

Ensuring effective and smooth flow of data throughout the data life cycle

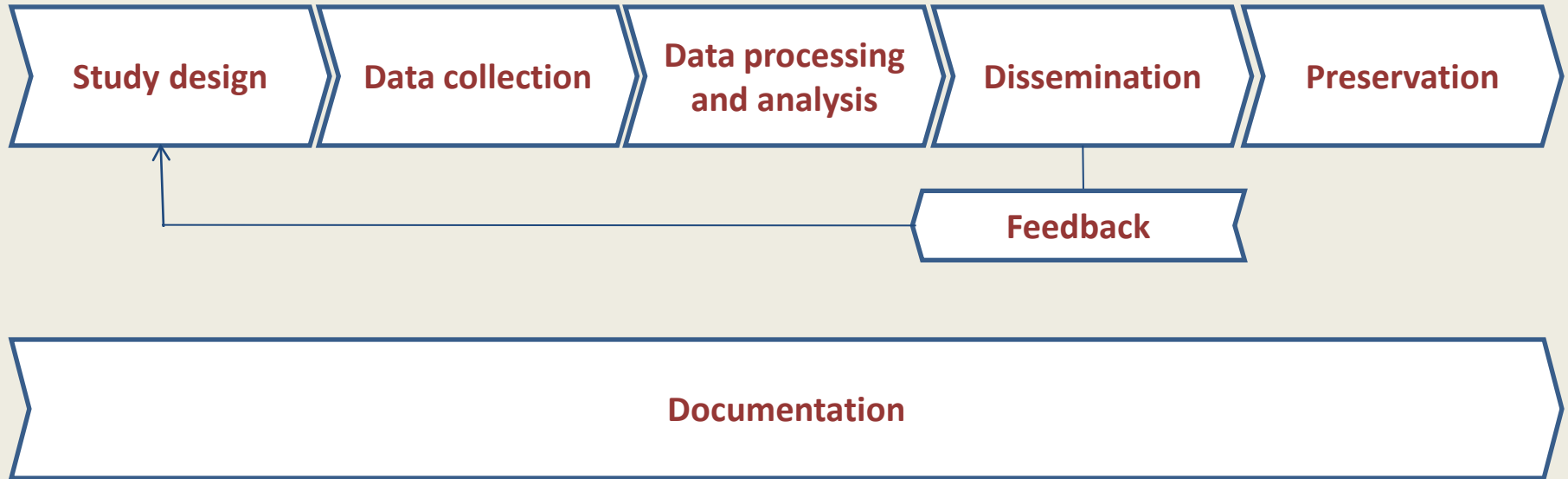
Standards, practices and procedures

Olivier Dupriez
World Bank / IHSN

*Eighth Management Seminar for the Heads of
National Statistical Offices in Asia and the Pacific
(3–5 November 2009)*

