United Nations ECE/CES/2013/27

Economic and Social Council

Distr.: General
18 March 2013

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-first plenary session
Geneva, 10-12 June 2013
Item 4 (a) of the provisional agenda
Drivers for micro-data access

Micro-data exchange and the challenges of Open Data and transparency

Note prepared by the United Kingdom and the Organisation for Economic Co-operation and Development

Summary

This paper addresses the challenges of Open Data for micro-data exchange. To serve the research community many statistical offices have established national data archives, on-site data laboratories and remote access facilities. However, serving the general public and the emerging industry of information entrepreneurs presents a different set of challenges. Producers of official statistics should take into account pressures arising from the challenge of anonymisation for Open Data. Offices will need a suite of solutions, including laboratory/remote access facilities, scientific use files, public use files, and Open Data products. In the last decade there has been a considerable improvement in services for the research community, and this improvement will continue. However, the development of public use files to the Open Data standards, to feed into the Big Data movement, is much less developed. This paper, therefore, considers these issues that may become an important area of work in the decade to come.

I. Introduction

1. This paper is based on the experiences of the Office for National Statistics of the United Kingdom, but is not a statement of the policy of that office or of the United Kingdom government. It discusses some of the policy and legislative initiatives for Open Data in the context of micro-data in national statistics institutes. The drivers and solutions for Open Data in micro-data are distinct from those for the exchange of confidential micro-data, but challenges in either activity cannot be addressed without an understanding of them both.

2. For the purposes of this report, "Open Data" are data (datasets) that are:

(a) Accessible to anyone and everyone, ideally via the internet;

(b) In a digital machine readable format that allows interoperation with other data;

(c) Available at reproduction cost or less;

(d) Are free from restrictions on use and re-use.

3. For the purposes of this report:

(a) Micro-data are the data that quantify observations or facts with respect to a particular reporting unit;

(b) Statistics are combinations of data which may reveal patterns in observations and facts.

4. Thus for the purpose of discussion of Open Data, the relevant micro-data are the individual records of direct observations and facts obtained through instruments of statistical inquiry (such as surveys), registers, or administrative sources. The data are modified before dissemination only where necessary to address obligations of confidentiality to the subjects of the observations. Such data might be called Open micro-data. The purpose of the paper is to explore expectations that micro-data exchange should increasingly be based on Open Data principles, and the issues that arise for producers of official statistics.

II. Drivers for Open Data

5. The drivers for development of micro-data as Open Data include the adoption of *scientific principles*, support for democracy, stimulus for social and economic growth, and response to legislation and public policy.

A. Scientific principles

6. Open Data supports the status of statistics as a scientific discipline.

7. The Fundamental Principles of Official Statistics[1] require the methods of dissemination for official statistics to be decided according to *scientific principles*. Users' interpretation of the data is facilitated when they are presented according to *scientific standards*.

[1] http://unstats.un.org/unsd/methods/statorg/FP-English.htm

8. What are these *scientific principles* and *scientific standards* for the dissemination and presentation of data? The Fundamental Principles suggest they pre-exist and are familiar.

9. The Royal Society published its report "Science as an Open Enterprise" in June 2012[2]. The report reflects on science as a self-correcting process. Theories can be independently corroborated, invalidated or improved. Findings, and the supporting data, are presented to the widest possible audience for further development.

10. Producers of Official Statistics should challenge themselves as to whether their published statistics - their assertions of patterns in observations and facts - can be independently corroborated, invalidated, or improved. Open Data allows this scientific principle to act on Official Statistics. Published methodology allows abstract challenges only, unless the data as used by the producer of the statistic is also available to another independent party. This can be achieved to some extent through channels of controlled access such as statistical peer review, or through research access to data. However, maximum transparency is achieved when no restrictions, selections or pressures (whether real or imaginary) are brought to bear upon the independent scrutiny of the statistics.

11. Public trust and confidence may suffer if the data that underpin scientific information and statistics are not freely available along with the methods used. For example, public trust in the publications of the Climate Research Unit of the University of East Anglia was undermined when some data and methods were withheld from the public domain. Two inquiries were held, and the Government's response was to recommend the full disclosure of the raw data along with the necessary computer programmes and methodologies to replicate the results.

> *"The disclosure of raw data and sufficient details of the computer programmes is paramount in encouraging people to question science in the conventional way, challenging existing work, enabling validation of it and coming forward with new hypotheses".*
>
> - Government Response to the Science and Technology Committee's First Report of Session 2010-12[3]

B. Indispensable element for democracy

12. The Fundamental Principles of Official Statistics position official statistics data as "...an indispensable element in the information system of a democratic society."

