Datapedia: a Yellow Brick Roadmap

Alan Freeman

Radical Demon

8. June 2008

Datapedia
A yellow brick roadmap

Alan Freeman

www.iwgvt.org

Abstract

This note lays out a roadmap to Datapedia: the goal is to share numbers with the same power and ease that the Wiki has delivered for documents. The goal is a system which, by analogy with Wikipedia can establish a world resource for reliable data. The paper discusses a process by which data providers and users can evolve a new set os systems for exchanging, describing and interacting with data to bring this about.

The first step would be *Datawiki*: an opensource system for recording revisions, changes and sources of data, allowing users to compare different revisions and versions of data with each other. It would be a set of protocols, and simple web tools, to help data researchers pool, compare, scrutinise, and revise datasets from multiple sources.

The first step towards Datawiki is *Wikidata:* rethinking the way that data itself is transmitted between people that collaborate on it a platform-independent standard for exchanging specifically numeric data. I show that the ubiquitous standard for exchanging data – the spreadsheet – is not up to the task of serving as a platform for Datawiki, and assess how alternatives can be developed.

The proposal centres on the metadata – additional descriptive data – that is associated with numeric data, and suggests how, in two cases – World GDP and Creative Industry Employment – data could be mapped in such a way that viable Datawiki platforms can be built.

The proposal also allows existing communities of users to start reshaping the way they exchange and handle data, to permit, and also to improve existing standards for collaborative use of data.

Datapedia
A yellow brick roadmap

Alan Freeman Tuesday, 20 May 2008

This note lays out a roadmap to Datapedia: the goal is to share numbers with the same power and ease that the Wiki has delivered for documents.

The eventual goal is a system which, by analogy with Wikipedia (perhaps even improving on some first-mover defects) could become a world resource for reliable data. Potentially, we can create repositories of numbers whose scope is wider than any dataset from an isolated researcher or a single institution and which are also better scrutinised, have less errors, are more accessible. The implicit assumptions behind numeric data can be made transparent, and the scientific community can explore alternative assumptions and compare the results. Multiple users can work together to improve quantitative data in the same way that they can now work together to improve text.

Improvement in ease of use are also easy to envisage: a range of applications in the public domain so that any user can interact with, revise, and comment on data using an interface of choice be it a spreadsheet, a database, a chart programme, a mapping programme, or new interfaces yet to come. We can anticipate a major expansion of what users can do with data. Charting, mapping, database and spreadsheet programmes will all visualise, interact with, and modify, the same datasets. The tortuous process of exporting and importing data in a baffling variety of formats will be abolished, saving enormous time and effort. Data integrity will be guaranteed by what I term *computational integrity* - rule-based data exchange eliminating laborious and time-consuming processes of checking for errors that do not need to be allowed into the system in the first place.

The first step: Wikidata

Along the road to Datapedia, the first step is *Datawiki*: an opensource system for recording revisions, changes and sources of data, allowing users to compare different revisions and versions of data with each other (as they can now with the Wiki). It would be a set of protocols, and simple web tools, to help data researchers pool, compare, scrutinise, and revise datasets from multiple sources.

The first step towards Datawiki is *Wikidata:* rethinking the way that data itself is transmitted between people that collaborate on it a platform-independent standard for exchanging specifically numeric data. I want to show that the ubiquitous standard for exchanging data – the spreadsheet – is not up to the task of serving as a platform for Datawiki. We are in the wrong branch of a QWERTY fork, stuck with a technology which is fit for single users but unfit for data sharing. The first step down the yellow brick road to Datapedia is to rethink, from the bottom up, the way we exchange numeric data with each other.

The main purpose of this paper is to explain Wikidata, and encourage providers and users to invest time and interest in building it.

