JISC

Project Document Cover Sheet

Project Information			
Project Acronym	DataShare		
Project Title	DISC-UK DataShare		
Start Date	March 2007	**End Date**	March 2009
Lead Institution	EDINA, University of Edinburgh		
Project Director	Peter Burnhill, Mark Brown		
Project Manager & contact details	Robin Rice, Edinburgh University Data Library, Main Library Bldg., George Square, Edinburgh EH8 9LJ R.Rice@ed.ac.uk, 0131 651 1431		
Partner Institutions	Universities of Edinburgh, Oxford, Southampton and London School of Economics (affiliate)		
Project Web URL	http://www.disc-uk.org/datashare/		
Programme Name (and number)	*JISC Repositories and Preservation Programme: Repositories Start-up and Enhancement projects strand*		
Programme Manager	Andrew McGregor		

Document Name			
Document Title	Data Visualisation Tools: Part 1 – Numeric Data in a Web 2.0 Environment		
Author(s) & project role	Stuart Macdonald, University of Edinburgh, Project Officer		
Date	17 December 2007	**Filename**	Numeric_data_mashup.pdf
URL	http://www.disc-uk.org/publications.html		
Access	☐ Project and JISC internal	■ General dissemination	

Document History		
Version	**Date**	**Comments**
1	14 January 2008	
1.1	8 September 2008	Title and header made consistent with Part 2.

Introduction

Data visualisation has been defined as:

> *The set of techniques used to turn a set of data into visual insight. It aims to give the data a meaningful representation by exploiting the powerful discerning capabilities of the human eye.*[1]

Part 1 of this briefing paper will highlight some examples of new collaborative web services using Web 2.0 technologies which venture into the numeric data visualisation arena. These *mashups* allow researchers to upload and analyse their own data in 'open' and dynamic environments. Broadly speaking the numeric data being referred to could be micro-data (data about the individual), macro-data[2] or country-level data, derived or summary data.

Part 2 will investigate and showcase examples of spatial (or geographic) data mashups using Web 2.0 technologies and how they can be utilised in a research environment.

This paper does not intend to conduct an investigation into the definitive merits of each utility but rather compare the functionality, 'openness' and usability of such utilities from the perspective of a researcher willing to share or analyse their data.

A word of warning - researchers will have to account for the inconstant nature of the web - resources such as those described above may not be around in two, five or ten years. Not only will there be further advances in web technologies but services merge, are bought out or indeed cease to exist. Services that start off open or free may become 'closed'. Resources may start up with a particular rationale but may evolve into a completely different service or resource. Thus it would be unwise for users to deposit the only copy of a dataset within such a utility, i.e. use it as an archive or trusted repository. Certain data held within such utilities may often lack any authority or obvious provenance. Indeed data deposited may well be open to edits and uses unforeseen by the depositor. Many issues regarding the legal, cultural and technical aspects surrounding the sharing and publishing of data produced from personal or academic research abound. This paper does not purport to find solutions but highlights possibilities!

Utilities covered

There is a range of proprietary and academic domain-specific data visualisation tools, such as Nesstar (social science), CORDA Centerview (business intelligence), GGobi (high-dimensional data visualisation), Crystal Xcelsius (business data), Ferret (oceanography / meteorology), Giovanni (earth sciences), to name but a few. There are also *Open Source* utilities such as the Prefuse visualisation toolkit which is a set of tools for creating interactive data models, visualisations and animations.

The data visualisation tools described in this briefing paper can be regarded as *Open Data* utilities in that they retain the ethos of other 'Open' initiatives such as *Open Source* and *Open Access* by allowing anyone to upload or use the data. They are generally commercial services embracing a Web 2.0 business model in which users pay to access additional features such as keeping data private. The tools described here have emerged on the World Wide Web in the last two years, namely: Data360, Many Eyes, Swivel, Gapminder, StatCrunch, Graphwise.

Note: Gapminder World has been included in this paper because the interactive data features are to be included in the forthcoming version of the utility.

---

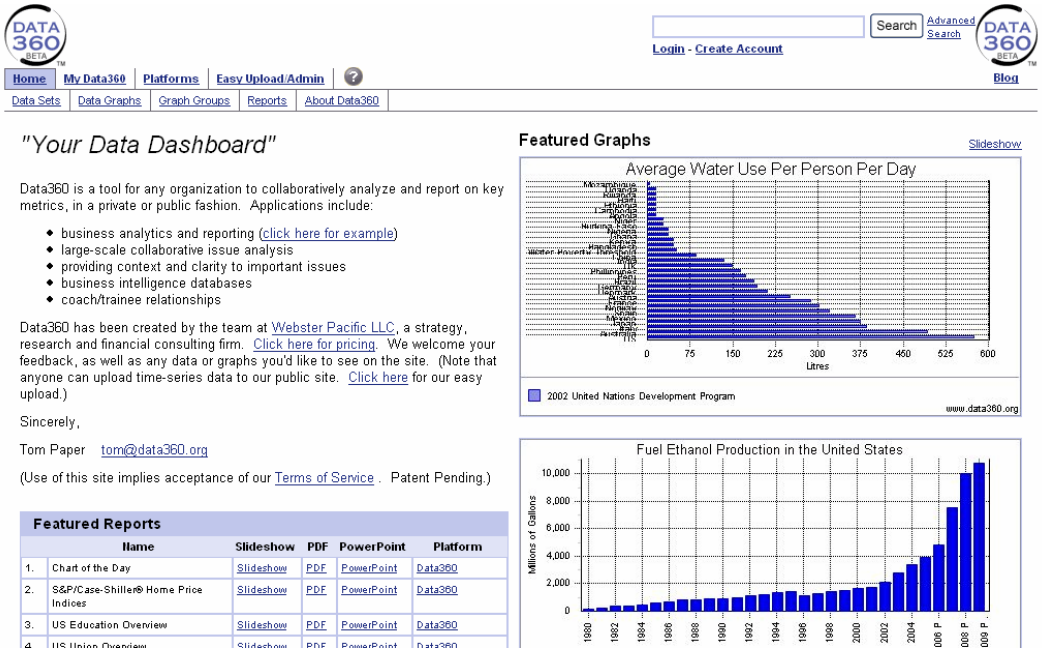
[1] Edinburgh Online Graphics Dictionary
http://homepages.inf.ed.ac.uk/rbf/GRDICT/grdict.htm

