"WORLD IN FIGURES": A STATISTICAL DATA DISSEMINATION SERVICE VIA THE NET, WITH AN ONLINE DATA ANALYZER [1]

Andras Vag, Heureka Research, Italy
Eva Hideg, Statistics Hungary

1 SUMMARY

"World in Figures" (WiF) is an online system of statistical databases, extended with a couple of basic and advanced analyzing tools. This online knowledge-base with direct access to a great amount of numerical data is a gateway to an integrated statistical database containing global, country and sub-country level information. The instruments and techniques of WiF help the user to discover associations between variables and manage "what-if" type questions related to socio-economic development, environment, governance and human behaviour. With its wide scope, the project provides a solid background to describe possible future alternatives.

The innovation of WiF can be described as (1) wide range of data about different fields of life; (2) thousands of variables and millions of records in a single database; (3) immediate access to results; (4) online statistical modules; and (5) advanced output to show or download the results. Apart from the data access, it offers modelling options, too. The online analyzer has basic and advanced functions, contains data presentation and visualization modules, forecasting functions, downloading options, etc. Data access is free, except for some research and business statistics. Moreover: the user can find specific time series and the methodological descriptions of the variables as well. WiF helps researchers and analysts in the field of science, education or business, with a new and effective online service, which facilitates the understanding of human life, accelerates the discovery of hidden causal relationships etc. WiF shows signs of integrating the „traditional" variable-based analyses with new modelling tools, like multi-agent-models or chaos models. Additionally, it creates one possible shape of the „floating" modelling philosophy.

This paper has four main parts: (1) introduction to World in Figures, (2) some words about world-modelling, (3) the picture of floating modelling philosophy and finally (4) the summary of the option developing World in Figures to an effective floating modelling tool.

[1] Supported by the Hungarian Scientific Research Fund (T35070 sz. OTKA program)

2 WHAT IS WORLD IN FIGURES?

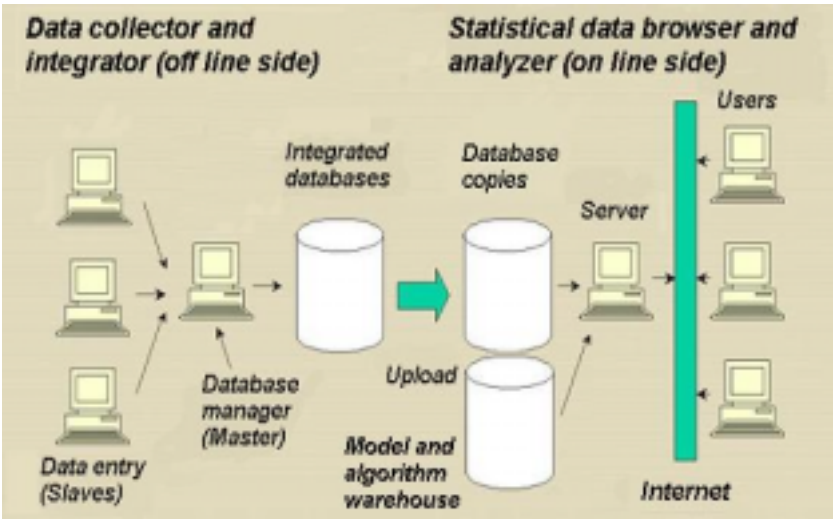
2.1 The concept

World in Figures (WiF) is an online knowledge–base with direct access to a great amount of statistical data. The data dissemination service is extended with a flexible statistical data analyzer and a basic online modelling tool. The project's portal[2] is a gateway to an integrated statistical database containing global, country and sub–country level data. WiF project has five objectives: (1) integrating statistical databases; (2) providing web-access to the project's integrated databases; (3) providing web-access to a variable–based statistical data analyzer; and (4) developing a new world-modelling tool, the details of which are below.

2.2 WiF's architecture

WiF's infrastructure has to main parts: the data collector and integrator (off-line side) on the left of the picture, and the statistical data browser and analyser on the right of the picture (on-line side), which is extended with a model and algorithm warehouse. The architecture of WiF is shown in Image 1.

Image 1.
The architecture of WiF



The two main parts of the architecture are described in the following paragraphs.