13. The Open Government Partnership is an initiative to make governments better. It was founded in September 2011 with 8 government members, and is now expanded to 55 government members. The Partnership declaration includes a commitment to increase the availability of information about government activities. The partners to:

> *"...proactively provide high-value information, including raw data, in a timely manner, in formats that the public can easily locate, understand and use, and in formats that facilitate reuse."*
>
> - Open Government Declaration, Open Government Partnership[4]

[2] http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

[3] http://www.publications.parliament.uk/pa/cm201012/cmselect/cmsctech/496/496.pdf

[4] http://www.opengovpartnership.org/open-government-declaration

14. Implementing the Fundamental Principles, most national statistical systems will incorporate statutory objectives that include reference to support for the scrutiny of public policy. The primary objective of the United Kingdom Statistics Authority and its executive office, Office for National Statistics (ONS) reads:

> *"The Board is to have the objective of promoting and safeguarding the production and publication of official statistics that serve the public good. [T]he reference to public good includes in particular informing the public about social and economic matters, and assisting in the development and evaluation of public policy."*
>
> - Statistics and Registration Service Act 2007[5]

C. Open Data for economic and social growth

15. The Vickery Study[6] conducted for the European Commission estimated that the direct and indirect economic gains when public sector information is open for re-use are in the order of 140 billion euros.

16. The Organisation for Economic Co-operation and Development (OECD) Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information (2008)[7] makes several relevant recommendations for Open Data in statistical micro-data, in order to "increase economic and social benefits in particular through more efficient distribution [of information], enhanced innovation and development of new uses".

17. Member countries are recommended to adopt its principles, which include these of relevance to statistical micro-data:

(a) *Openness*. Maximising availability, with open access as the default rule;

(b) *Transparent conditions for re-use*. Non-discriminatory conditions, eliminating exclusive arrangements and removing unnecessary restrictions;

(c) *Asset lists*. Inventories of Open Data, published online;

(d) *Quality and integrity*. Use of best methods in data preparation and protection from misrepresentation;

(e) *Pricing*. Free of charge, or cost recovery only;

(f) *International*. Facilitation of cross-border use and interoperability.

18. These recommendations provide a framework for National Statistical Institutes (NSIs) to assess their adoption of Open Data principles in micro-data.

D. Legislation and public policy for Open Data

19. Legislation and policy for Open Data take two basic forms. First, those laws and public policies that oblige the public sector to *push* Open Data. And second, those laws and public policies that entitle users to pull Open Data from the public sector.

[5] http://www.legislation.gov.uk/ukpga/2007/18/section/7
[6] http://ec.europa.eu/information_society/policy/psi/index_en.htm
[7] http://www.oecd.org/internet/interneteconomy/40826024.pdf

1. 'Push' legislation and policy

20. Many NSIs are subject to provisions in law and policy that establish an expectation that they will push Open Data to the public under a regulatory regime.

21. For example, in the European Union the public sector information market was first regulated by the Directive on the Re-use of Public Sector Information[8]. The Directive is now under revision to strengthen its provisions and to respond to the digital world. The Directive does not, strictly speaking, oblige the production of public use documents, but provides a positive regulatory framework for adoption in the legislation of member states, with an assumption that public sector information should be Open within reasonable limitations. For example, the Directive enshrines the principle that the total income from supplying and allowing re-use of data should not exceed the cost of collection, production and dissemination and a reasonable return on investment.

22. Implementation of the Directive has been achieved in all member states, and some have gone beyond its requirements in both national law and national policy. For example, the UK has adopted national Public Data Principles[9] binding on the public sector that include an obligation to actively encourage the use and re-use of departments' public data.

23. The public sector copyright rules in the United Kingdom ("Crown Copyright") have been modified to establish a 'push' principle. The default copyright licence for the UK public sector is the Open Government Licence (OGL)[10]. The OGL is a positive, proactive, 'push' licence:

> *"[The government] grants you a worldwide, royalty-free, perpetual, non-exclusive licence to use the Information... You are free to copy, publish, distribute and transmit the Information; adapt the Information; exploit the Information commercially for example by combining it with other Information, or by including it in your own product or application."*
>
> - UK Open Government Licence

24. A United Kingdom public body that wishes to obtain an exception to marginal cost recovery must seek accreditation to the Information Fair Trader Scheme[11], submit a business case, and be subject to the scrutiny of the Office for Public Sector Information.