[2] See Appendix 1 for visualisation tools for macro-level or country-level data

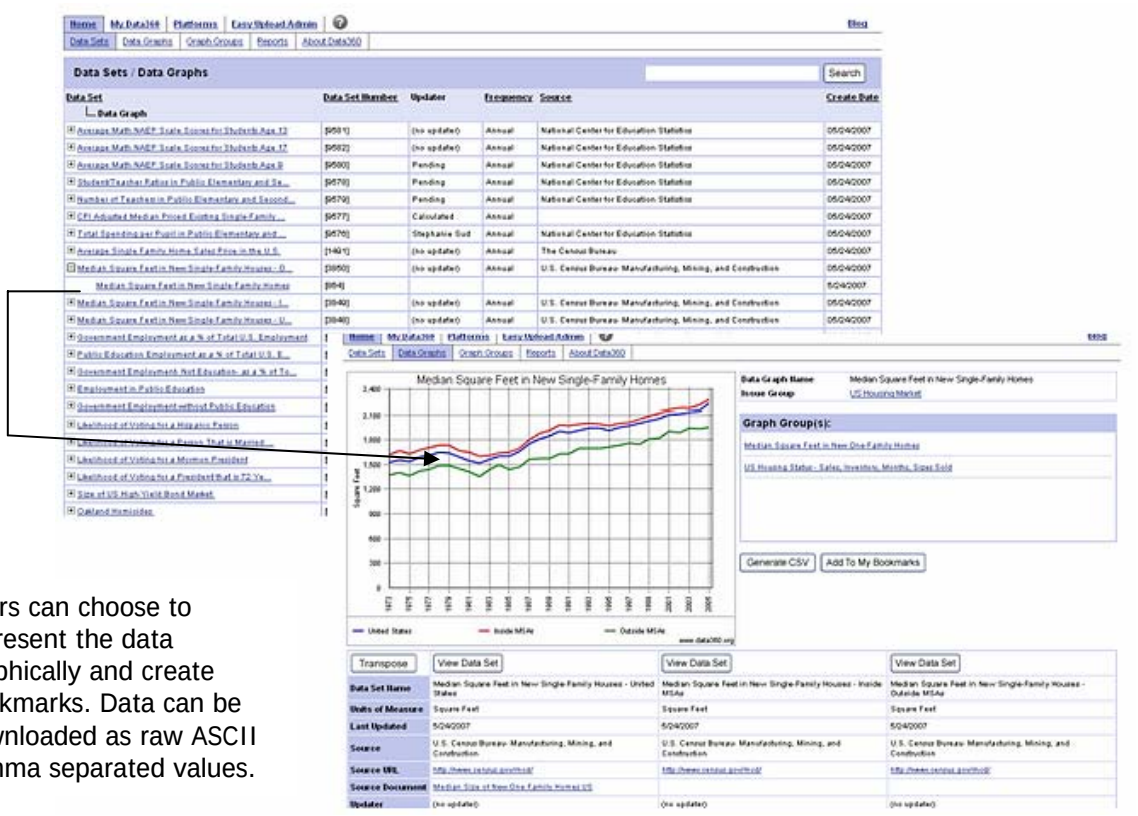
Data360

http://www.data360.org

Data360 embraces Web.2.0 concepts of participation and collaboration 'to provide clear context on important cultural, environmental, social and economic issues'. By using 'collaborative trend tracking' Data360 allows organisations and users to create their own Data360 space, as well as to collaborate with others. The datasets held are predominantly aggregated and from the US, with sources including: *US Environmental Protection Agency, National Science Foundation, US Geological Survey, Gallup, Federal Reserve, US Census Bureau, US Department of Energy.* There are also datasets sourced from *the World Bank, United Nations, the World Factbook Transparency International*. The frequency of the datasets is mainly annual although some daily, weekly, monthly and quarterly data are also available.



Data360 spaces can be built for free, as long as users are willing to a) share all their data openly and b) potentially (although not in the current version) have small advertisements on their pages. For those who want the software and access to the public data on the site, but still want to keep their data private, the software is available for a fee.