Even this first step could have considerable implications for both data users and providers. Producing high-quality data is very labour-intensive, and many users today simply rely on the data provided by statistical authorities or ‘data agencies’. But, as the regular data user knows and the providers acknowledge, agency data is far from

unproblematic. Agencies work with raw or primary sources that are far from perfect. They have to 'correct' or clean the data as best they can, and render it in conformity with international standards. Along this road, many hidden and contested assumptions are incorporated which means that conclusions drawn from the data depend as much on these assumptions, as on the actual information contained in the final numbers. A Datawiki can transform this process: by comparing data compiled under different assumptions and from different sources it would become possible both to *pool* this information, *validate* it, and *harmonise* it. Users and providers could aggregate and compare data, test the impact of different assumptions on the conclusions drawn, and reaching agreement, through transparent discussion, on the best numbers to be used. But even the simpler step of *providing the data in a form that permits such comparisons* will transform the present situation.

Of course, data pooling and validation happens now to some extent. It is the bread and butter work of professional statisticians. But the sheer volume of work needed for such simple tasks as comparing even two releases of the same data – let alone two different classification systems – limits the extent to which this can be done and, importantly, enormously restricts both the number of people that can participate and the range of options that can be explored.

Ward Cunningham, the inventor of the Wiki, describes it as 'the simplest online database that could possibly work' and explains the goal of the wiki as editing text *quickly* (Wiki was taken from the Hawaiian for 'quick'). The basic reason there are no data wikis is that with existing technology, we cannot *quickly* do with data what we can with text

I'm hoping a concentrated assault on a small area can kick-start an overdue process faster than grand synthesis and, therefore, I'm starting with economic data. There is no reason to stop there: it's just a case in point. However I am hoping that my fellow economists will be enthused by the merits of what I propose. This note, therefore, faces two ways: it endeavours to talk to both economists and technical enthusiasts. Forty years as a programmer and twenty as an economist suggest to me that though the two groups know they need to talk to each other, they have yet to evolve a language to do it in, and I hope responders will bear this in mind.

The state of play

Recent advances make change possible. Surprisingly, it hasn't been done. I can see two reasons: the intrinsic limitations of the 'spreadsheet model' of data sharing, and the cult status of the expert, which acts as a drag on transparency in describing data. However it is very important to review what exists so far because it makes much of what needs to be done a great deal easier, and also, makes it more likely that the changes I am proposing can be part of a standard, which will increase the uptake.

The Open Document Format(ODF) has defined a world standard for documents, endorsed by the International Standards Organisation (ISO) and, grudgingly, Microsoft. ODF implements XML (Extensible Markup Language), a standard on which a successful datawiki could be based.

ON the application side, web-based applications like Google Spreadsheets, Hans Rosling's GapMinder, Zoho, and server-side databases like MySQL, allow people to publish numeric data on the web, visualise it in a variety of ways such as tables, maps and charts, interact with it, and discuss and improve it. Embedded clients like SQLite, the 'engine' in many web browser extensions, have also provide dramatic

improvements in the 'intelligence' of web browsers, such as Zotero, a canonical new bibliographic application. Projects like Luc Anselin's Geoda have put advanced geospatial mapping at the disposal of the open access community.

A variety of 'general purpose' projects are taking off to expand the Web's capacity for tracing heterogenous (mixed) data such as the Dublin Core and RDF initiatives. Finally, a number of projects are under way to standardise path-tracing and referencing in web documents. This will be important for numeric data in which, I will argue, the structure of the *computational references* that give rise to a number are the key to representing the assumptions incorporated into it.

Relatively little attention is focussed on standards for explicitly quantitative data. The *de facto* standard is to publish and exchange numbers using the simple table layouts provided by spreadsheets. These have hardly changed since the 1990s, and the web community has outgrown them. I think it's time to move on.

A wishlist

The following 'wishlist' for a true 'Datawiki' shows what needs to be done, and highlight the limitations of the above advances, important though they are.

(1) Each data element should be explicitly associated with the full range of attributes or 'metadata' that describe it and explain what it 'means'.

(2) Metadata should include source, revision history, dependencies, and methodology. Users should be able to trace the effect of revisions, source variations, and variant methodologies on any final result, and examine the alternatives just as Wiki users compare many versions of the same text.

