UNIVERSITÉ PARIS-EST

École des Ponts ParisTech

IFRIS

MANCHESTER 1824

The University of Manchester

MIoIR

Toward new 'positioning' indicators: In between semantics, open data, Computer power & theory-based approaches

Philippe Larédo

OECD-EC workshop on structuring data for STI policy analysis

A long interest for ‘written traces’

- The 1980s, semantics, and the development of **Leximappe** (Callon & Turner) → Strong limitations due to power computing
- The rebirth in the 2000s with 2 developments
  - Medialab in Science Po (more for training purposes)
  - Cortext in IFRIS focused on researchers
- **CORTEXT**
  → a public good available freely on line for all public researchers
  → an ‘open approach’ (in particular to connect with all existing visualisation software)
  → a permanent tension between going further & serving users
  → today over 200 researchers worldwide use it every month for enriching, treating & visualising their textual corpuses

Building datasets: some hard lessons

- An illuminating experience for many of us here in this room: the early 2000s and the shared experiment on nano sciences & technologies
- Issue 1: How to identify relevant 'information': From keywords to complex 'queries' & to ontologies
  → see work in Knowmak & D. Maynard's presentation
- Issue 2: Researchers with multiple different corpuses (from a few thousands to hundred thousands publications or patents)
  → the need for 'reproducible approaches'
- Issue 3: how to move to unstructured data, and how to retrieve 'relevant' data within big data
  → see again Knowmak and work on 'non technological' innovations around grand challenges

Analysing datasets: rediscovering old lessons from economics (1)

- International, multinational, global
  → asymmetric distribution of actors
  → remember monopolies & cartels & their importance for policy!
  →The limit of statistical categories & the need to identify these 'powerful actors'
  →congruent activities: large firms (cf. ITPS or CIB in RISIS), universities & public actors (rankings but even more ETER & OrgReg), but also funders (BMG…)

  A note in passing: what about communication infrastructures (See VERA scenarios & our backcasting approach to policy lensing)

Analysing datasets: rediscovering old lessons from economics (2)

- Start-ups, proximity, RIS, smart specialisation
  → the role of place
  → nano and agglomeration phenomena, below existing statistical categories (countries & regions)
  → new development to geolocate activities (OECD and FUA, clustering software, RISIS & systematic efforts to geolocate key datasets)

Our understanding: the move to positioning indicators

IFRIS

- Keeping the notion of indicators, i.e.
  - a theory-based representation of phenomena under analysis
  - a dynamic view (longitudinal and not cross-sectional)
  - robustness of design & implementation
- But enabling to
  - delineate key actors' strategies
  - and dynamics of networks / places
- Implications:
  - capacity to retrieve relevant information within 'big/open data' → specific platforms to help researchers in doing so (see RISIS effort with SMS)
  - capacity to integrate heterogeneous datasets → the on-going debate about ways ahead
  - and a word of caution: open, yes… but the more we experiment, the more we face a growing privatisation of datasources

A last word about RISIS

- A first experiment based on 2 dimensions:
  - open previously a number of research datasets to the research community (attractiveness shows that there is a demand – often for more classical datasets!)
  - provide platforms to support researcher work: SMS under test, CORTEXT widely used far beyond our experimentations!
- Second round (we hope) for
  - building an architecture for systematic 'controlled' distant access
  - creating a 'living' repository of datasets of interests
  - going on developing approaches & tools for integration
  - increasing support to colleagues (training but also in indicator building)