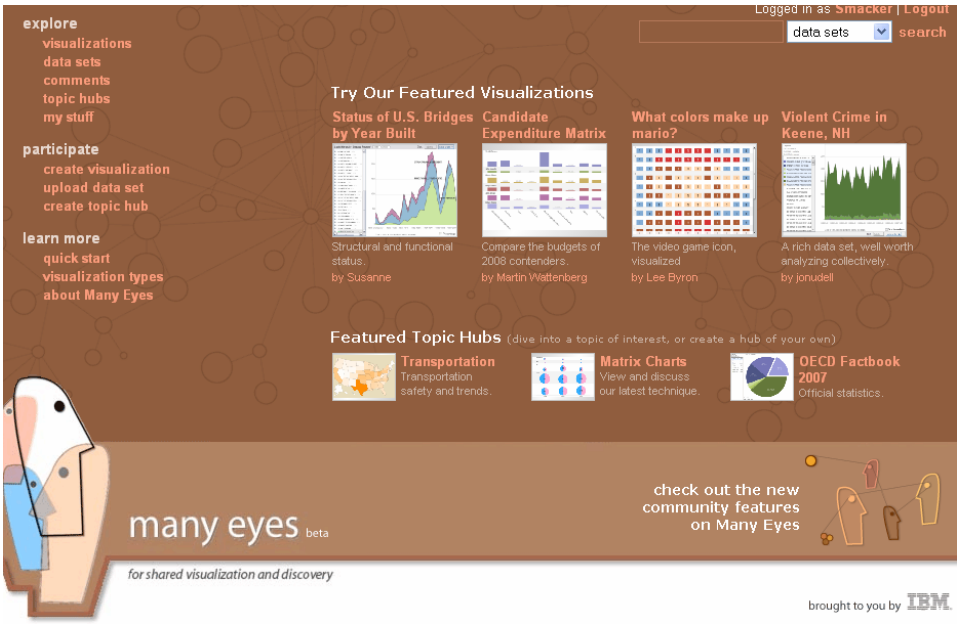


Users can choose to represent the data graphically and create bookmarks. Data can be downloaded as raw ASCII comma separated values.

Many Eyes

http://services.alphaworks.ibm.com/manyeyes/home

This IBM utility aims 'to "democratize" visualisation and to enable a new social kind of data analysis'. Many Eyes allows registered users to upload their own data on the understanding that the data is made available to all. It also has the facility to visualise numeric data in a spatial manner (using the proprietary mapping utility World Map). Users can view (or create) topic hubs which are pages devoted to a particular topic or theme. It serves as a way to collect related visualisations and data sets.



Data Visualisation Tools: Part 1 - Numeric Data in a Web 2.0 Environment

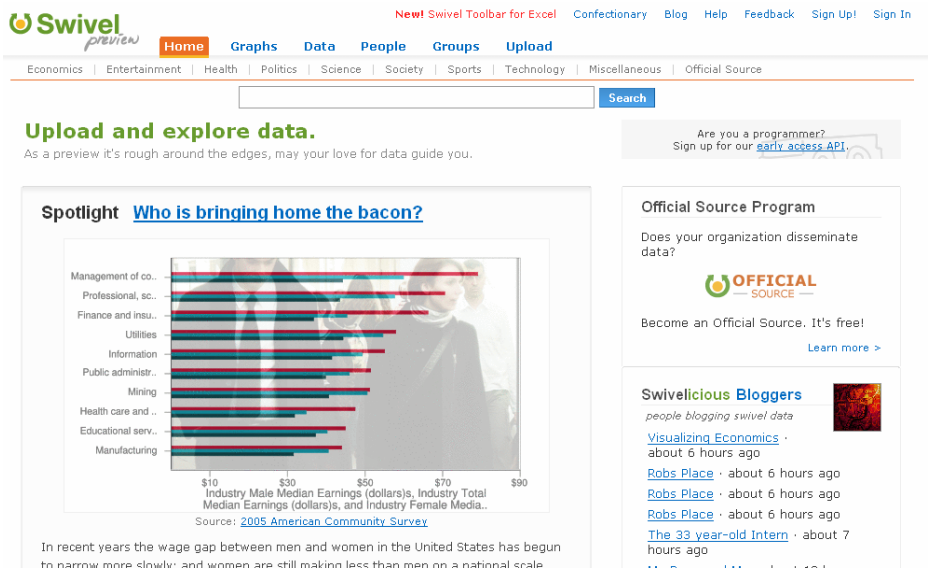
There is a range of visualisation tools with a range of functions, namely:

- Compare values (bar charts, block histograms, bubble and matrix charts)
- Tag clouds (to view word popularity in given text)
- Data point relationships (network diagrams, scatter plots)
- Parts of the whole (pie charts, tree charts)
- Track rises and falls over time (line graphs, stack graphs)