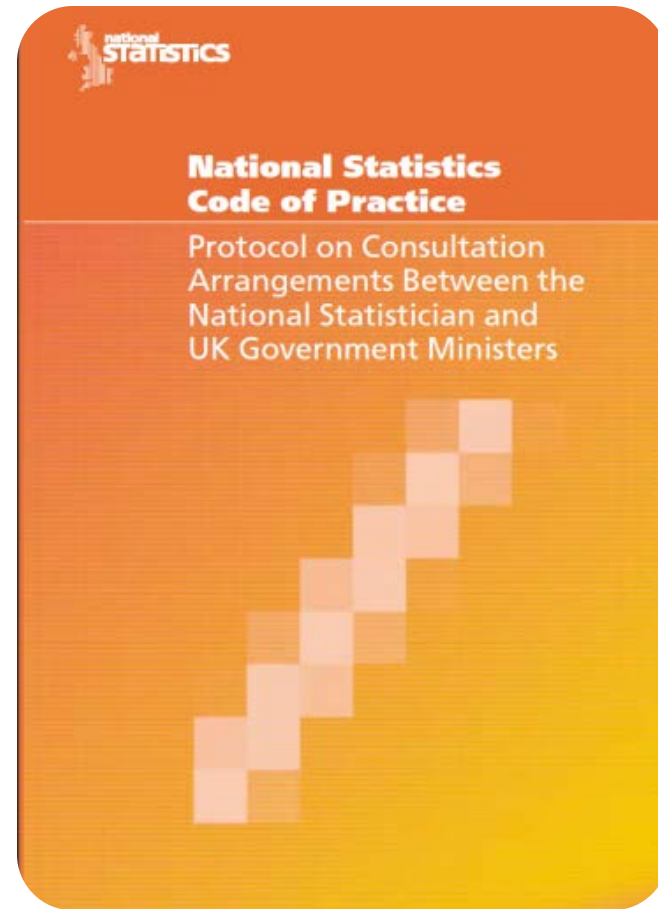
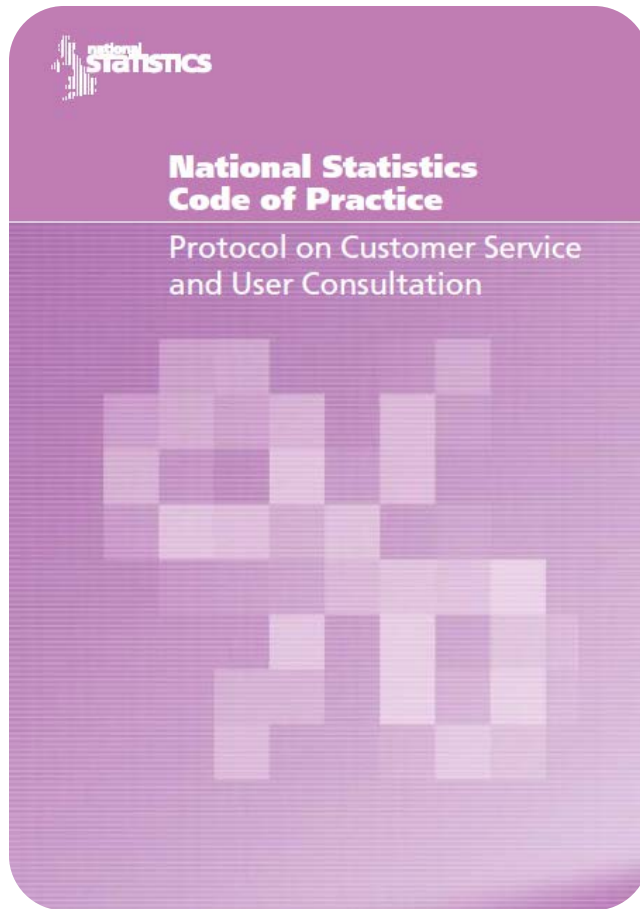
The great thing about standards is that there are so many to choose from ...

The data life cycle



RELEVANCE

- Count what counts
- Know who your users are and what their needs are
- Consultation with users can be formal or informal (solicited or not)
- Many ways to communicate with users: conferences, visits, correspondence, analyzing feedback, website usage statistics, etc.
- Engagement with users must be inclusive and coordinated
- Regularly assess relevance and publish outcome



CONSISTENCY, INTEGRATION

- Data production is fragmented, vehicle-driven. This causes redundancies, inefficiencies, disharmonies
- Solution: integration
 - Use of common classifications, geographic referencing standards, definitions, questions and instructions across the statistical system (keep the option to diverge from standard, but with clear and explicit justification).
 - Take advantage of international good practices (SNA, etc)
 - Maintain a corporate inventory of holdings (metadata)
- Integration requires better communication within the system, but makes communication with suppliers and users much easier.

(Inter)national standards for integration

Source of drinking water in country [X]: 9 surveys, 9 different ways of collecting data.
Definition of “household” and of “urban/rural” also varies from survey to survey.

DHS 1999

11 Piped Into Residence/Yard/Plot
12 Public Tap
21 Well in Residence/Yard/Plot
22 Public Well
31 Spring
32 Stream
33 Pond/Lake
34 Dam
41 Rainwater
51 Tanker (Truck)
52 Tanker Vendor
61 Bottled Water
71 Borehole
96 Other

2006 Census

1 Pipe borne inside dwelling
2 Pipe borne outside dwelling
3 Tanker Supply/Water Vendor'
4 Well
5 Bore-hole
6 Rain water
7 River/Stream/Spring
8 Dugout/Pond/Dam/pool
9 Other

DHS 2003

11 Piped Into Dwelling
12 Piped Into Yard/Plot
13 Public Tap
21 Open Well In Dwelling
22 Open Well In Yard/Plot
23 Open Public Well
31 Protected Well/Borehole In Dwelling
32 Protected Well/Borehole In Yard/Plot
33 Protected Public Well/Borehole
41 Spring
42 River/Stream
43 Pond/Lake
44 Dam
51 Rainwater
61 Tanker Truck
71 Bottled Water
96 Other

CLFS 2000

1 Tap/pipe Inside house
2 Tap/Pipe Outside House
3 Tube/well
4 Manual Well Protected
5. Handpump
6.Ponds/Stream/River/Rain water

CWIQ 2006

1 Pipe borne water treated
2 Pipe borne water untreated
3 Bore hole/hand pump
4 Protected well
5 Unprotected well
6 Rain water
7 River, lake or pond
8 Vendor, truck
9 Other