2.3 Statistical data collection and integration

WiF works with off-line stored statistical data (databases, tables, time-series) and not with on-line interlinks of databases. Data integration is solved with an intelligent tool, having two parts, called „Master" and „Slave". These programs automatically combine statistical tables and databases, although the contribution of professionals is necessary. The result of the integration process is a single database, which can be further manipulated. The fields of the

[2] www.worldinfigures.org

databases contain all relevant information for further processing, like value, variable name, time period, territory, data source and comment.

The main functions of the Slave side of statistical data integrator software are the following: (1) importing statistical tables from different source files (XLS, CSV, HTML, TXT, PDF) or from the clipboard and from source databases (MDB); (2) generating Slave database; (3) merging input and Slave databases, (4) editing headers and other fields, (5) facilitating the merge of source records and the Slave database; (6) setting some important sub-functions, like managing data conflicts (e.g Source Priority), handling different wordings of the same fields from different sources, treating "Multiple Variables" to accelerate the merging process of different variable structures, etc.; (7) sending the integrated database to the Master. The input capacity of Slave is 20-100 thousand of records/day (depending on table-complexity) from tables and millions of records from databases. (see Image 1.)

The main functions of the Master are: (1) managing Master databases; (2) importing databases prepared by Slaves; (3) merging imported Slave databases and the Master database; (4) adjusting Master database structure and editing Master database content; (5) slave administration; (6) submitting Master database to the server, (7) filtering sub-databases from the Master database. (see Image 1.)

2.4 A web-based statistical data source

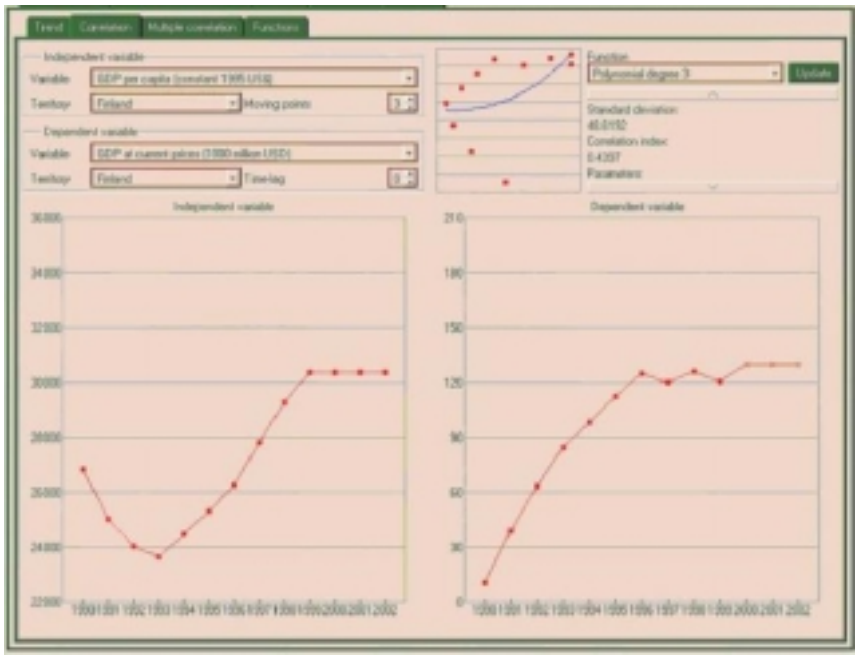
WiF is a comprehensive data source accessible via the project's portal (www.worldinfigures.org). Since there is practically no memory limit to store the information, the database management concept is to collect as much information as possible in a single database to facilitate researchers' work (e.g. to compute correlation within a minutes between the selected variables). Hence the main database of WiF (the global database) contains: economic, social, environmental, ecological, political, opinion, public policy, behavioral and many other data. Beside the global database, WiF has other databases too (e.g. country and sub-country databases), which contain the same types of statistical variables. The time-interval is one year, but any other time-interval (quarter, month, day) can also be treated easily. The database content types are the well-known statistical dimensions: frequencies, volumes, values, percentages, etc.

The user may search and select statistical variables, territories and periods with the help of the statistical data browser. This is an advanced Java application, which runs within the client's browser. The program offers keyword search, directory search and some other options to accelerate the execution of data queries. The results are presented in a very flexible downloadable 3D tables and diagrams. The statistical data browser also serves as a variable selection tool for the analyzing and modelling modules.