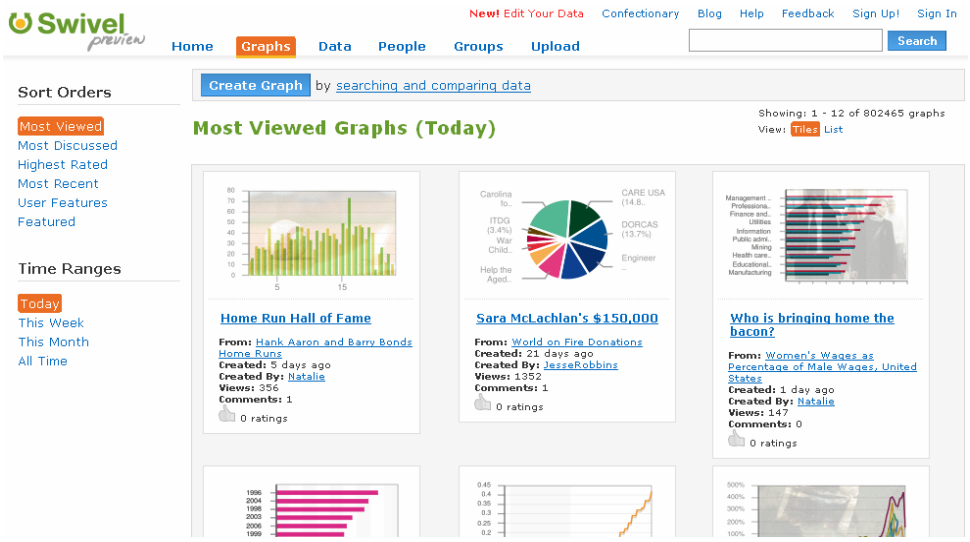
Swivel

http://www.swivel.com/

Swivel aims to 'liberate the world's data and make it useful so new insights can be discovered and shared'. Swivel charge for any restrictions a user may impose on the data. Swivel Geography or Swivel G allows users to zoom in and out and move seamlessly from one part of the globe to another. Swivel G also includes interactive features that enable users to drill down into specific data points in more detail.



Non-Government Organisations are directly uploading their data into Swivel. Swivel are also ingesting freely available data directly from NGO websites. This exposes their data (and ultimately their commercial products) to a wide audience in addition to providing users with access to trusted data from recognized and official sources. [3]



Gapminder World

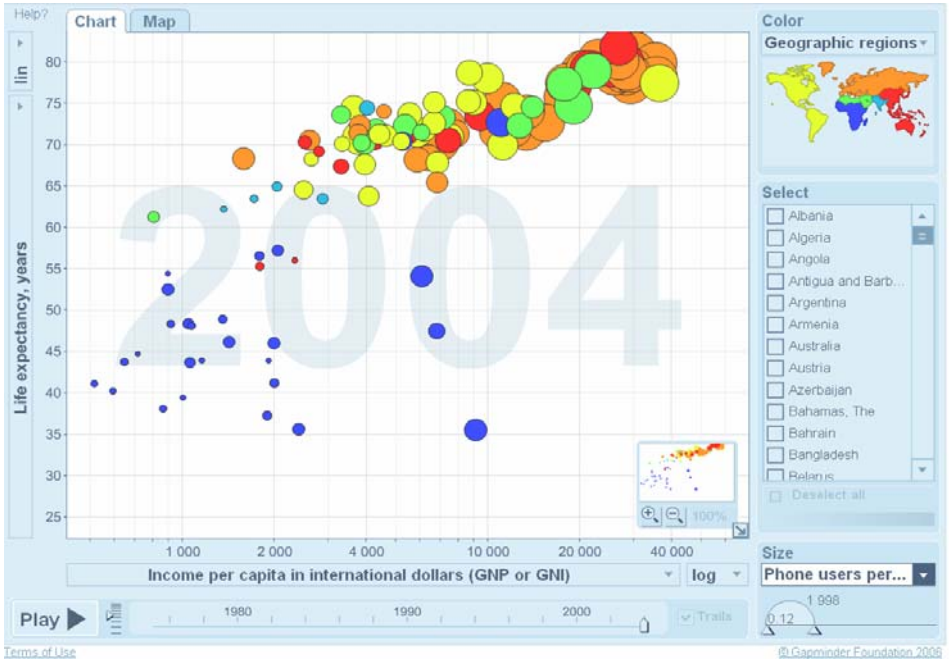
http://www.gapminder.org/downloads/applications/

Gapminder was originally funded by the Swedish International Development Cooperation Agency but has recently been acquired by Google. As with Data360, Gapminder's strength is the delivery of time-series data, in this case at the country level. The underlying *Trendalyzer* software imports data from Excel and shows moving graphics on the screen in a range of formats including PowerPoint.

At present Gapminder World contains only 16 variables. However collaboration is planned with the UN Statistics Division to visualise the Millennium Development Goals, and the UN Common Database with Trendalyzer on a test website, with a view to opening up the software for wider use.

A range of advanced features are under development, including: data importation, chart customisation, export charts as pictures/movies, share facility.

[3] In December 2007 Swivel announced the creation of a new Firefox search engine extension which searches for data at Swivel directly from the Firefox toolbar. Swivel also announced the release of Swivel Private Edition (invitation only) which enables sharing and collaboration on private data.