GHS 2006

1 Pipe borne water treated
2 Pipe borne water untreated
3 Borehole/hand pump
4 Well/Spring Protected
5 Well/Spring Unprotected
6 Rain Water
7 Streams/Pond/River
8 Tanker/Truck/Vendor
9 Other

MICS 2007

11 Piped into dwelling
12 Piped into yard or plot
13 Public tap/standpipe
21 Tubewell/borehole
31 Dug well/Protected well
32 Unprotected well
41 Protected spring
42 Unprotected spring
51 Rainwater collection
61 Tanker-truck
71 Cart with small tank/drum
81 Surface water (river, stream, dam, lake, pond, canal, irrigation channel)
91 Bottled water
96 Other (*specify*)

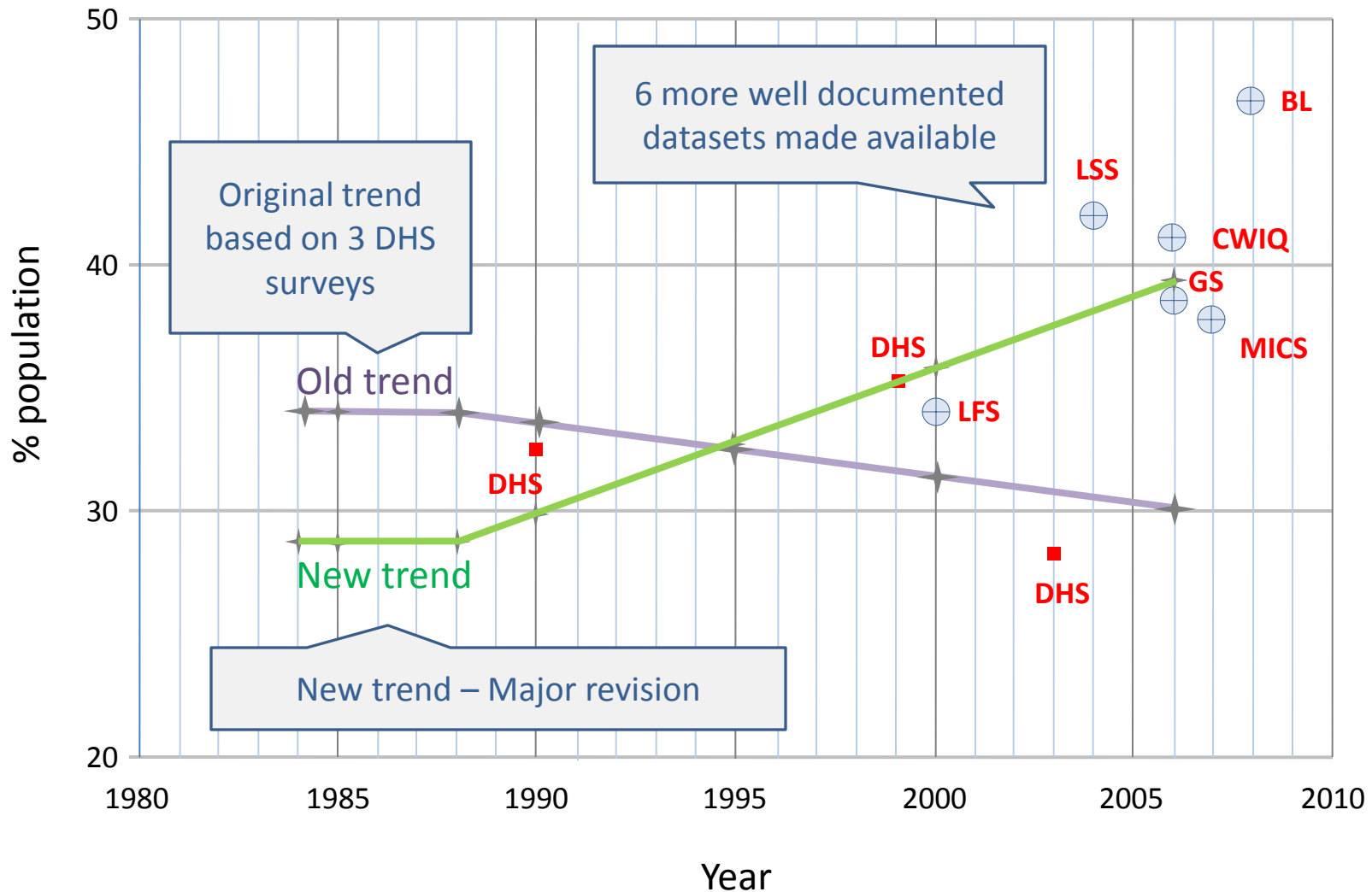
Sector Baseline 2008

Protected
a. Household Connections
b. Boreholes with hand pump
c. Motorized borehole
d. Protected Dug well
e. Public Standpipe
f. Rain water harvesting
g. Protected Spring
Unprotected
a. Unprotected Traditional hand dug wells
b. Unprotected wells
c. Vendor provided water
d. Bottled/sachets water
e. Tanker truck provided water
f. Streams
g. River
h. Pond
i. Broken pipes

