data seldom stands on its own: need to be qualified, enriched & stored into databases
  - most of the times not in classical quantitative forms, requires textual/semantic analyses and also other types of visualisation

Building new datasets from open data: important considerations

- The ability to build ‘problem-based’ ontologies
- wide but selective search (enabling updating) & the role of linked open data
- The need of supports for enrichments: reference databases (e.g. universities, cf. ETER), dedicated tools (e.g. geocoding & clustering, cf. Geoclust)
- The capacity to mobilise, & interface with, other ‘structured’ resource (with issues of accessiblity, quality & harmonisation)
- The possibility to treat, interpret & visualise other aspects than numbers to support indicator production & policy reflection

What next

- Building a computer architecture that
  - enables systematic distant access
  - provides working spaces for researchers to work
  - enables access to all shared resources for accredited researchers
  - organises specific access to ‘hybrid data’ (based on project agreement)
- A very important note: a common good for European Research, that is research to be published in academic journals (no consultancy, no work for ‘restricted use’)
  → a RISIS code of conduct about use of resources

RISIS

Now time for questions and a small demo