(3) Data attributes should be standardised by the communities that need them. In some areas, such as GIS data, standards are already emerging. In economics these might include the indicator (GDP, population, employment, etc), the measure (real GDP, nominal GDP, PPP-based GDP, etc), classification (SIC, NACE, NAICS including the version), and so on. The standard should be extensible so additional attributes, and child attributes, can be defined flexibly.

(4) Data should be application-independent. The user can view, and modify, the same data using a spreadsheet, a database, an interactive chart, or a map. Every means of *visualising* data should also be a means of *modifying* it.

(5) It should be possible to visualise the data in the user's mode of choice: as a table, flat file, pivot table, chart, or map.

(6) Any number of 'dimensions' or repeating attributes should be allowed, such as time, territorial unit, industrial classification and so on, providing for the full multidimensional functionality found in pivot tables and OLAP cubes, for example 'slicing and dicing' and dimension switching.

(7) Dimensions should permit hierachies, for example the division of a country into its regions, or of a time unit such as a year into subunits such as months.

(8) Applications should be able to impose rules on repeating calculations to impose *computational integrity*: if the data for the elements of a hierarchy must add up to a third (for example, sales for each months add up to annual sales) it should be possible to include this specification in the metadata.

This calls I think, for a *data exchange* standard. It should define what goes in the files containing the data, not how the data is displayed. Users can then 'abstract' from the platform or programme which use the data. This is the route that has been taken for text and bibliographic information in Zotero, for example.

The first steps

Two red shoes are not needed to set out on a yellow brick road. The road probably passes through a range of ODF-compliant schemas or DTDs. Long before that, groups of researchers who use and publish data can start agreeing on limited sets of 'standard metadata fields' to include in the data they share with each other. Plugins to publish and read compliant data can be developed for popular applications, particularly open-source applications. And if a critical mass builds up, developers can work on applications to facilitate genuinely collaborative interaction with a growing mass of compliant data.

Initially, however, data can be circulated as flat files – as much data is today. The crucial point is what the file contains and, in particular, the *metadata* associated with it. Metadata is 'data that describes data'. For example, the GDP of Britain in 2006 was $2,357 billion, according to IMF figures released in September 2006. In this statement, '2357' is data. Everything else is metadata: GDP, Britain, 2006, billions, the IMF, and the release date, September 2006. Without the metadata, we don't even know what the number means, let alone whether it is the right one or what to do with it. Without metadata, data makes no sense. Yet, as this paper will show, the great bulk of metadata is either *absent* – not given by the provider – or *implicit* – given in such a way that, before the user can associate it with the data, she has to spend time and effort. In many quite trivial cases this effort is enormous, as when an agency neglects to supply metadata and has to be pursued relentlessly on the telephone or by email until the information is yielded. In less trivial cases it is routinely time-wasting,. To take a simple case, the source for the data in a table is invariably placed at the bottom of the table. This is an excellent way of *displaying* the data but a terrible way to *transmit* it; the user who wishes to compare the same data from different sources, or different releases of the same data, or different measures of a given indicator (real vs nominal GDP, etc) must laboriously put the metadata back where it belongs, next to the data, or devise some complicated technical device to conduct the comparison, such as spreadsheets with a set of aligned worksheets.

In a Datawiki file, each record would contain one data item along with its associated metadata. This would free the data of application dependency. Each data item would also be tracable back to its source, and a history of such revisions could then be maintained – exactly as can now be done with a text Wiki, but for each individual number. It would spare the data user the immense and wasteful labour of rummaging around to find the metadata and associate it with the data. It would lay the basis for simple web-based applications to conduct the great bulk of repetitive time-consuming operations such as comparing, splicing, pooling, validating, and so on. It would lay the basis for Datapedia.

My initial aim is, hence, to see how much agreement can be reached on the metadata a 'good' provider should put into a datawiki file in one restricted field: economic data.

Why spreadsheets are not enough

The first reaction I expect is: why go to all these lengths, when we have spreadsheets? The answer is that we use spreadsheets for two entirely different purposes which conflict with each other. We use them for data *exchange* – sending data to another researcher or another programme, and for data *processing* – making calculations.

