Session 3 – Dissemination platforms to make data more accessible and interpretable

**UNdata - an innovative way to provide easy access to UN System data**

Stefan Schweinfest and Mary Jane Holupka, UN Statistics Division

**Introduction**

**UNdata** is an internet-based data access system which was developed by the United Nations Statistics Division (UNSD) of the Department of Economic and Social Affairs (DESA). It offers a single entry point from an easy-to-remember URL (http://data.un.org/) through which users can search for and download a variety of statistical resources (both data and metadata) of the UN System organizations free of charge without having to subscribe or register.

One of the driving forces for the development of UNdata was the concern to overcome the fragmentation of the UN System databases. It was designed to provide users with easier access to the wealth of information collected by the various members of the UN family, and, thus, to promote its use for policy decision making and analysis. One of the platform’s strengths is its ability to pool together data from a large variety of sources. This feature responds to the increasing demand by users for the capability to retrieve and combine data from different spheres (economic, social and environmental) to measure, for instance, progress towards the Millennium Development Goals.

The paper’s primary focus is on the quality dimension of accessibility. It gives an overview of UNdata in its current state of development, highlighting the various options it offers for users to gain access to statistical information. It furthermore discusses how UNdata has the potential to contribute to the improvement of the UN System’s data with respect to other quality dimensions, such as relevance, accuracy, comparability and timeliness. Finally, the paper also points to some of the planned future developments of UNdata and the challenges ahead.

**Current status of the development of UNdata**

For over 60 years, the organizations of the UN System have been collecting statistical information from Member States on a wide variety of topics. Needless to say, it is close to impossible for a regular user to be aware of the full array of statistical information that the UN System has been compiling, or to know exactly which data series are available in which of the different data libraries of the UN System organizations. To complicate things further, the data are
typically stored in different proprietary databases, each with a different web interface (if a web interface has been developed at all) and unique access and dissemination policies, so users have to move from database to database to access the information desired. This is unfortunate since in a globalized world, users are increasingly interested in performing analyses involving cross-sectoral data.

Therefore, in order to bring the wide range of data resources of the UN System to the public, the Statistics Division, in fulfilling its mandate as coordinator of UN statistical activities, started in early 2006 to develop UNdata. The development of this database service is part of a project called "Statistics as a Public Good", whose overall objective is to increase the dissemination, use and understanding of international statistics. This involves not only providing free access to the statistics of the UN Statistics Division and the UN System, but also assisting national statistical offices of Member States to strengthen their data dissemination capabilities. The project is being implemented in partnership with Statistics Sweden and the Gapminder Foundation, with financial support from the Swedish International Development Cooperation Agency (SIDA).

UNdata is, in fact, essentially a natural extension of UNSD’s historical role in UN data dissemination. Through the Statistical Yearbook, and later on the Common Database (CDB), UNSD has for many years worked closely with the data specialists of the UN System in order to supply users with a unified platform of data from many different sources. However, the shortcomings of these modes of dissemination are obvious: They are work-intensive, subject to space constraints and allow the users little flexibility in the access to information.

The innovative design element of UNdata is the fact that data are decentralized; they remain in their “native environment” so-to-speak, close to their data owners, who retain control over, and, as a consequence, responsibility for their specialized databases. The underlying concept, thus, is one of a federated data system, which is linked through a powerful search mechanism, allowing users to search and access UN datasets included in UNdata in a variety of simple ways.

Currently, UNdata contains more than 55 million data points from over 15 data domains, with more to be added. UNdata is being developed in a phased approach:

- Phase 1 covered the range of statistics compiled by UNSD, i.e. Key Global Indicators (formerly called the Common Database (CDB)), MDGs, Energy, Gender, Industry, National Accounts, Population and Trade. Selected UN Population Division (UNPD) data have also been included in this stage. Phase 1 has been completed and the UNdata team is working with the various UNSD and UNPD data owners on the first round of new updates;

- Phase 2 covers data from UN specialized agencies, programmes and funds (FAO, ILO, ITU, UNDP, UNESCO, UNFCCC, UNHCR, UNICEF, UNWTO, WHO so far). This phase is ongoing, as data providers continue to add series to those initially supplied and new partners join the platform.