NLSS 2004

1 Pipe borne water treated
2 Pipe borne water untreated
3 Borehole/hand pump
4 Protected Well
5 Unprotected

Country X - Rural access to improved drinking water sources



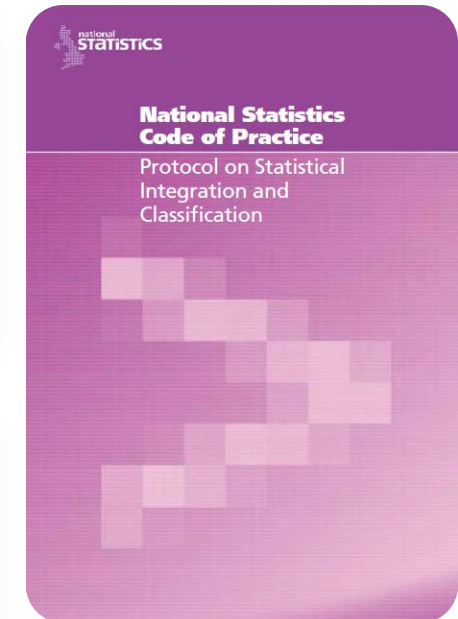
Standards and classifications should be accessible on your website, with guidelines for their application.



The screenshot shows the top section of the Statistics Canada website. It features a blue header with a red maple leaf logo and the text "Statistics Canada" and "www.statcan.gc.ca". Below the header is a navigation menu with links for "Français", "Home", "Contact Us", "Help", "Search", and "canada.gc.ca". A search bar is visible on the right. The main content area has a white background with a red sidebar containing the text "Definitions, data sources and methods". The main heading is "Definitions, data sources and methods" and the sub-heading is "Our purpose is to provide information that will assist you in interpreting Statistics".



The screenshot shows the top section of the Australian Bureau of Statistics website. It features a green header with the text "Australian Bureau of Statistics". Below the header is a navigation menu with links for "Home", "First Visit?", "Statistics", "Services", "Census", "Themes", and "Methods & Standards".



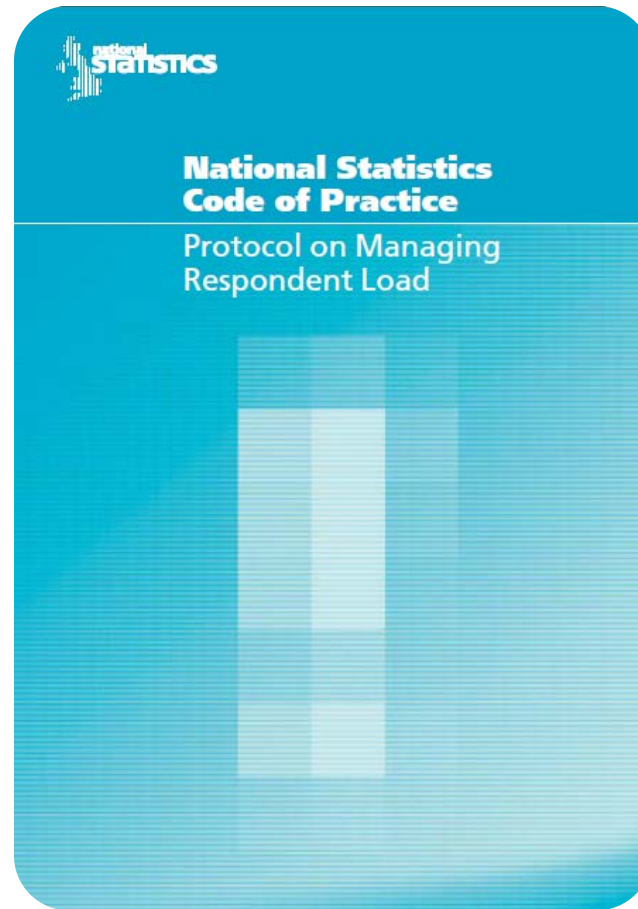
IN SEARCH OF DATA INTEGRATION: NO MATCHES FOUND

Gordon E. Priest, Statistics Canada

<http://www.amstat.org/sections/sgovt/priest.pdf>

TRUST

- Respondents will provide more honest information when the trust is high and the burden is low
- Persuasion is better than obligation
- Respondents must be informed of the intended uses of the data and be convinced that there is a clear benefit
- A guarantee must be given to respondents that no statistics will be produced that are likely to identify them unless specifically agreed
- Laws and regulations do not provide a “user friendly” set of principles. Important to have a code of practice and related protocols to communicate with data suppliers.