The second reaction I expect is: why attach such a vast amount of information to every single number that changes hands? The short answer is 'because we can'. Ten years ago it seemed inconceivable to maintain, for ever, every single edit ever made to a learned article. It's now routine. In ten years' time it will be the same with numeric data. A slightly longer answer is that we *need to know where the number comes from if we want to understand the conclusions that people draw from it*.

I'm going to try and illustrate both points starting from table 1 below

Table 1: Gross Domestic Product (GDP)

Country	2000	2001	2002	2003	2004	2005	2006
Algeria	73	75	78	84	88	93	96
Argentina	214	204	182	198	216	236	255
Austria	212	214	216	218	224	228	235

Source: United Nations

This is a fairly innocuous table which everyone who has worked with economic data will recognise. Each row is a country, each column is a year, and every cell is a number, which tells us the GDP of that country in that year.

In fact, it isn't so simple. The row and column titles are *metadata*: they tell us what the numbers mean. However, they don't tell us very well. For example, agencies can never agree on what to call a country. This makes it difficult to *pool* data; we have to spend ages squaring the names with each other. Second, it's easy to violate *data integrity*. I can mess with the metadata just as easily as the data, for example turning Austria into Australia with little more than the slip of a finger. Thirdly we cannot *switch dimensions:* the layout limits us to looking at different years, or different countries, but no other source of variation.

Table 2 makes explicit the metadata that is implicit in table 1. In table 1, we can work out that '4,227' is the GDP of Algeria in 2001 but we have to make use of the row and column titles. Table 2 unambiguously says that '4,227' is the GDP of Algeria in 2001.

Table 2: flat file layout

Country	Year	Item
Austria	2001	214
Austria	2002	216
Austria	2003	218
Argentina	2001	204
Argentina	2002	182
Argentina	2003	198
Algeria	2001	75
Algeria	2002	78
Algeria	2003	84
Source: United Nations		

Table 3: flat file with source info

Source	Country	Year	Item
UN	Austria	2001	214
UN	Austria	2002	216
UN	Austria	2003	218
UN	Argentina	2001	204
UN	Argentina	2002	182
UN	Argentina	2003	198
UN	Algeria	2001	75
UN	Algeria	2002	78
UN	Algeria	2003	84

However there is further implicit information in the table. The 'source' information at the bottom is actually also part of the metadata. Table 3 tells us, for every item of data,

not only its country and year, but who supplied it – a piece of information that bibliographers, for example, could not possibly do without.

But there is still more implicit metadata. How do we know that these numbers refer to GDP? This, too, is part of the information associated with the data. It is 'tacit' metadata, concealed in the mind of the expert, or in the file name, or some such distant place. Worst still, there is no such thing as a single, agreed definition of what GDP *is,* even though it is one of the most standardised indicators known to social science. At least four measures are given out by international agencies: nominal in local currency; converted into dollars at market exchange rates; in 'real' or deflated national currency units; in real dollars; or even in 'Purchasing Power Parities' which attempt to allow for price variations between countries. And each of these measures has a range of variations: for example we can convert into dollars at the rate that holds at the end of the year, using an average, using the official rate if this is different from the market rate, and so on.

It doesn't stop there: other people supply the same data – for example the World Bank and the International Monetary Fund (IMF). They don't give the same numbers: in 2006 the World Bank said Belgium's GDP was $310bn and the IMF said it was $304bn. The IMF gives out two datasets: the International Financial Statistics (IFS) and the World Economic Outlook (WEO). In the second, data supplied by national statistical offices are 'cleaned' corrections are applied to make sense of it. But without comparing the two datasets, we cannot know what these corrections were.