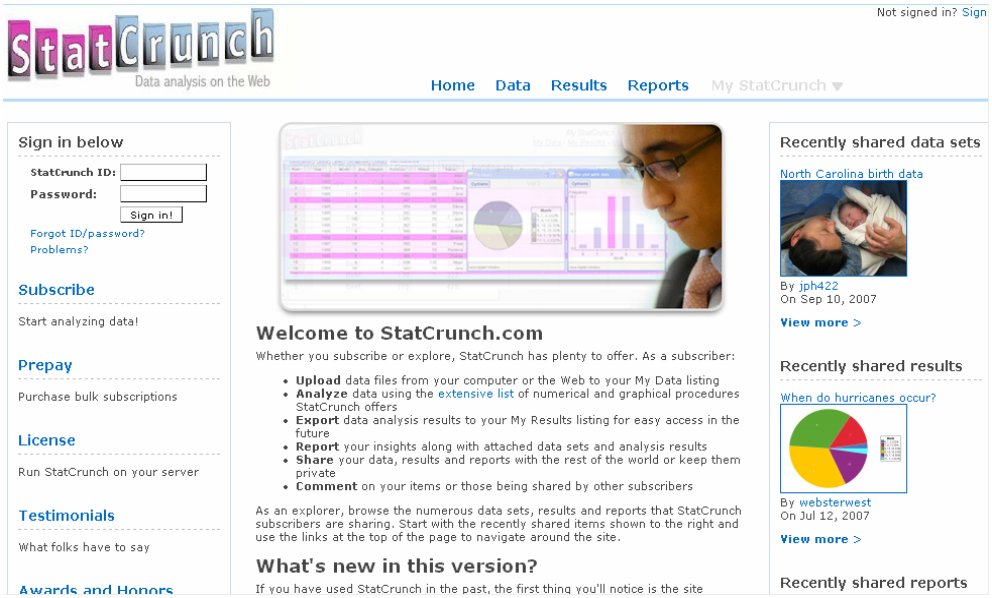


StatCrunch

http://www.statcrunch.com/

StatCrunch started life as a web-based statistical analysis tool, though recently it has been deploying more and more Web 2.0 functionality culminating in StatCrunch 5.0, launched in August 2007. It offers users the options to browse data sets, results and reports. However it differs from the 'open data' models described above in that users have to subscribe for a nominal monthly or annual fee in order to upload and analyse data files, create reports, export and share.

Academic institutions can also purchase a site licence to install the StatCrunch software on an institutional server. This model has similarities with the 'private data' fees provided by abovementioned resources.



Graphwise
http://www.graphwise.com/

Graphwise describes itself as 'a beta release of a search engine for table data'. Upon registration, users can search for and upload tabular data from their local machine or 'scraped' from an internet web page in a number of formats including Excel (.xls), comma delimited (.csv), HTML (.html), PDF (.pdf), Lotus spreadsheets.

Users can also share data by: publishing graphs in paper and on-line documents, creating a graph log ('glog') to share with others, pasting a live graph into their website or blog. A newer version of Graphwise, with a social networking portal, will be released in the near future. That version will allow users to publish their graphs to be searched and viewed by others.

How do the data visualisation utilities compare?

Table 1 aims to measure the functionality, versatility and *fitness for purpose* with regards to the deposition and utilisation of 'open' data held within the resources in question. One point has been assigned to a feature deemed present within the utility. A zero defines a missing function. 0.5 points have been attributed to a feature deemed as having neither a positive nor negative impact on the effectiveness of the utility, or a function that has been advertised via the resource as 'coming soon'. The exceptions are 'Contains adverts' and 'Privacy/Subscription fee' which have been assigned zeros.

From the objective criteria chosen it is evident that each data visualisation utility has associated strengths and weaknesses with regards to functionality and features, with each having a particular focus and/or business model.

Table 1: Functions and Features

Functions and features	Data360	Swivel	Many Eyes	Gapminder	StatCrunch	Graphwise
Contains adverts	Coming soon = 0.5	Coming soon = 0.5	N = 1	N = 1	N = 1	Y= 0
Bookmark facilities	N = 0	Y = 1	Y = 1	N = 0	N = 0	Y = 1
Uploads NGO/IGO data	Y = 1	Y = 1	Y = 1	Y = 1	N = 0	Y = 1
Privacy/subscription fee	Y = 0	Coming soon= 0.5	Y = 0	N = 1	Y = 0	N = 1
Offers spatial visualisation	N = 0	Y = 1	Y = 1	Y = 1	N = 0	N = 0
Data download e.g. via Excel plugin	N = 0	Y = 1	N = 0	N = 0	N = 0	N = 0
Blog	Y = 1	Y = 1	Y = 1	N = 0	N = 0	N = 0
Registration required	Y = 0.5	Y = 0.5	Y = 0.5	N = 0.5	Y (Fee) = 0.5	Y = 0.5
Plugin required e.g. Java, Flash, SVG	N = 0.5	N = 0.5	Y (Java) = 0.5	Y (Flash) = 0.5	Y (Java) = 0.5	Y (SVG) = 0.5
Help/FAQ available	N = 0	Y = 1	Y = 1	Y = 1	N = 0	Y = 1
Creative Commons Licence	N = 0	Y = 1	N = 0	N = 0	N = 0	N = 0
Keyword tagging facility	N = 0	Y = 1	Y = 1	N = 0	Y = 1	Y = 0
Application Prog-ramming Interface	N = 0	Y = 1	N = 0	N = 0	N = 0	N = 0
User comment facility	N = 0	Y = 1	Y = 1	N = 0	Y = 1	Y = 1
Tutorial / Getting started	Y = 1	Y = 1	Y = 1	Coming soon= 0.5	Y = 1	Y = 1
User topic hubs	N = 0	Y = 1	Y = 1	Coming soon = 0.5	N = 0	N = 0
Subject groups	Y = 1	Y = 1	N = 0	N = 0	N = 0	Y = 1
Report download	Y = 1	N = 0	N = 0	N = 0	Y = 1	N = 0
Data edit facility	N = 0	Y = 1	N = 0	Coming soon= 0.5	N = 0	N = 0
Syndication (RSS, Atom)	N = 0	Y = 1	Y = 1	Y = 1	N = 0	Y = 1
Search facility	Y = 1	Y = 1	Y = 1	Y = 1	Y = 1	Y = 1
Share facility	Y = 1	Y = 1	Y = 1	Coming soon = 0.5	Y = 1	Y = 1
Data upload	Y = 1	Y = 1	Y = 1	Coming soon= 0.5	Y = 1	Y = 1
TOTAL	**9.5**	**20**	**15**	**10.5**	**9**	**12**