<http://www.ons.gov.uk/about-statistics/ns-standard/cop/protocols/index.html>

REPLICABILITY

- We must know the exact process by which the data were generated and the analysis produced
 - "The **replication standard** holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author."

<http://gking.harvard.edu/files/replication.pdf>

- Crucial to defend your results, train new staff, etc.
- Importance of documentation and preservation (must be imposed to all including consultants)

CONFIDENTIALITY

- Everyone must be aware of the obligation to protect confidentiality and of the fact that this obligation continues after completion of service (including consultants)
- Data identifying respondents will be kept physically secure

TIMELINESS

- Release calendar and arrangements must be open and pre-announced
- Statistics will be released as soon as practicable once they are judged fit for purpose
- Release the data to all interested parties simultaneously. Early access only in exceptional circumstances, and not for personal advantage
- Statistics must be released separately from statements by ministers (and before)
- Timing not to be influenced by the content of the release

ACCESSIBILITY

- Promote equality of access
- As far as possible, the price should not be a barrier to access
- The web is the primary means of providing general access, but other forms of dissemination must be maintained (paper, CD-ROMs, etc)
- Choice and flexibility in the formats (monitor the demand !); respond to changing expectations

QUALITY, CLARITY, USABILITY, PORTABILITY

- Disseminate data with lots of metadata:
 - To help users **understand** what the data are measuring and how the data have been created
 - To help users **assess** the quality of the data
- **Metadata standards** and XML technology are convenient ways to ensure completeness and portability of metadata (provide “checklist” of elements)
 - SDMX for time series data (ISO)
 - DDI for microdata *<ddi>*




VISIBILITY


- Metadata also helps users **find** the data; any cataloguing system is based on metadata
- “Discovery metadata” should be made available in a comprehensive catalogue covering all national statistics
- Monitor the demand. Make use of log files and usage statistics of your website