Table 4: pooled data exchange file

Indicator	Measure	Method	Currency	Conversion year	Units	Revision year	Source	Country	Year	Data Item
GDP	C$	MX	USD	1990	BN	2006	UN	Austria	2001	214
GDP	C$	MX	USD	1990	BN	2006	UN	Austria	2002	216
GDP	C$	MX	USD	1990	BN	2006	UN	Austria	2003	218
GDP	C$	MX	USD	1990	BN	2006	UN	Argentina	2001	204
GDP	C$	MX	USD	1990	BN	2006	UN	Argentina	2002	182
GDP	C$	MX	USD	1990	BN	2006	UN	Argentina	2003	198
GDP	C$	MX	USD	1990	BN	2006	UN	Algeria	2001	75
GDP	C$	MX	USD	1990	BN	2006	UN	Algeria	2002	78
GDP	C$	MX	USD	1990	BN	2006	UN	Algeria	2003	84
GDP	C$	MX	USD	2000	BN	2005	WB	Algeria	2001	55
GDP	C$	MX	USD	2000	BN	2005	WB	Algeria	2002	57
GDP	C$	MX	USD	2000	BN	2005	WB	Algeria	2003	61
GDP	C$	MX	USD	2000	BN	2005	WB	Argentina	2001	272
GDP	C$	MX	USD	2000	BN	2005	WB	Argentina	2002	242
GDP	C$	MX	USD	2000	BN	2005	WB	Argentina	2003	263
GDP	C$	MX	USD	2000	BN	2005	WB	Austria	2001	192
GDP	C$	MX	USD	2000	BN	2005	WB	Austria	2002	194
GDP	C$	MX	USD	2000	BN	2005	WB	Austria	2003	196

The agencies also produce new data or *revisions* to the data. These can be very large: in April 2008, the IMF revised China's GDP, in PPP terms, downward by 40%. We also need to know when the data was published, as with a book citation. A table which even pretends to be complete would thus look something like table 4.

I have used some abbreviations: C$ means constant dollars, MX means Market Exchange, UN means United Nations, and so on. These abbreviations can be included

with the data, and as we shall see, can become a part of the exchange standard by encoding them in the DTD or schema.

The additional metadata is needed to assess whether researchers' results are *robust*: do they arise from reality, or the way the data was constructed?

The table is much larger, and when completed is larger still, including rows for Argentina's GDP sourced from the IFS, WEO, etc.; it includes GDP measured in terms of PPP, nominal or real GDP, revision years 2000, 2001,…2006, and so on and so on. Not every data provider offers all this information, but if every provider includes the same metadata fields, such large tables will arise as data is *pooled* – added together from files created by a diverse community of researchers.

We can now start to see the *effect* of changing the source and conversion year. In 2001, Algeria's GDP was $75bn according to the UN at 1990 dollars, but $55bn according to the WEO at 2000 dollars. Is this entirely due to the change in year, or is there also a difference between the sources? The simplest way to study this is to consider an alternative tabular layout shown in Table 6.

Now something interesting happens. Laid out in this way, we are under no obligation to preserve the spreadsheet ordering: Indeed relational database files have no implicit ordering at all. It can be looked at it as in table 5.

Table 5: as table 4, re-ordered to illustrate dimension switching

Indicator	Measure	Method	Currency	Conversion year	Units	Revision year	Source	Country	Year	Data Item
GDP	C$	MX	USD	1990	BN	2006	UN	Algeria	2001	75
GDP	C$	MX	USD	2000	BN	2005	WB	Algeria	2001	55
GDP	C$	MX	USD	1990	BN	2006	UN	Argentina	2001	204
GDP	C$	MX	USD	2000	BN	2005	WB	Argentina	2001	272
GDP	C$	MX	USD	1990	BN	2006	UN	Austria	2001	214
GDP	C$	MX	USD	2000	BN	2005	WB	Austria	2001	192
GDP	C$	MX	USD	1990	BN	2006	UN	Algeria	2002	78
GDP	C$	MX	USD	2000	BN	2005	WB	Algeria	2002	57
GDP	C$	MX	USD	1990	BN	2006	UN	Argentina	2002	182
GDP	C$	MX	USD	2000	BN	2005	WB	Argentina	2002	242
GDP	C$	MX	USD	1990	BN	2006	UN	Austria	2002	216
GDP	C$	MX	USD	2000	BN	2005	WB	Austria	2002	194
GDP	C$	MX	USD	1990	BN	2006	UN	Algeria	2003	84
GDP	C$	MX	USD	2000	BN	2005	WB	Algeria	2003	61
GDP	C$	MX	USD	1990	BN	2006	UN	Argentina	2003	198
GDP	C$	MX	USD	2000	BN	2005	WB	Argentina	2003	263
GDP	C$	MX	USD	1990	BN	2006	UN	Austria	2003	218
GDP	C$	MX	USD	2000	BN	2005	WB	Austria	2003	196