Table 2: Number of Data sets

	Data360	Swivel	Many Eyes	Gapminder	StatCrunch	Graphwise
Start Date*	Jan 2006	Dec 2006	Jan 2007	Feb 2005	Aug 2007	Apr 2007
No. Public Data sets**	10619	7844	10364	16***	3421	N/A

***** *in present format*
** *as of December 2007*
** *under review. Currently no public data upload*

Table 3: Visualisation options*

Data360: line graph, stack graph, line – log Y-axis, bar regular, bar stacked, pie chart, scatter plot, bubble chart, marimekko

Swivel: bar chart, horizontal bar chart, scatter plot, line graph, transpose

Many Eyes: world maps, country maps, line graph, stack graph, bar chart, block histogram, bubble chart, matrix chart, scatter plot, network diagram, pie chart, tree map, tag cloud, word tree, transpose

Gapminder: world maps, bubble chart, line – log X/Y-axis, change-over-time visual

Statcrunch: bar plot, box plot, chi square, column statistics, contingency table, correlation, covariance, dot plot, frequency table, histogram, index plot, Kruskal-Wallis, logistic regression, Mann-Whitney, means plot, multi plot, multiple linear regression, one/two sample T, one/two sample Z, one/two sample proportion, one/two sample variance, pairs plot, parallel co-ordinates, pie chart, QQ plot, R chart, row statistics, scatter plot, sign test, simple linear regression, stars plot, stem and leaf, Wilcoxon signed ranks, x-bar chart, x-bar, R chart, c chart, p chart, u chart

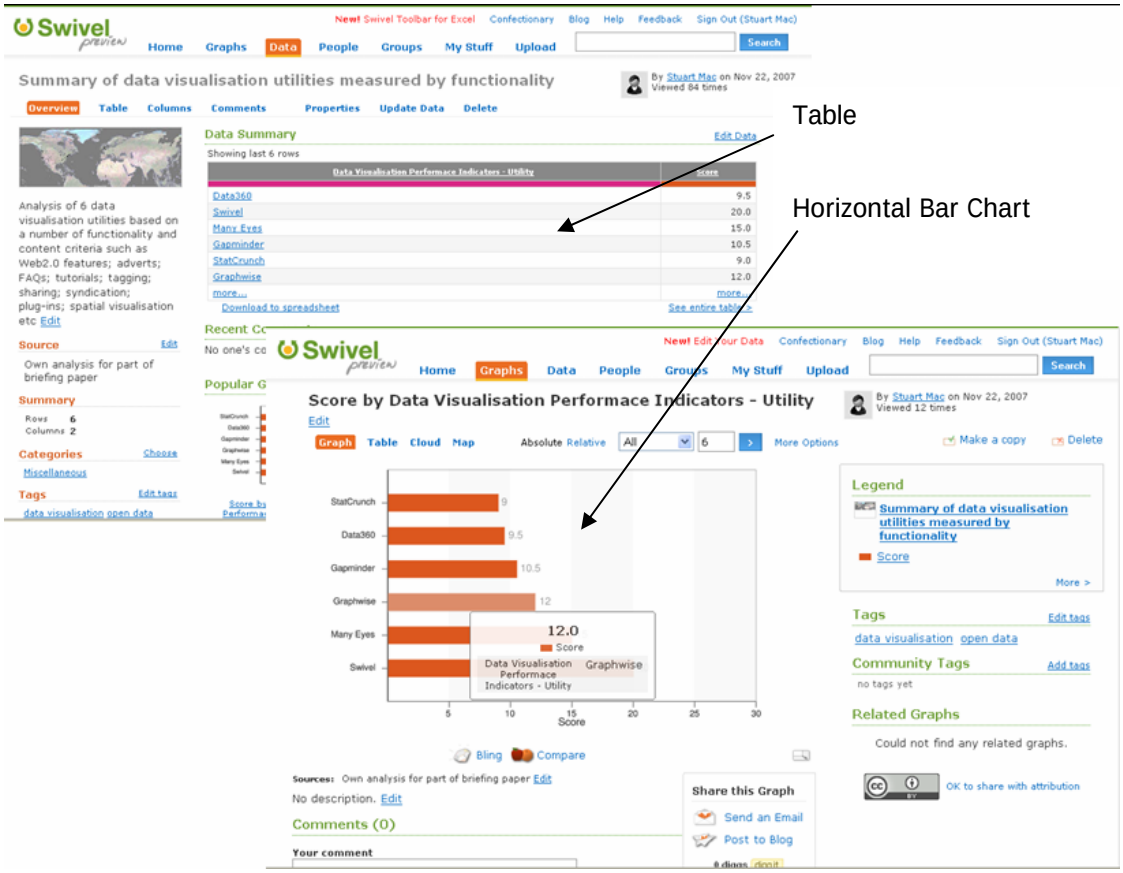
Graphwise:** bar chart, line graph, pie chart