Example: monitoring web usage using Google analytics


Site Usage


 **2,943** Visits

 **60.35%** Bounce Rate

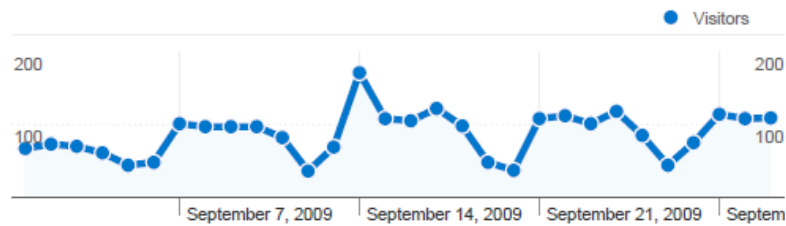
 **7,278** Pageviews

 **00:02:45** Avg. Time on Site

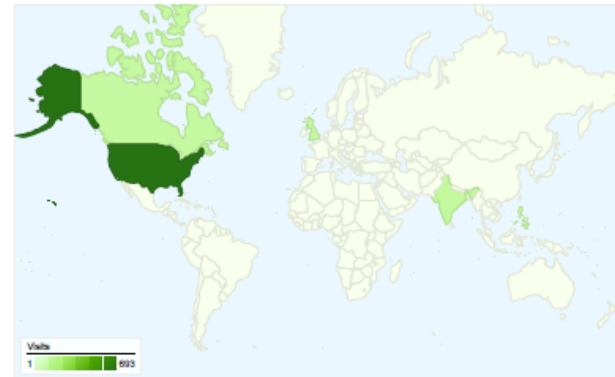
 **2.47** Pages/Visit

 **72.14%** % New Visits

Visitors Overview



Map Overlay world



Traffic Sources Overview



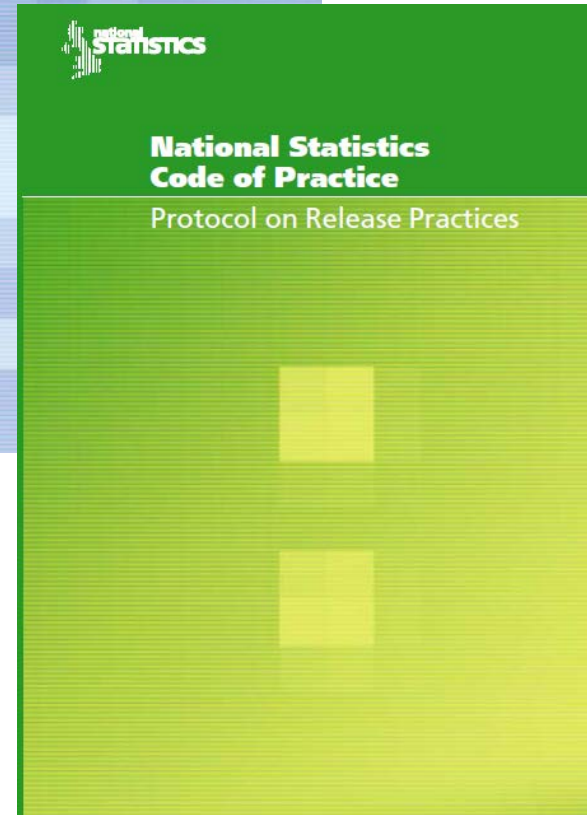
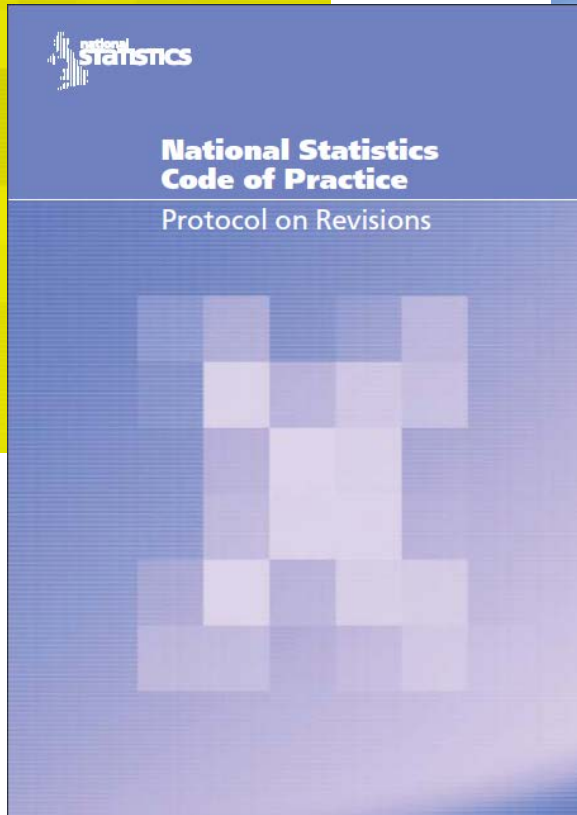
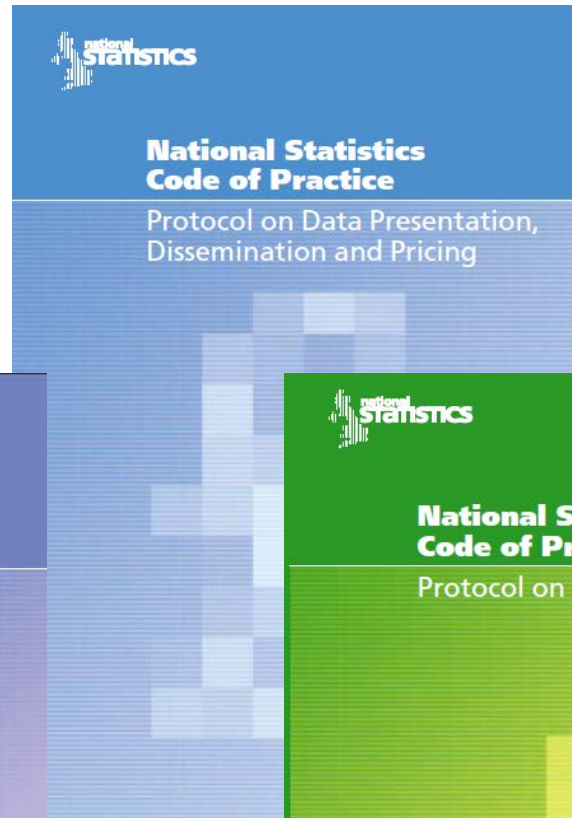
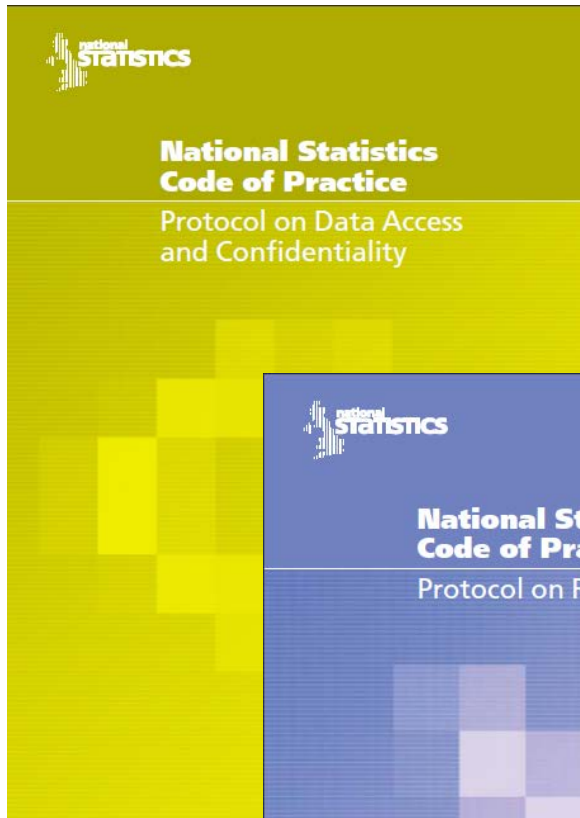
- **Referring Sites**
1,304.00 (44.31%)
- **Direct Traffic**
825.00 (28.03%)
- **Search Engines**