The differences, we can now see, are not just due to the different year of conversion. For Algeria, the UN estimate of GDP is larger, but for Argentina, the World Bank's is. This is probably due to a reason the data does not yet reveal: for example because the growth rates of the two countries have been estimated differently, or because they have recalculated their rates of inflation. It is likely to be related to the sudden devaluation of the peso on 2000. And it will have significant effects for any conclusions that the unwary might draw from the data, since they will get a different answer to the same question, depending on whether they use the World Bank or the United Nations as a source. Without pooling and re-dimensioning the source

information and laying it out in this way, all but the most dedicated would not even know the question existed, let alone wonder about the answer.

Table 6: dimension switching

Country	Year	UN	WB
Algeria	2001	75	55
Argentina	2001	204	272
Austria	2001	214	192

Dimensions, data hierarchy and data interrogation

Dimension-switching is familiar to pivot tables users. In this case it was made possible by pooling data from two sources. But most data, even from a single source, is implicitly multidimensional. The World Bank provides, as noted, at least four different measures of GDP, and reports each of these measures for every country and every year. This gives us a handle on what a 'dimension' might mean in a Datawiki:

A dimension is a metadata field that repeats in such a way that no other metadata field changes.

Not every metadata field qualifies as a dimension. For example, it is desirable to record the permission status of a data item – whether it is confidential, the subject of Intellectual Property, covered by creative commons, and so on – but there is no reason to build tables of data with different permission values. Useful also are things like format information (fixed or floating point, precision, preferred display form, etc). Again, these won't be used to construct tables. Similarly information about units (whether recorded in billions, thousands, etc) is needed, but not dimensional.

Hence, as part of the Datawiki standard it is helpful to try and agree which metadata fields do constitute dimensions in economic data. This, I think, is probably the core of the specification. My initial suggestion is in table 7. In each case I have grouped the metadata with square brackets because metadata has a structure, as table 8 indicates for GDP. This table is not meant to be exhaustive, restrictive or even rigorous. I've drawn it up to make two points:

(1) Any group of people who want to share data will need to agree metadata schemas to define the way they to 'talk about' it.

(2) when this is done, it will become clear which are the 'true' data dimensions – those that correspond to the nature of the object, as the researchers' consensus defines it.

This list immediately introduces a second issue which takes us beyond the spreadsheet model: data *hierachies*. Table 7 contains two examples where a basic category has subcategories: 'Indicator' and 'Territorial Unit'. The primary relation between the two is *dependency* or instantiation: concept is a 'type' of indicator and 'country' is a type of territory. For 'territory' however we have something additional which is specific to numeric data: 'additivity'. The GDP of the countries in a region add up to the GDP of the region. This applies only to certain indicators, of course and, for example, their GDP per capita cannot be aggregated in this simple way. Nevertheless, the key point is that data hierarchies, once numbers are involved, contain additional constraints that a simple parent-child relation, as specified in XML, does not cover.

Hierarchy leads naturally to a second requirement for handling repetitive data: *dicing*, which takes place when we step down a data hierarchy. To illustrate this I turn to a

second dataset, relating to the Creative Industries (CI), and given in table 8. This table was used in constructing figures for 'creative intensity' which is related to what Higgs term 'embeddedness' – to show how many creative workers are employed within the creative industries, and how many non-creative workers.