25. In the future 'push' policies and structures can be expected to develop, often in response to popular campaigns[12]. The United Kingdom is now launching the Open Data Institute[13] as a public/private partnership. The Open Data Institute (ODI) is co-chaired by Nigel Shadbolt and Tim Berners Lee. It is established expressly to push Open Data out of the public sector:

> *"...a collaboration between our leading businesses and entrepreneurs, universities and researchers, government and civil society to unlock enterprise and social value from the vast amount of Open Government Data now being made accessible."*
>
> - About the ODI, www.theODI.com

[8] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32003L0098:EN:NOT
[9] http://www.data.gov.uk/library/public-data-principles
[10] http://www.nationalarchives.gov.uk/doc/open-government-licence/
[11] http://www.nationalarchives.gov.uk/information-management/ifts.htm
[12] http://www.freeourdata.org.uk
[13] http://www.theodi.org/people/nrs

2. ‘Pull’ legislation and policy

26. Many NSIs are also subject to examples of laws and policies that empower the public to pull data from the public sector. Examples include the Environmental Information Regulations, and the Freedom of Information Acts of many countries. It is usually the case that statistical data are subject to the pull power of these laws and policies, with exemptions only where it can be shown that the data are confidential or scheduled for future publication in the form requested. The impact of such laws on statistical data can be significant:

(a) ‘Pull’ laws typically have a time limit for compliance. NSIs in the United Kingdom must respond with either the data requested or an explanation of why an exemption applies within 20 days of the request;

(b) In combination with the Open Government Licence (or equivalent), limitations and conditions on use and re-use cannot be imposed;

(c) The applicant typically has a right to appeal a decision to refuse or to comply on in part to the request for data. The appeal is typically heard by an Information Tribunal or other such non-statistical authority. Whilst the statistics office may present its evidence for why the data should not be disclosed (for example, for reasons of confidentiality), the decision lies with the Information Tribunal;

(d) Typically, the application for data is ‘purpose blind’. The applicant does not need to explain why the data are request, nor what their intended uses are for the data;

(e) Often there is no exemption for data that are of poor quality;

(f) The release of data under ‘pull’ legislation is often of a precedent-setting nature;

(g) Unless handled carefully, release of data under ‘pull’ legislation can look like privileged access;

(h) Release of datasets can undermine confidentiality through ‘mosaic attack’;

(i) Disclosure control issues are often non-obvious to the uninitiated.

27. The push and pull laws and policies are now enabling information entrepreneurs, from financial analysts to App designers, to invest in confidence in expert in data exploitation. The industry resulting exerts its own pull on the public sector data owners.

III. Challenges of producing Open (micro)Data

28. How are national statistics offices responding to the challenges of Open Data? With their statistics, we might argue the response is very good. Fundamental practices in statistical presentation and dissemination are by their nature compatible with Open Data principles. However, where the challenges of Open Data are applied to the ‘raw’ micro data of national statistics offices, the response is currently less convincing. Typically, a statistics office will assign their data assets into three categories:

(a) “Statistics” for publication or for use in public administration - usually aggregate data or visualisations;

(b) “Research datasets” for statistical research uses under controlled conditions - usually micro-data and usually marked ‘confidential’;

(c) “Production data”, being the micro-data or aggregate data within statistical production systems and which is not expected to be used for any purpose other than the derivation of statistics and the derivation of research datasets.

29. The concept of 'public use files' is familiar, but current practice is usually to assign them to the "research datasets" category. Usually, that is the correct classification, because most public use files cannot be called Open Data, in that there are terms and conditions of use, and there are issues of timeliness, completeness, granularity, format, and interoperability. Neither NSIs nor the users of public use files currently have the expectation that such files could be used to replicate the statistics produced by the NSI - not least because the NSI is not using the public use file themselves to produce the statistics. Thus the scientific principle of independent corroboration, verification, or improvement of the statistics is not achieved through public use files in the manner in which they are currently produced.