Content Overview

Pages	Pageviews	% Pageviews
/home//home/index.php	1,314	18.05%
/home//home/index.php?lv1=to	1,019	14.00%

CONFIDENTIALITY

- No statistics produced that are likely to identify an individual, unless consent provided by respondent
- Agency should publish information setting out its arrangements for maintaining confidentiality of data
- When identifying data are to be given by law, they must be released under the personal responsibility of the national statistician



Box 2.1. The Four Dimensions of the GDDS

1. **The Data—Coverage, Periodicity, and Timeliness.** Dissemination of reliable, comprehensive, and timely economic, financial, and sociodemographic data is essential to the transparency of macroeconomic performance and policy.

The GDDS therefore recommends dissemination of data as described in Table 3.1.

2. **Quality.** Data quality must have a high priority. Data users must be provided with information to assess quality and quality improvements. The GDDS recommends:

- Dissemination of documentation on methodology and sources used in preparing statistics.
- Dissemination of component detail, reconciliations with related data, and statistical frameworks that support statistical cross-checks and provide assurance of reasonableness.

3. **Integrity.** To fulfill the purpose of providing the public with information, official statistics must have the confidence of their users. In turn, confidence in the statistics ultimately

becomes a matter of confidence in the objectivity and professionalism of the agency producing the statistics. Transparency of practices and procedures are key factors in creating this confidence. The GDDS therefore recommends:

- Dissemination of the terms and conditions under which official statistics are produced, including those relating to the confidentiality of individually identifiable information.
- Identification of internal government access to data before release.
- Identification of ministerial commentary on the occasion of statistical releases.
- Provision of information about revisions and advance notice of major changes in methodology.

4. **Access by the public.** Dissemination of official statistics is an essential feature of statistics as a public good. Ready and similar access by the public are principal requirements. The GDDS recommends:

- Dissemination of advance-release calendars.
- Simultaneous release to all interested parties.

Box 1.1. Key Dimensions and Elements of the SDDS

The four dimensions of the SDDS are shown in bold, with corresponding monitorable elements in italics.

The data: coverage, periodicity, and timeliness. Comprehensive economic and financial data, disseminated on a timely basis, are essential to the transparency of macroeconomic performance and policy. Countries subscribing to the SDDS are to:

- *Disseminate the prescribed categories of data with the specified periodicity and timeliness.*

Access by the public. Dissemination of official statistics is an essential feature of statistics as a public good. The SDDS calls for providing the public, including market participants, ready and equal access to the data. Countries subscribing to the SDDS are to:

- *Disseminate advance release calendars for the data.*
- *Release the data to all interested parties simultaneously.*

Integrity. To fulfill the purpose of providing the public with information, official statistics must have the confidence of their users. In turn, confidence in the statistics ultimately becomes a matter of confidence in the objectivity and professionalism of the agency producing the statistics. Transparency of its practices and procedures is a key factor in creating this confidence. The SDDS requires subscribing countries to:

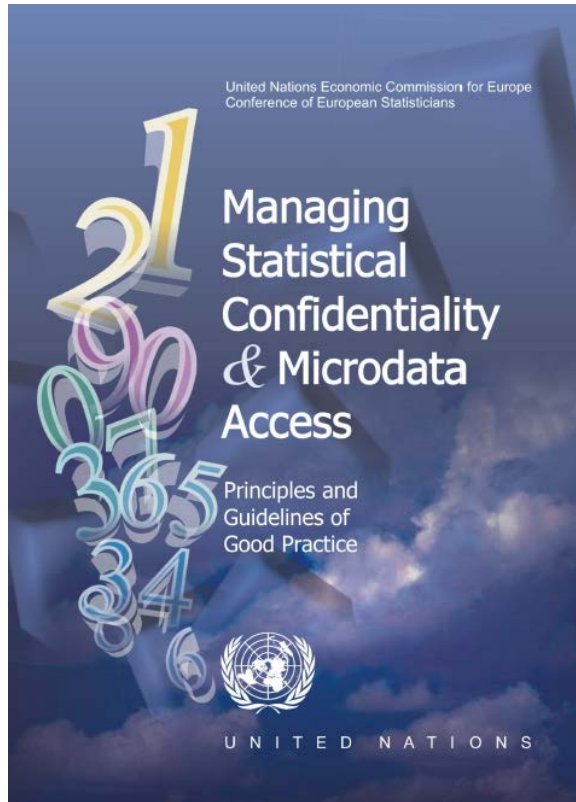
- *Disseminate the terms and conditions under which official statistics are produced, including those relating to the confidentiality of individually identifiable information.*
- *Identify internal government access to data before release.*
- *Identify ministerial commentary on the occasion of statistical releases.*
- *Provide information about revision and advance notice of major changes in methodology.*

Quality. A set of standards that deals with the coverage, periodicity, and timeliness of data must also address the quality of statistics. Although quality is difficult to judge, monitorable proxies, designed to focus on information the user needs to judge quality, can be useful. The SDDS requires subscribing countries to:

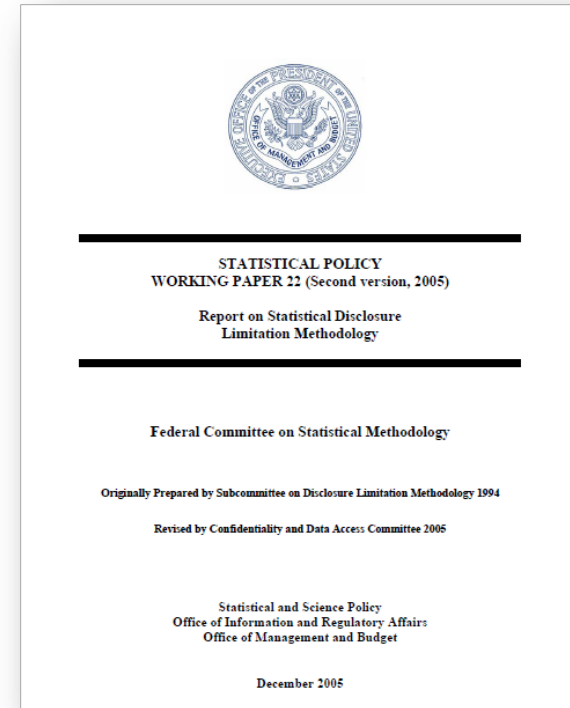
- *Disseminate documentation on methodology and sources used in preparing statistics.*
- *Disseminate component detail, reconciliations with related data, and statistical frameworks that support statistical cross-checks and provide assurance of reasonableness.*

SPECIAL ISSUE: MICRODATA DISSEMINATION

- Publish formal microdata dissemination policy and procedures (agency-level policy and dataset-specific policy)
- Provide very detailed metadata
- Anonymize datasets (no direct identifiers; reduced risk by controlling quasi-identifiers)
 - No “standard” practice
 - Common practices (e.g. USA Working Paper 22)



www.ihsn.org



Federal Committee on Statistical Methodology, **Statistical Policy Working Paper 22 (Revised 2005)- Report on Statistical Disclosure Limitation Methodology**

<http://www.fcsm.gov/working-papers/spwp22.html>

OPENNESS

- Provide easy way for users to give input and feedback
- Welcome comments, even criticism and complaints
- Respond (preferably openly) to enquiries
- Record and analyze feedback
- Data producer can also provide feedback to users, especially by commenting on erroneous interpretation and misuse of statistics.

COMMUNICATION WITH FUTURE GENERATIONS OF USERS AND STAFF

- Data are non-renewable (irreplaceable) resources. Statistical agencies must ensure their most effective use by present and future generations
- IT gives a false sense of security against loss
- A preservation policy is needed to ensure that data and metadata are preserved against hardware or software obsolescence, media failure, and other physical threats
- Preserving digital information demands constant attention



National Statistics Code of Practice

Protocol on Data Management,
Documentation and Preservation

<http://www.ons.gov.uk/about-statistics/ns-standard/cop/protocols/index.html>



Digital Preservation Management:

Implementing Short-term Strategies for Long-term Problems

www.icpsr.umich.edu/dpm/

Digital Preservation at ICPSR

enabling social science research over time.

www.icpsr.umich.edu/DP/policies/



IHSN

International Household Survey Network

www.ihsn.org