HEALTH SERVICES RESEARCH (VIA E-MAIL)

Making Open Data Real: A Public Consultation

A response from Laurence Moseley, MBCS, CITP

Definition of Open Data

Data cannot be open unless they are easily, and cheaply (normally free of charge) available to the general public. A critical feature is that the data should be in a form in which they can easily be analyzed. That means that they must be provided in a format which is accessible to most forms of analytical software, both commercial and open-source. The government should maintain a list, which should be updated at least annually, of formats in which public bodies should provide their data. That list will vary over time, and between users, but a starting point might include:

SPSS
STATA
SAS
Statistica
Oracle
Access
MySQL
XML
PPML
Open/Libre Office
CSV
PSPP
Epidata
Rapid Miner

That list reflects my own experience and interests but the consultation should go into the detail of what such a list ought to comprise. It is encouraging that for most academic purposes ESDS and UKDA provide their data sets free or charge in at least 3 of those formats (SPSS, STATA, CSV). That should be encouraged and extended.

As new analytical software emerges it should be a research priority to consider, and where necessary provide software for, translation systems to make data easily available. Many packages do that already. For example, I regularly read Excel data into SPSS and export the results back to it. Such inter-transferability should become the norm, not the exception.

Definition of Information

I was glad to see that “information” was distinguished from “data”. I usually define “information” as “data selected, transformed, and organised so as to directly inform a decision”. The words “inform” and “information” are intimately related and both are most easily visualized when they are used in relation to the taking of decisions.

If my definition is not to be used I am reasonably happy with the definition in the consultation document as long as it is made clear that “added value” does not imply only a financial value. The recent attempts to derive indices of a non-financial kind to measure types of well-being are to be welcomed and should not be vitiated by restricting the concept to finance alone.

Discussion of tests

Clearly, one test would be the utility of data. That shows itself particularly clearly when an independent analysis challenges pre-existing prejudices. The making available of data to analysis by competent members of the public makes it possible for such challenges to succeed. Indeed, I might well argue that one of the purposes of open data is to make life uncomfortable for prejudiced or incompetent public servants or elected officials at the national, regional, or local level. Of course, the objective and competent ones would have nothing to fear.

It matters that we define the tests to be applied. However, it is equally important that we ask “who should apply those tests?” As we unfortunately no longer have the UK Statistics Commission we shall need alternative sources of wisdom and expertise. The UK Statistics Authority, ONS, and the RSS should routinely be involved when the revision of such tests is under consideration. There may be other bodies which represent particular areas of expertise (e.g. health, planning, education, or other professional bodies). The one group of

people who should NOT make the decision to apply particular tests is politicians.

Charging

There will be some commercial organisations which will wish to use public data. If they are to make a profit from the use of such data then it is fair that they should pay a reasonable charge. However, when non-commercial organisations or individuals use the data (especially voluntary bodies) any charging structure would undermine the whole purpose of open data. The system should contain a clear definition of who should be charged. However, that would consist of a list of exceptions. That implies that there will be a default that charging would not apply – unless an applicant is on that list.

Without such a default position, it is difficult for voluntary bodies or private individuals to challenge, on evidential grounds, conclusions which have been drawn by well-funded groups (e.g. developers) or by tax-payer funded groups (e.g. local authorities). Voluntary bodies frequently have greater experience and expertise in data analysis than do such funded groups. Making data open to challenge can only improve government. That is why charging would in general be counter-productive.

The scope of open data

At the moment it is impossible to define the scope. We should start with a presumption that <u>any</u> public data should be open and that there should be a list of carefully worded exclusions from that default position.

Anonymity matters for reasons of privacy and confidentiality of individuals. Naturally, one should start with a rule of anonymity. However, at times that is not enough, especially when it comes to small area statistics. Knowing that there is a left-handed, red-headed man of 63, who is a smoker, has pancreatic cancer, and lives in a particular SOA may be enough to infringe confidentiality. I see no general problem in following the Census practice of randomizing small numbers in local areas. The one exception would be for one-off research projects where one may wish to identify individuals for further sampling. One could easily have licences with special conditions for such cases. For example, applicants may be permitted to access the data only at special locations and under controlled conditions.

Encouraging compliance

Why not simply make it a legal requirement? As usual, a presumption of open data, but with clear rules about when that presumption does not apply, should be easy enough to devise and update. The FOIA currently works well, although it can at times be cumbersome. That usually applies when a request is issued to, for example, all local authorities, or universities nation-wide. Perhaps there should be two procedures, one for requests to individual bodies, and a second one for what are effectively trawling exercises. I would not wish to make either more difficult, but it is clear that occasionally what I have called trawling exercises have not been thought out, and are unjustifiable. For one-off research projects we have Ethics Committees. Why not for the use of public open data when the usefulness of a given proposal has been challenged?