2.5 WiF's online statistical data analyser

The online data analyzer of WiF is a Java based application. It is integrated with the statistical data browser and appears in a separated tab of the program. The analyzer is built-up by modules, which mainly include functions for analyzing time-series. It is impossible to give a description of the analyzer in this article. To illustrate the functionality of the online statistical data analyzer, a dialog panel of it is shown in Image 2.

Image 2.
A dialog window of WiF's statistical data analyzer



3 WiF's BENCHMARKS

WiF can be benchmarked in three fields: (1) the online statistical data sources; (2) the online data analysing tools and (3) the online world-models.

3.1 Statistical data on the web

One may find four categories of direct statistical data sources on the web: (1) global data sources (for example UN, World Bank and OECD); (2) European statistical data sources (for example Eurostat); (3) national statistical agencies; (4) vertical statistical data sources (e.g. UN organizations); and others, like the sources of alternative indicators (e.g. Redefining Progress). Taking into account all global and regional horizontal portals, the global vertical portals, the country level horizontal (statistical agencies) and vertical portals, the total number of statistical portals providing valuable, highly aggregated data is over 200. The number of freely accessible, highly aggregated, country-level statistical data per year can be estimated to be around 5-15 x 10^6 records (overlaps excluded).

WiF contains more than 2 millions records, which qualifies it as one of the biggest direct access online data source. Compared to the other free access statistical data disseminators, WiF's database is rather complex. As it was mentionned above its scope is very wide, which is realy unique on the web. The relative weakness of WiF's databases - because of global statistical data dissemination policies - is that statistical data in many sectors are not updated.

3.2 Online data analysers

There is a significant number of online accessible and free data analyzers, most of them serve educational purposes. The online accessible analysing tools have two types: (1) downloadable desktop data-analyzers; and (2) online running data-analyzers The most popular

downloadable desktop analysers are Dataplot[3], Instat Plus[4], Scilab[5], Vista 5.5[6], Mx[7], and R[8]. The number of online running analysers is lower. Three of those worth to mention: WebStat 3.0[9], Rweb[10], and Statiscope[11].

Almost all of the a above listed analyzers are specialized to a specific type of problems. Similarly to those, the analyzer of WiF is also focusing on a specific approach, especially the analysis of time-series data. Additionally to this WiF's analyzer has some innivative graphical solution to facilitate the work with figures and charts.

3.3 World-models on the web

Comparing world-models, more precisely numerical world models to statistical data analyzers one may find significant differences and similarities. The description below provides a very short introduction to the world-models.

There is no explicitely elaborated definition of "world modelling". We can lean on the conventional word–usage. These provide a sufficient basis to see "world models" as simulating tools of multiple– and humanity–related phenomena. For a first sight, for a model to be "global" in scope it must be meant to characterize some features of human thought, behavior and the human environment that are held to be typical if not universal, across cultures and history. Specifically, it must represent these features in logical and mathematical forms amenable to describing dynamics. It must offer a causal explanation for the interactions modeled. It must be "testable" in the sense of empirical investigation. Finally, it must be intended to be useful in practical application, that is, have variables or parameters that a user of the model can identify as factors amenable to being manipulated to produce desired effects in some aspects of the world being modeled. Another definition is Brecke's, who suggests the features of global models to be: geographically global in scope, long duration (25-50 years), and integrative of diverse sectors, like population dynamics, economic dynamics, politics and the environment (Brecke 1995). Chadwick focuses on scientific universality, logical and mathematical construction, explanatory power, testability, and practical, applied usefulness as characteristics of this type of simulation, and left the geographic and sectoral scope, and time-horizon as issues to be addressed rather than as criteria for exclusion. (Chadwick)

The first period of computer–based world–modelling lasted from the start of the seventieth to the mid of the eightieth of the last century. The results of this glorious decade are well known among professionals and modelers, but the large audience of the 21st century is not aware of the findings and details of these models. The first dynamic world–model was developed by Jay Forrester, based on his previous efforts on system dynamics, industrial dynamics and

