Institute for the Study of Labor

IZA

EC public consultation on open research data

European Commission, Brussels, July 2

N. Askitas

idsc.iza.org

IDSC
INTERNATIONAL
DATA SERVICE CENTER

Q & A in a nutshell

Society seems to have an increasingly insatiable appetite for research results: policy recommendations, causality claims, stylized facts, fraudulent or erroneous claims, ideological premonitions, partial results, etc all pollute our fact space with uncertainty. This pollution may damage research credibility as a whole. Data with properly associated metadata and paradata (i.e. documentation) is our insurance against this danger.

Q: How can we define research data and what types of research data should be open?

•If “research data” means “data suitable for research” then the definition lies in the question: “ What kind of crime scene clues can help a forensics specialist?” and is: ALL DATA IS (potentially) RESEARCH DATA. If the world we live in is our crime scene then scientists are forensic experts. To find the perpetrator (causality) any and all clues can be crucial.

•An alternative definition is “any data which has been used to answer a research question is research data”. Data openness separates scientific hearsay from scientific assertion. Empirical research whose data is not available (open) at terms and conditions no worse than its researchers is simply scientific hearsay. Downside: it is harder for data owners to open it to any one researcher if they know a priori that they have to potentially open it to others.

Q: When and how does openness need to be limited?

•Openness does not mean free. Even in the case of publicly funded data it does not mean data should be free. It should be available at a reasonable cost though (like road toll or museum fees etc).

•On the opposite side proprietary does not mean closed. Data, like profits, are made using publicly financed infrastructure and just like companies pay monetary taxes it could be put to debate that they pay data taxes, with data. For example some form of aggregate reporting suitable for global economics understanding.

Q: How should the issue of data re-use be addressed?

•Re-using data can both increase return on investment (same data different research question) and safeguard scientific integrity (replicability, fraud detection/prevention). Re-use does not mean that “the same survey or experiment should not be done multiple times”. The collection of all instances of the “same” experiment may help improve our understanding of how to contact them and how to do science in general. Re-use for teaching is also a very important aspect for research in the long run. For data re-use proper metadata is a sine qua non.

Q & A in a nutshell

Q: Where should research data be stored and made accessible?

•Journals, libraries, individuals, centrally, distributed?

•The library of Alexandria which was to collect all known knowledge perished. Putting all our eggs in one basket may not be the best way to do things.

•Many of the ancient manuscripts survived because data experts (monks) in distributed specialized data centers (monasteries) provided curation with built-in backup and recovery features (same text several locations).

•If I had to place a bet I would think as an evolutionary biologist: While a large central repository may be logistically simpler to place data in it may become cumbersome for locating and getting the data out and it may not survive adverse shocks in the future. I would feel more comfortable with distributed, domain specific data curation: every research department, institute should have a domain specific repository offering curation, documentation and data education services. For that we need a new class of professionals: data experts. They are part researchers, part technologists, part librarians. Their job has a name and a title, there are degrees and career paths for them. They are what paralegals or paramedicals are to lawyers and doctors. (e.g. embedded informationalist)

Q: How can we enhance data awareness and a culture of sharing?

Openness is not necessarily about sharing (which itself is a priori a good thing). Where competition is a good thing sharing may be counterproductive. We need a scientific culture which promotes “put your data where your claims are”! If you publish results your (well documented) data must be made available easily. We cannot expect any more data sharing than sharing in general. We can promote norms of scientific contact and be clever about investments (re-use).

IZA – Shaping the Future of Labor

Nikos Askitas

IZA, 53072 Bonn, Germany

Tel: +49 (0) 228 - 38 94 525

Fax: +49 (0) 228 - 38 94 510

E-Mail: askitas@iza.org

Info: www.iza.org/home/askitas

www.askitas.com

Web: **www.iza.org**