It is important to stress that UNdata also provides access to all accompanying metadata. This is not only limited to “cell specific” micro metadata, such as footnotes, but extends to all relevant supporting information, most important of which is the attribution and reference to the exact data.
source. A UNdata feature entitled “Wiki” contains information on relevant glossaries, guidelines, definitions, methodologies, classifications, etc. We consider Wiki to be an essential quality feature of UNdata, as it provides users with the necessary tools to understand, interpret and assess the data with respect to “fitness” for their particular intended use. This is all the more important as UNdata, by design, has the capability to juxtapose seemingly identical data series from different data sources.

Assessment of quality - accessibility

This section briefly describes how UNdata facilitates access to information and provides the user with a number of simple tools to interpret and use the data.

Getting in:
http://data.un.org

A single integrated entry point to many data sources and datasets

UNdata’s entry point is a clear and relatively simple screen, which follows the best practices and principles of web 2.0 design, offering a Google-like search for data, with a search bar prominently placed for users to start from there.
The system is designed to respond to a variety of searches, anticipating different information needs and different degrees of sophistication of users.

For instance, entering simply a country name (see screenshot below) returns a map and an abridged country profile from the *World Statistics Pocketbook* (with a link to the full profile). Furthermore in the center field, links to all the data series available for this country are offered. Not surprisingly, this rather general search criterion yielded 576 results involving all 15 databases currently contained in UNdata. The system also provides a link to the country’s national statistical office website on the right hand of the screen under the heading *related links*.

There are a priori no limitations on what type of criteria may be entered in the search bar. Instead of entering a single country name as in the example above, a user may enter a single keyword (e.g. population) or a more sophisticated string of criteria (e.g. one or more country names, and/or topics and/or years). The results are generally returned in the same layout as above (without the country profile box), namely source databases on the left, data series names in the center, and related links on the right side of the screen. As with general web-search tools, the more specific the criteria entered, the fewer the number of matches found. From the first round of feedback received, we found that users characterized this type of access as “intuitive” and quickly learned to refine their search criteria, in order to access the desired specific data items.

It is noteworthy that, for those users who have more specialized-sector specific needs, the system provides for two short cuts that point the user to the relevant sector datasets. First, the UNdata entry page (see page 3) offers, as an alternative access route to the data, a listing of the various underlying datasets, which are clickable, and allow the user to further explore or drill down to
the series level. Second, even if a user has chosen the search bar as the main access route to the data, an option to restrict the search to one or more specific datasets is provided through the application of filters (see screenshot page 4, left-hand panel).

Once the user has selected a particular data series, the screenshot below shows the standard data display screen. This default display may be changed in a number of ways, to suit the user’s preferences. Data can be sorted and filtered, columns can be pivoted and relevant metadata can be directly accessed. A customized view can be saved and bookmarked, and the data can be downloaded in several formats.
As briefly demonstrated above, UNdata eliminates many of the obstacles to accessibility that users face when looking for data, since it:

- is available free of charge, with no need to subscribe or register;
- provides one interface from which all underlying datasets can be searched and browsed easily and quickly;
- is easily searchable by either using the search box or clicking on a particular dataset name and browsing its contents;
- requires few clicks to get where one wants to go;
- presents data clearly, in simple tabular presentation formats, which can then be customized and further refined by sorting, filtering, including/excluding columns and pivoting;
- links to or provides relevant accompanying metadata;
- provides various download formats (xml or value separated for data received in relational database format, or excel tables if the original data were provided in that format).

Not surprisingly, the most repeated request from the users who have sent in feedback, is for more: more data from the current providers, new data from additional data providers and more up-to-date data. In order to continuously improve the accessibility dimension, work is underway to increase UNdata’s database coverage with several additional UN agency data providers in the pipeline for contributing their data series to the portal.
One particular “accessibility challenge” in the context of an international database is multilingualism. We would anticipate that through the provision of multilingual access tools, the use of the system could be significantly increased and outreach to certain user-groups enhanced, especially in specific regions of the world. Whilst it does not seem to be realistic to aim at making the entire retrieval system multilingual, UNSD is currently exploring how to provide at least limited multilingual access features.

Other quality dimensions – relevance, accuracy, comparability and timeliness

UNdata’s decentralized approach strengthens the role of the data producers/owners, and leaves it up to them to provide their most relevant data series, since they are in the best position to know which of their data series are in greatest demand and considered to be most important to their users, and of the highest quality, in terms of accuracy.