---

[3] www.itl.nist.gov/div898/software/dataplot/
[4] www.rdg.ac.uk/ssc/instat/instat.html
[5] http://www-rocq.inria.fr/scilab/
[6] http://forrest.psych.unc.edu/research/
[7] http://views.vcu.edu/mx/
[8] www.r-project.org
[9] http:/webstatsoftware.com
[10] http://www.math.montana.edu/jeff-bin/Rweb1.03/
[11] http://www.df.lth.se/~mikaelb/statiscope/statiscope.shtml

urban dynamics at MIT (Massachusetts Institute of Technology). The origin of system dynamics was the so called “feedback control mechanism” researched in the “Servomechanism Laboratory” during the World War II. Forrester took part in this early work. The language of system dynamics was DYNAMO, first published in 1959. WORLD2 as a real, system-dynamics based, computerized world-model mapped associations between world population, resources, pollution, industrial production and food. (Forrester 1971)

The well known publication of the Club of Rome, titled “The Limits to Growth” (Meadows et al, 1972) contained the results of the simulation of the world using the system dynamics of WORLD3. The project was led by Dennis Meadows, a former Ph.D. student of Forrester. This world-models drew the attention of the public to the so called world problems. The spreading of system dynamics as a modelling concept and the increase of the number of functions and variables within the models were enabled by the rise of computer technology. Nevertheless, the knowledge base of the first world-models were really small, compared to those of the end of the century. First world-models contain the sprout of interactivity, by allowing developers to adjust initial parameters of the models, as means to run different scenarios. In this period, world-models were accessible neither by the research community nor by the large audience. The last decade was determined by the worldwide spreading of personal computers and the option of running world–models in PCs. Parallel to this, models became more and more advanced, new modules were introduced and regionally detailed. The research interest center of gravity moved away from economy towards environmental and climatic issues.

Today’s world models are colorful and methodologically innovative. The traditional modelling philosophy is extended with additional concepts, higher level of integration and new application packages. Meanwhile, the scope of simulated phenomena have also been widened or changed, for example by linking micro- and macrolevels of societies and economies. Mainstream macroeconomic modelling also became widely used in world-modelling. Nowadays, world-models are running on desktop computers, some of them are downloadable from the net, some others – not the complicated ones - are running online. A good indicator of the methodological innovation is the appearance of multi-agent-models, which do not include data (e.g. time-series) and the model-structures are set up by the user. Anyway, the methodological innovations and the accelerating computer processors lead to a brand new world-model philosophies.

The number of ever built and published world-models is around 80 - 100. There’s no room to indicate all important types of world–models in this paper. I rather focus on human behavior–environment related (economy, society, politics, demography, energy, pollution, etc) global or regional world–models. The best examples for updated world-models or new models based on traditional philosophies are International Futures and Globesight. Both deal with different aspects of the country-behaviors and attributes, like economy, environment, energy, natural resources, demography with the good old system dynamics approach.

International Futures was the first global model that could run on PC. This model still had no regional or country differentiation. The next generation became a full–scale PC world–model. It also improved the representations of energy, and food systems, demographic, and new socio–political and environmental content were added. Forecasts are produced by time–series and functions. Subjects are analyzed at global, regional and national level. IF can not be used

for demonstrating complex adaptive systems, since there is no emergence and no generative modelling included. Another concern is that the multitude of parameters and variables can lead to difficulties in separating out what factors drive the outcomes. This can cause problems in understanding policy projections. The late version of IF has the following detailed sub–modules: economy, agricultural energy, socio–politics, international politics, environment and implicit technology. The student version of IF is downloadable from the net.

For the end of the twentieth century models became highly integrated and extended to analyze new, emergent phenomena. These models put great emphasis on scenarios, which can be treated as parts of the methodological frameworks of each world-model. At that point, the pure methodology closely interlinks with both the model concepts and expected results. Compared to the first world–models, today's macroeconomic and world–models are significantly more flexible. New approaches and methods appear, advanced algorithms are used and the knowledge bases are notably more extensive. Additionally, the interest of the researchers have also extended: ecology, environment, climate change, politics became integrated part of the world–models. Parallel to these innovations, new simulation packages and other tools have been released to facilitate world–modelling and forecasting. Moreover, we can observe the spreading of multi–agent models (which help simulating emergent phenomena and interlink the micro and macro levels) and the appearance of free access models and online applications. Although the number of widely used packages is much higher, I can only introduce four packages: IMAGE 2.2, CORMAS and NetLogo.