30. It is not common current practice to recognise Open Data as a category for micro-data and assign information assets to that category for dissemination under those conditions. This may be because of the many challenges presented by producing Open micro-data

A. Confidentiality

31. Confidentiality issues are the most frequently cited reason for not producing Open micro-data. Very often it will be an entirely legitimate reason. However, it is common practice for a NSI to consider all its unpublished data to be confidential data either by default, or by virtue of its status as 'data held for the purpose of official statistics'. This practice is unlikely to survive the push or pull legislation and policies for Open Data. First, it should be recognised in the policies of the NSI that some data assets are inherently less likely to raise confidentiality issues than others - for example, public sector budget and expenditure data, or prices data, may be identified micro-data but not confidential due to the information being already available in the public domain. Second, a more systematic and critical analysis of whether a data asset is truly disclosive or not should be established in NSIs. This may involve so-called 'penetration tests', whereby a trusted party is provided with a candidate dataset and allowed to see whether, under a reasonable test, any private and confidential information about identified individuals can be discovered from the data. If such personal information can be discovered, the NSI has obtained independent evidence to present to an Information Tribunal or equivalent, helping them to sustain a challenge against them withholding the data from an applicant. If such personal information cannot be discovered through the penetration test, the NSI may have a candidate Open micro-dataset.

32. Data Protection Supervisors recognise the difficulty of distinguishing between personal and non-personal information. The Article 29 Working Party Opinion on the concept of personal data[14] helps NSIs with the task of identifying what is, and what is not, personal data. The United Kingdom Information Commissioner has recently published a Code of Practice on anonymisation of personal data and managing data protection risk[15], with the help of Office for National Statistics and many others.

B. Data quality and reputation

1. Of all the factors that affect public confidence in Official Statistics, data quality may be the most important. A reputation for good quality statistics is hard won, and easily lost. It is entirely understandable that Open Data, and especially Open micro-data, is seen

[14] http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf

[15] http://www.ico.gov.uk/news/latest_news/2012/~/media/documents/library/Data_Protection/Practical_application/anonymisation_code.ashx

as a threat to a reputation for good quality statistics. The threat may arise from two scenarios. First, the data may well be of poor quality. If the data are 'pulled', they may be released before quality assurance and before additional input data comes in to the NSI. Second, the data may be of good quality, but may be interpreted differently or even wrongly by the other party. The NSI may have to explain the differences and justify them to a public that may be prefer the non-government party to be right. These issues are to be dealt with by excellent metadata, a readiness to comment on the quality of other the party's analysis, and a campaign of education for the public on matters of quality and analysis. The cultural issue can be addressed by reference to good scientific practices – good quality statistical methods have nothing to fear from the reuse of the underlying data in an Open Data world.

C. Dissemination standards

34. The standards for dissemination of Open micro-data are demanding. Few NSIs will achieve the expected standards in the short term. The United Kingdom's public sector is challenged by the 5 star Open Data expectation[16]. In particular, the use of non-proprietary formats and uniform resource identifiers (URIs) for all Open micro-data releases is a challenge, and may not even be welcomed by current users.

D. Using websites designed for other purposes

35. Compared to aggregate statistics, Open micro-data may be very large in volume and may have a format that is incompatible with existing dissemination channels such as NSI websites. Reengineering NSIs websites for Open Data may be a low priority, especially if there are hard limitations such as bandwidth, file size, and proprietary format restrictions. A solution is to adopt alternative dissemination channels designed specifically for this challenge. For example, the United Kingdom has established www.data.gov.uk as a shared service available all public sector institutions. It should also be born in mind that, as Open Data, restrictions on government information security standards are not applicable, allowing private sector dissemination channels to be used. Google Public Data is just one example, currently hosting Open Data from OECD, Eurostat, the IMF, the World Bank, DeSTATIS, and the United States Census Bureau. Although Google Public Data is primarily an aggregator of data released through other channels, it could be used as a channel of first release of Open Data.

E. Authenticity and attribution

36. NSIs are the attributed original source of Open micro-data, but legitimately may have concerns about modifications to the content of the data that affect the products of analysis but which are unrecorded and/or unexplained by users. In other words, the authenticity of the data is lost, but the attribution of 'Source: NSI' persists. The solution is the use of 'persistent identifiers' by the producer of the Open Data. Every Open Data asset should be associated with a unique Direct Object Identifier/ Uniform Resource Name, in combination with a permanent Universal Resource Locator and a unique citation for electronic publications using the (International Organization for Standardization) ISO 690-2 standard. This will allow the NSI, and users, to readily identify and retrieve the original and uncorrupted source data.