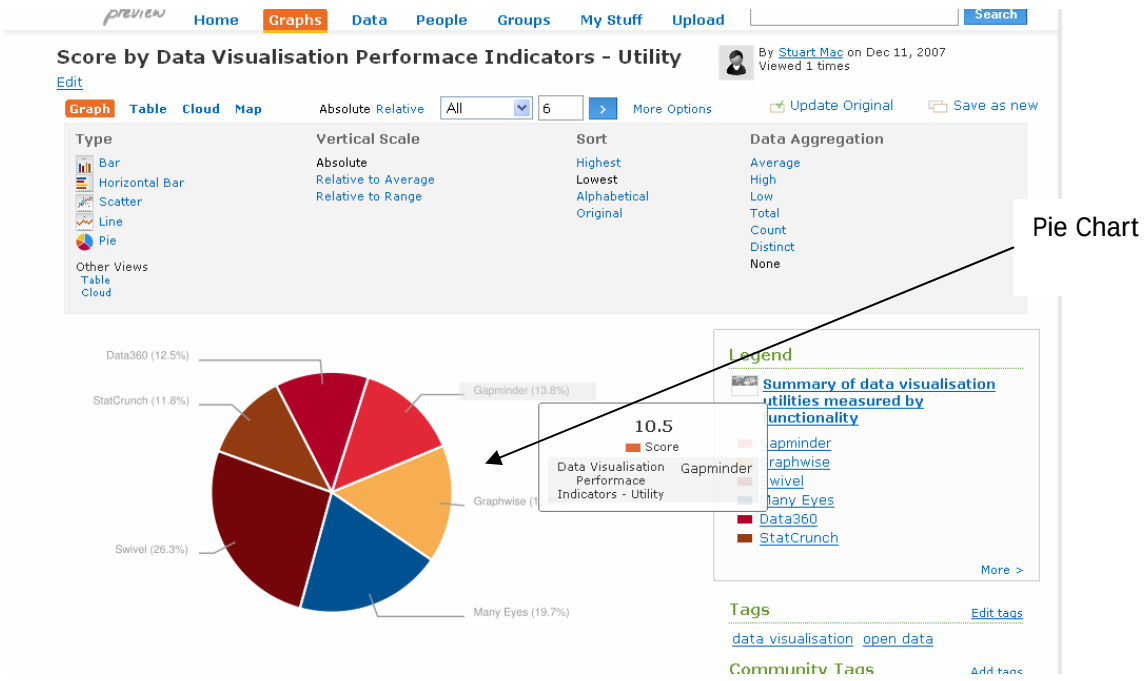
***** descriptions are as detailed within each specific utility
** new features to be released soon

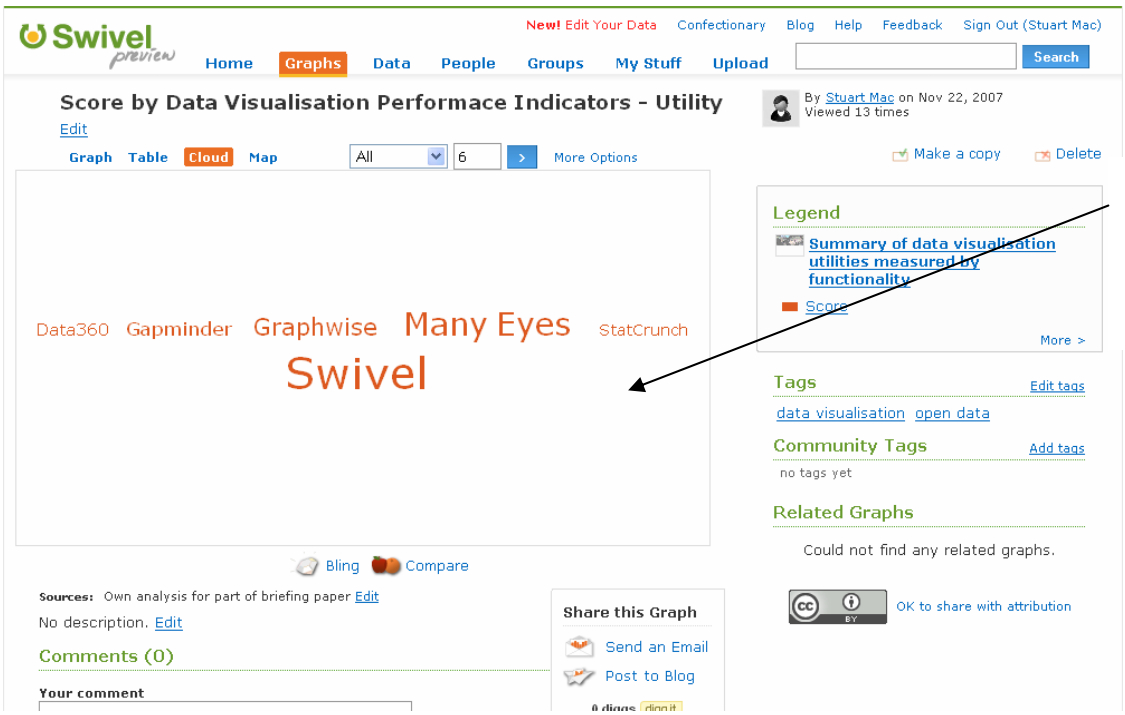
Swivel does appear to perform best with regards to overall Web 2.0 functionality; however this may be down to the fact that other utilities are more focused and hence have greater strength in other areas. For example, StatCrunch functions as a standalone statistical analysis package with a comprehensive array of graphing and charting functions, Graphwise places emphasis on its search engine functionality, and Gapminder focuses specifically on human development data.