Scope and objectives

At the moment this section refers to “non-personal” data. For most domains such a definition would undermine the purpose of making data open. Government at all levels interacts with people and the data need to be kept about people. Of course, the analysis will combine and segment the data in a wide variety of ways. However, it is vital that the data are available at the lowest level possible – leaving each analyst free to combine, transform, select, and present results in a form which they think is relevant, not which the data provider thinks is relevant.

I have already mentioned safeguards for confidentiality.

With regard to the devolved administrations, I do not agree that it is up to such administrations to determine which of the data should be open, or even which data should be gathered. One of the justifications for devolution was that one could have experiments in one part of the country from which other parts could learn. Indeed, one might even have one part being an experimental area while another could be a control group. Such experiments have been a major part of the development of social policy in the United States ever since the Moynihan Act, and have provided some of the most relevant, robust, and useful results anywhere in the world. The UK could learn from that experience.

If such comparisons are to be made, then comparable data must be gathered from each of the constituent administrations in the UK. That does not preclude an administration gathering data of particular interest to it, but it does mean

that all the administrations should gather certain agreed data. Such agreement should be very detailed, down to definitions of variables and units of measurement.

Opportunities for improvement

I welcome this listing of some of the barriers to access.

Cost

As this section concurs with my view that there should be a presumption of date being made available at no charge, naturally I support that element of it. However, the “business case” justification for exceptions to that presumption will have teeth only if it monitored by an independent group of professionals from say NSO or RSS. Any cost implications should also cover benefits, including non-monetary benefits. I believe that point is well made in section 6.4 (e) which should be stressed in any report on this consultation.

Additional points

I have two additional points to make both of which apply to higher education and the skills which it provides. If neither of these points are met I see very little chance of the open data initiative succeeding – or indeed of there being any move to more rational policy-making.

Skill and software for data mining

One should take the view that data sets are the corporate memory of the State (or the NHS, local authorities, etc). They are large, containing many cases, many variables, and many values. It is likely that this will mean that they will be hard to analyze, with many confounding variables. They are, and will be, incomprehensible to any human intellect without some quantification. I believe that there should be an ongoing process of data mining of such data sets. At the moment the expertise to mine them is in very limited supply in the UK, especially in the area of public services. This is in marked contrast to what is happening in private industry, especially in the USA.

In private business, it is now commonplace for data mining and other artificial intelligence software to be applied to corporate records. Those records may be financial, personnel, performance, and either structured or free text data

(indeed, the next step forward may be in the robust analysis of free-text data, the very opposite of what many social scientists call “qualitative” data). It is already being used by bodies like Amazon, Google, The Air Transport Safety Board, and many other organisations to learn lessons from, for example, calls received by call centres or help desks or from printed accident reports. They often involve analyzing terabytes of data – or more. Such applications are at the moment less common in the public services.

In the UK we have very few people trained to use currently available software. Even if we were to restrict our estimate of the need for such personnel to one per organisation (local authority, NHS Trust, University, etc), it would still come to a shortfall of over 1,000 – and that is probably a marked under-estimate. The government should, as part of its manpower planning, encourage universities to start to train people with that sort of expertise. That will be impossible unless we start to turn out more students who leave secondary school with at least a modicum of A-level Mathematics.

However, the expertise will be nugatory if there is no modern software to use. The government should therefore ensure that each university has available for general use free-of-charge at least one such data mining software package, and that funding should be contingent upon the teaching of such software to most students. There are at least four front runners at the moment and perhaps the government should initially plump for one of them. It would be a modern equivalent of the former concept of the well-found laboratory.

The contribution of education

I have asked above for an increase in the number of trained people in data analysis, which normally covers Statistics and Data Mining. One would want such people to have a good knowledge of society as well. They can be trained only if we have a core of social science teachers who are competent in quantitative methods. It would be remiss of me not to draw attention to the 2009 report of the ESRC Strategic Advisor on Quantitative Methods (MacInnes J [2009]. Proposals to support and improve the teaching of quantitative research methods at undergraduate level in the UK. University of Edinburgh. Report to ESRC 12 October 2009) in which he estimated that fewer than 1 in 10 sociology or social policy academics in the UK was capable of giving a basic methods course (let alone the advanced training for which I am calling). It is a matter of urgency that the government does all that it can (e.g. by financial inducements or penalties) to improve the calibre of our social science

academic staff. That in turn will require a substantial movement in secondary education towards encouraging a much greater number of pupils to study the sciences and Mathematics.

Overall, I welcome the initiative of the government in trying to make data more open and available. In an open society like our own it should be a given that wisdom and expertise is not a prerogative of public servants. There is a large stock of experience and expertise among the general population, including the relatively untapped group of retired professionals. That should be used. If the result of the current consultation were to be the greater use of such latent expertise, our country would benefit and our grand-children would thank us.

_____ Contact details _____
Laurence Moseley
Emeritus Professor of Health Services Research
Glyntaff Campus
University of Glamorgan
Pontypridd CF37 4BD