[16] http://5stardata.info

F. Metadata

37. The metadata standard most relevant for National Statistics is the Statistical Data and Metadata Exchange (SDMX), but this is optimised for aggregate statistical data rather than Open micro-data. If the NSI does not have expertise in metadata standards for Open micro-data it might seek assistance to pass this obstacle from partners such as data archives, in particular the members of the Council of European Social Science Data Archives (CESSDA).

G. Compliance with Code of Practice undertakings for scheduled releases and equality of access, versus obligations in Open Data standards for timelines.

38. NSIs will want to ensure that statistics scheduled for future release, according to the good practice for predictable and pre-announced release dates, are not undermined by similar statistics derived from re-use of Open micro-data. Where Open micro-data are pushed, this is easily achieved by proper scheduling of the two releases. However, where Open micro-data are pulled, there may be conflict with release schedules. A solution might be to bring forward the scheduled official release where possible, or to proactively explain the potential availability of unauthentic and unauthorised statistics from other sources in advance of the official publication.

39. Where Open micro-data are pulled from NSIs it is important to preserve the principle of equality of access. This is an important statistical principle, but it is an important Open Data principle too. If Open Data or Open micro-data are provided as (for example) attachments in an email to an applicant, this will give the appearance of privileged access. The use of the data may appear as a 'scoop' for the user. NSIs should establish a disclosure log[17] within their website, used to present ad-hoc releases of Open Data that may have been pulled by a particular user but is clearly simultaneously available to all. The user who pulled the data should be provided with a link to the URL only when the data resource is available to everyone.

H. Allocation of resources

40. The resource use profile of Open micro-data is different to research use files. Initial resource costs are high, despite the raw information already existing, by definition. The analysis of disclosure risk, the preparation of metadata, the assembly into an open format, etc, are all up-front costs. By definition, the income from Open Data cannot exceed costs plus a reasonable return on investment into their production. However, once produced, the on-going costs of Open micro-data are negligible, especially if the dissemination channel is a shared resource or a cost to another party. In contrast, the resource use profile for research datasets is low initially, as the data are in effect unchanged from their status in the production environment. The resource costs for research datasets are high in the maintenance of a secure research environment, the accreditation of researchers and research projects, and the checking of outputs. NSI budgets are, typically, oriented towards meeting on-going costs and easier to allocate, even if in aggregate over a number of years the burden on resources of research datasets is greater than the investment in Open micro-data of equivalent value to society.

[17] http://www.ons.gov.uk/ons/about-ons/what-we-do/FOI/foi-requests/index.html

IV. Response to the challenges - Draft recommendations

41. This report discusses the drivers and challenges to the exchange of micro-data under the emerging standards of Open Data.

2. The benefits of achieving the Open Data standard for official micro-data are clearly very substantial:

(a) Micro-data as Open Data allows the *scientific principles* of corroboration, validation, and improvement of the Official Statistics derived from the same sources;

(b) Open micro-data can be exchanged without costly and bureaucratic administrative obstacles;

(c) Open micro-data can be used for any purpose, including those never envisaged when they were produced;

(d) Open micro-data allow NSIs and third parties from all sectors to cooperate and collaborate fully on a shared information resource;

(e) Open micro-data are an additional information asset category for a NSI, making the NSI an important and (hopefully) valued partner in the public sector for the future of modern Information Societies;

(f) Open micro-data encourages the development of information entrepreneurs, fostering economic and social growth.

43. The obstacles and challenges are equally substantial:

(a) Confidentiality risks, and concepts, have to be addressed;

(b) Logistical issues arise, presenting challenge to the business architecture of NSIs;

(c) Authenticity and identification of assets must be addressed;

(d) The expected standards for Open Data are high.

A. Recommendations

- NSIs should adopt Open Data standards for their routine statistical production. This ensures the NSI becomes familiar with Open Data challenges before the particular challenges of Open micro-data are tackled.
- NSIs should include a category of Open micro-data in their information asset registers.
- NSIs should collaborate to spread the costs of developing methodologies for creating Open micro-data.
- NSIs should explore alternative dissemination channels for Open micro-data if existing architecture is unsuitable.
- NSIs should work closely with their national data protection supervisor on anonymisation standards and the concept of personal data.
- NSIs should consider in advance how obligations under Codes of Practice for statistics can be upheld when a Open micro-data are produced.

- NSIs should use the skills and experience of data curators, computer scientists, knowledge and information management professionals, data archives, and national libraries, to assist with the preparation and dissemination of Open micro-data.
- NSIs should prepare a communication strategy with key statistical users, and the public, to address novel Open Data products.