Visualisation examples in Swivel

As a fairly straightforward 'proof of concept exercise' the data collected in Table 1 have been summarised, converted to CSV and uploaded into Swivel [the 'winner' of the function and feature comparison]. The following are the resultant data visualisation examples:







Summary

Much has been written about geographic visualisations using well publicised tools such as Google Earth and Microsoft Virtual Earth. Part two of this briefing paper will discuss such examples. There has been less publicity regarding the visualisation of numeric data, once thought the remit of domain experts. This paper has described a number of utilities that, to varying degrees, visualise data and allow users the opportunity to interact with and share data in an open environment.

Swivel appears to have a comprehensive range of functions and capabilities as measured by a range of Web 2.0 criteria and was used to highlight the visualisation of some simple data collected about 6 utilities. The visualisations produced give a flavour of how a user can quickly and simply turn raw data into a graphical representation, which enhances its value and versatility whilst stimulating discussion.

The examples showcased are by no means a definitive representation[4] but rather provide an overview of the trends, functionality, content and technologies pertaining to those utilities visualising numeric data. As Web 2.0 continues to evolve and transform into Web 3.0[5], we

[4] Examples of numeric data visualisation utilities not showcased include Dabble, Open Economics, Insight, Numbrary

[5] Web experts' definitions of Web 3.0

'The World Wide Database' (wide-area distributed database interoperability enabled by Semantic Web technologies) – Minding the Planet, Nova Spivack, October 2005 (http://novaspivack.typepad.com/nova_spivacks_weblog/2005/10/towards_a_world.html)

'Applications being pieced together' – Seoul Digital Forum, Eric Schmidt (CEO Google), May 2007, (http://sdf.sbs.co.kr/en/index.sdf)

'the distinction between professionals, semi-professionals and consumers will get blurred, creating a network effect of business applications' – Technet Summit, Jerry Yang (founder and Chief of Yahoo), November 2006, (http://www.technet.org/transcript1/)

'Giant Global Graph' – Decentralized Information Group (DIG) Breadcrumbs, Tim Berners-Lee's Blog (http://dig.csail.mit.edu/breadcrumbs/blog/4)

are seeing the boundaries between websites and web services blurring, as more and more web content becomes *remixable*. Many of the resultant applications can be achieved with no more than a basic understanding of the underlying technologies, which in turn enables the creation of further innovative resources. Thus the entire system is turning into both a platform and the global database (see www.twine.com; www.freebase.com etc) with data visualisation being but a tiny component of the continual evolution that is the web.

Appendix 1

Measuring societal progress through macro-data visualisation

One of the criteria used to evaluate the data visualisation resources was whether data from Non-Governmental Organisations (NGO) or Inter-Governmental Organisations (IGO) were being utilised within them. The statistical output of NGOs and IGOs are predominantly at macro-level and in the form of official statistics. There are a range of cutting-edge data visualisation and dissemination tools available for country-level statistics which span both the numeric and spatial data arenas. At present users cannot upload and publish their own data, although this may not always be the case. Examples include:

OECD Country Statistical Profiles – OECD statistics in SDMX (Statistical Data Mark-up eXchange) put into Adobe Flex
http://stats.oecd.org/nawwe/csp/default.html

IMF Data Mapper
http://www.imf.org/external/datamapper/index.php

Stat@tlas Europe - An online atlas of European statistical themes hosted by the Federal Statistical Office of Switzerland
http://atlas.bfs.admin.ch

Eurostat Tables Graphs Maps (TGM) – an interactive user interface to Eurostat data
http://epp.eurostat.ec.europa.eu/portal/page?_pageid=3013,65310704&_dad=portal&_schema=PORTAL

Other examples include Statistics Canada's Portrait of the Canadian Population in 2006: data tables, figures, maps and animations; Statistics Switzerland's regional and international comparators; Statistics Netherlands' Neighbourhood Information with Google Earth (see part two of this paper).

Following the World Forum on Statistics, Knowledge and Policy in Istanbul (June 2007) the Istanbul Declaration* has been adopted by the representatives of the European Commission, Organisation for Economic Co-operation and Development, the Organisation of the Islamic Conference, the United Nations, the United Nations Development Programme and the World Bank. The Istanbul Declaration develops further the Fundamental Principles of Official Statistics adopted by the United Nations in 1994 by creating a dynamic *to measure societal progress through statistical indicators*.

We are already seeing the data visualisation utilities participate in such a movement with, as detailed above, Gapminder World's plans to collaborate with UN Statistics Division to visualize Millennium Development Goals in addition to visualizing the UN Common Database. It can be argued that the dissemination, sharing and analysis of macro-data in Web 2.0 utilities (be it Swivel, Many Eyes or other proprietary solutions) can contribute to such a dynamic.

* http://www.oecd.org/dataoecd/14/46/38883774.pdf