In addition to the above external type of relevance check, a system-internal check exists as well. UNdata’s administration application feature (coupled with the use of Google Analytics) allows for rather detailed analysis of user behaviour. This provides the UNdata team with valuable information on whether users are actually finding the data series they are interested in. Information on most searched series, time spent on particular pages and searches returning zero results, as well as the direct feedback received from the users, provide valuable pointers for the team, in terms of which direction to take in order to continuously enhance the relevance of the data provided.
Giving users the capability to easily search across different datasets, and link to the dataset providers’ own online data on their websites, is expected to lead to greater scrutiny by users in terms of data accuracy and comparability. Finding different data in different sources referring to seemingly similar data series will inevitably lead to questions and possibly some confusion. However, this juxtaposition of data series has the potential to create the necessary positive “market pressure” towards the elimination of duplication or discrepancies that do not have a substantive basis. In this context, UNdata’s Wiki feature plays a key role, as it provides a platform to present data definitions, methodological explanations and other metadata information which should shed light on why data could legitimately be different, for instance due to definitional or methodological reasons. It is hoped that the need to address user questions will provide data owners with an additional incentive to cooperate in order to increase the comparability of data from different data sources, in terms of content and presentation (e.g. country names/classifications).

The timely incorporation of new data into the UNdata portal will remain one of the key challenges in order to keep users satisfied. UNSD is working closely with its partners in order to manage the electronic data transmission as effectively as possible. The development and implementation of SDMX in an increasing number of international databases will certainly facilitate this task in the future. Furthermore, whilst UNdata in its current version only contains annual data, UNSD is currently exploring how quarterly and monthly data can be added to the system in an effort to make more timely information available.

**Challenges ahead**

There are certainly numerous ways in which UNdata can be extended, some of which have already been mentioned in the text above. In the following, we would like to highlight some envisaged future developments, which are of course mainly motivated by the intent to improve the quality dimensions discussed above.

Several initiatives are under way to continuously add to the data and metadata content: In addition to the regular updates that will keep UNdata “fresh”, we are encouraging our data partners to review their databases in order to determine whether more data series could be made accessible through UNdata. In general, partners are invited to provide information up to a “reasonable” level of subject-matter breakdown. Highly disaggregated data, which are likely to be the object of interest only for very specialized users within a particular domain (e.g. trade data), are considered to be better left to a subject-matter specific dissemination context (such as Comtrade). As an element of quality control in terms of accuracy, partner agencies normally provide those datasets that have gone through a solid internal vetting process, such as the ones earmarked for paper and/or web publication.

One particular kind of extension of the system refers to the possible inclusion of country databases. Technically the system is capable of accommodating national datasets. Juxtaposing national and international datasets in one easily accessible framework does seem to have a number of attractive features, such as bringing even more data volume to the fingertips of users, potentially accelerating the international delivery of national data in a significant manner and...
empowering the users to raise the level of scrutiny with respect to perceived discrepancies between national and international data. However, a number of practical challenges would need to be addressed, not the least of which is that there is no such thing as “one national dataset”. UNSD is currently exploring the possibility of experimenting with datasets where a certain degree of harmonization in content and presentation has been achieved, as is the case with the UN DevInfo datasets, which contain major national development indicators.

With respect to enhancing the functionalities of the UNdata system, one of the key areas is to offer an effective linkage with attractive visualization tools. In order to better convey the “underlying messages” of the data, advanced visualization tools, such as graphs, charts, maps and dynamic bubble charts of the type developed by Gapminder, are required. Making such visualization tools an integral part of UNdata, however, would go beyond the scope of the system. Nonetheless, offering features that would allow users to extract datasets in pre-specified formats, which could then feed easily into standard visualization tools, will certainly be appreciated by the user community.

Another functionality that requires attention is the continuous fine-tuning of the search algorithm to ensure user-friendly priority ranking of the matches encountered. The current system functions on the basis of some pre-defined selection criteria. Over time, the search and selection history defined by the users will add useful criteria.

Some of the above-mentioned possible improvements are certainly ambitious sub-projects in their own right, and it will take considerable time to move forward on them. However, the authors are confident that with the necessary commitment, enthusiasm and support of the partnership network, the exciting potential of UNdata can be developed much further. Past and expected future user feedback will thereby be a helpful guide in setting appropriate priorities in terms of system development.