IMAGE 2.2 (Integrated Model to Assess the Global Environment) or rather its antecedent was presented in 1994. IMAGE 2.2 is a multi–disciplinary, integrated model designed to simulate the dynamics of the global society– biosphere– and climate system[12]. The objectives of the model are (1) to investigate linkages and feedbacks in the system, and (2) to evaluate consequences of climate policies. Dynamic calculations are performed to the year 2100, with a spatial scale ranging from grid (0.5 x 0.5 degrees latitude–longitude) to world and regional level, depending on the sub–model. The model consists of three fully linked sub–systems: (1) Energy–Industry, (2) Terrestrial Environment, and (3) Atmosphere–Ocean. Four standard scenarios are included for selected aspects of the society– biosphere– climate system including primary energy consumption, emissions of various greenhouse gases, atmospheric concentrations of gases, temperature, precipitation, land cover and other indicators. These scenarios, computed with IMAGE 2.0, are presented for selected aspects of the society–biosphere–climate system including primary energy consumption, emissions of various greenhouse gases, atmospheric concentrations of gases, temperature, precipitation, land cover and other indicators. (Alcamo 1994).

CORMAS is devoted to the applied modelling of the relationships between societies and their environment. The developer team is part of the "Renewable Resources and Viability" program (TERA department) of CIRAD, France. The project is busy in developing multi–agent systems about integrated natural resources management. The computer tool, which is an agent–based simulation framework can be downloaded from the CORMAS Internet portal[13]

[12] http://arch.rivm.nl/image/
[13] http://cormas.cirad.fr/indexeng.htm

free of charge. The user will also be able to access application examples (with a models library) and publications.

NetLogo is a programmable modelling environment for simulating natural and social phenomena, developed at the Northwestern University, USA. It is particularly well–suited for modelling complex systems developing over time. Modelers can give instructions to hundreds or thousands of independent "agents" all operating in parallel. This makes it possible to explore the connection between the micro–level behavior of individuals and the macro–level patterns that emerge from the interaction of many individuals. NetLogo lets students open simulations and "play" with them, exploring their behavior under various conditions. It is also an "authoring tool" which enables students, teachers and curriculum developers to create their own models. NetLogo is simple enough that students and teachers can easily run simulations or even build their own. And, it is advanced enough to serve as a powerful tool for researchers in many fields. (Wilensky 1999) NetLogo is downloadable from the project's portal[14].

4 THE "FLOATING" WORLD-MODELLING PHILOSOPHY

4.1 Conceptual framework

Floating modelling philosophy[15] summarizes the most significant upgrades compared to the preceding world–models. The basic notion is that the structure, the functions and the equations of most of the traditional world–models were "frozen" to the programs. All runs gave the same results and the users could only study the built–in processes. The evolving "floating" world-modelling concept is derived from the simple extrapolation of the current world-modelling trends. Meanwhile, this philosophy contains some emergent modelling features too. The pillars of the trends are (1) the dramatically developing computer capabilities; (2) the accelerated information exchange via the Internet; and newly introduced algorithms and simulation models. On the basis of these phenomena (3) models are becoming more complex and more integrated; (4) access to data and to models are much more extended than previously; and (5) finally model development and model usage can be interlinked to one process. Consequently world-modelling methods of the 21st century will provide new tools for the humanity to reflect the potential futures.

4.2 Differences compared to other world-models

As I indicated the tendeny above, world-modelling are becoming relatively independent from modelling philosophy barriers and their complexity exceeds earlier limits. Next step, floating world-modelling approach describes, is the further improved independence from methodological boundaries. That means modeller will be able to choose the best fitting modelling philosophy to the simulate a problem. Conceptual-, algorithmic- or computer speed limits of the modelling tools determine modellers creativity when they construct models with the intention of simulating real-world problems. This limits are stretching and the modellers may choose more and more flexibly from algorithms, modules and approaches. They may build their own models, because structures are not anymore burned-in the models. This does

---

[14] http://ccl.northwestern.edu/netlogo/

[15] The word "floating" is adopted from Karl Mannheim's "free-floating intelligentsia" concept (Mannheim, 1991).

not mean that the user only has this option. It is possible to open, modify or extend existing models, imported from a models library.

Nowadays, the term “complexity” is one of the most popular expressions in science. In our context its meaning, for the first sight, is relatively simple: future world–models will work with much more data than the preceding ones and the inventory of methods will also be more advanced. The result will be an increased level of model complexity. Emphasis will be put on the in–depth analyses of sectorial and/or regional problems, where approaches of different disciplines (e.g. ecology, socio–economic research, climate research) will be managed in joint projects.

4.3 New features

Since the 1980s and the 1990s, the new methodological philosophies and applications (for example non–linear modelling approaches, analysis of dynamic phenomena, chaos researches, multi–agent methods and other evolutionary algorithms, pattern recognition and event data analysis) have dramatically improving importance. In short we are the wittnesse of the rise of the employment of the tools of artificial intelligence in world-modelling. These methodological concepts and algorithms have turned into practice in social sciences since the mid of the 1990s and it is plausible that they will be widely used in future world–models. The “new” methods will be integrated with “traditional” ones (like system dynamics, differential equations, time–series analysis, multivariate and clustering methods, etc) to advanced software packages. These kinds of integration will contribute both to the improved forecasting value and the increased accuracy of simulation processes.

As it was mentioned, analytical and simulation tools can be flexibly selected, depending on the problem, which means that the modelling tool implies high level of interactivity. This option provide further opportunities and access to the results by the large audience. Interactivity will facilitate the permanent project development, data supply and discussions, which may have political implications as well. So, the “floating” philosophy can be seen as a new synthesis of the above listed trends.

System dynamics and equilibrium models have lost their privilege in model building and meanwhile the relative importance of GIS (geographical information systems) and MA (multi-agent) is growing. This methodological shift is accompanied with the multiple usage or integration of the methods. The methodological integration means the appearance of different modelling principles within one model. This kind of integration is increasingly popular. For example multi-agent models are integrated with system dynamics, with equilibrium models, GIS applications and time-series analysis. In the same time the growing number of ASPs (Application Service Providers) indicates the explicit need for online data access, data processing and analysis. Additionally, the fast downloading options let researchers, decisionmakers and even the citizen easily access much more complex world–models and modelling ressults, than those used 10 years ago.

4.4 Extended knowledge-base

The keyword in knowledge-base extension is statistics or rather enormous statistical information. Data owners, public organizations and researchers have realized the need of integrating statistical databases. For example, a great number of EC funds indicate this

expectation of the European societies. From a historical perspective, the emphasis of the early world–models was on economy. A bit later, the importance of politics, sociological factors and human issues became more and more significant. For now models are further extended with ecological, climatic and geographical information. In the future we can count on the expansion of the need for both the horizontal and vertical data. This will be expressed by the radical increase of accessible statistical (or rather numerical) information and by the usage of decreasing aggregation level statistical data. Both of these factors underline the need for huge integrated databases.

5 WIF'S UPGRADE TO A FLOATING WORLD-MODEL

World in Figures aims at upgrading its system to a floating world-model. This process is a continuous and complex task. The current version, as described above, is a statistical variable-based solution, but for now it has some modules which are frequently used by chaos models. The project is on the way to interlink WiF's variables or time-series and a multi-agent model, containing GIS options too. This will be a real floating world-model, since the researcher can build a model for demonstration and education purposes or for the simulation of any real-world phenomenon. Data feeding can be solved either by a direct upload from the client side or by queries sent from the WiF's statistical data browser.

6 REFERENCES

Alcamo, J.R. (ed.) (1994): Image 2.0. Integrated Modelling of Global Climate Change. Dordrecht: Kluwer Academic Publishers.Brecke 1995

Brecke, P: http://www.inta.gatech.edu/peter/globmod.html

Chadwick, R.W. (2000): Global Modelling: Origins. Assessment and Alternative Futures. Simulation & Gaming. Vol. 31. No. 1: 50–73.Cormas homepage

Forrester. J.W. (1971): World Dynamics. Cambridge MA: MIT Press

Meadows, D.H et al. (1972): The limits to growth. London and Sidney: Pan Books.

Mannheim, K. (1991): Ideology and Utopia. London: Routledge.

Wilensky, U. (1999): NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer–Based Modelling. Evanston IL: Northwestern University.