Table 7: shortlist of economic metadata

- Indicator, including
  - Concept eg GDP, population, exports
  - Measure eg current, constant/real
  - Units eg local currency, PPP, dollars
- Time
- Territorial Unit, including
  - World region eg Latin America
  - Country
  - Subregion eg California, Baroda, Queensland
- Classification (eg SIC, SOC, NAICS, NACE)
- Provenance (eg Raw source; primary source; provider; date of revision).

Here the data hierarchies are (partly) exposed in the table layout. The 'London' and 'Non-London' jobs *add up* to British jobs. Across columns, the jobs inside, and outside, the creative industries add up to total jobs across all industries. The 'main' and 'second' jobs add up to the total number of creative jobs.

Aggregating rows, similarly, all the jobs whose occupations belong to the DCMS sector 'Advertising' (SOC codes 1134, 3433 and 3543) add up to the total number of workers whose occupations fall in this sector.

Tables 9 and 10 are 'condensed' tables that summarise this detailed information in two different ways. The two tables address completely different questions and the 'dimensions' are therefore different. But they relate to the same underlying dataset. This 'slicing and dicing' – the ability to fold, and unfold, elements in a data hierarchy, is a fundamental requirement of data interrogation.

The issue which again surfaces is the following: the 'diceability' of the data is an *attribute of the data* and it should not be the prerogative of the application to permit, or not permit it, or for the user to choose, or not choose to do it. In specifying the metadata, a specific requirement has to be incorporated that says 'working inside or outside London is a decomposition of working in Britain'.

Once this is done, a number of immediate improvements result of which the most important is *data integrity*. When data users construct tables like table 8, they have to insert formulas into the spreadsheet which it is easy to get wrong. In fact, these formulas are a consequence of the metadata structure and neither should it be at the user's discretion to override it, nor should the user have to waste time working out which formula to use to incorporate it.

In this respect, Datawiki needs to go a stage beyond the data integrity protection imposed by the relational database model, which does take precautions to ensure that the metadata is consistent. The next step is to ensure that not only the data, but the

formulae connection them, are consistent, with the dimensional and hierachical structure of the metadata.

Table 8: creative industry employment in London

		Main Job				Second Job			
		London		Non-London		London		Non-London	
SOC	DCMS sector	Within CI	Outside CI	Within CI	Outside CI	Within CI	Outside CI	Within CI	Outside CI
0000	Not creatively occupied	623,585	2,797,419	3,689,219	18,740,807	35,358	136,070	203,570	1,178,952
1134	Advertising	12,759	11,412	9,827	13,157	151	-	-	979
3433	Advertising	2,666	8,676	2,620	16,349	334	292	337	712
3543	Advertising	14,802	20,445	24,578	56,502	484	189	405	2,425
2431	Architecture	14,539	1,085	21,119	6,936	-	340	114	1,061
2432	Architecture	157	1,366	4,371	10,529	-	-	-	-
3121	Architecture	1,604	388	6,377	4,385	-	-	368	-
5491	Crafts	383	929	2,070	22,118	219	-	-	583
5492	Crafts	449	1,816	5,331	39,862	-	212	426	457
5493	Crafts	-	-	309	4,615	-	-	-	-
5494	Crafts	327	350	390	2,849	-	-	-	-
5495	Crafts	-	2,134	1,970	6,026	-	-	98	228
5496	Crafts	1,641	-	8,346	123	-	-	210	348
5499	Crafts	508	1,437	1,420	12,569	-	-	259	1,728
8112	Crafts	-	-	-	11,671	-	-	-	-
9121	Crafts	-	14,733	3,419	154,387	-	1,048	734	4,675
2126	Fashion	1,490	2,186	12,799	53,591	-	434	359	643
3411	Fashion	5,048	518	19,180	3,849	831	592	317	3,078
3421	Fashion	15,476	5,603	31,805	26,817	-	209	632	1,781
3422	Fashion	5,392	8,085	12,164	17,532	602	261	318	1,677
5411	Fashion	-	-	573	7,547	-	-	-	118
3434	Film & video	15,282	1,548	24,568	6,488	577	1,040	1,354	2,421
1136	Leisure software	17,536	39,644	54,748	133,980	-	765	1,453	1,231
3412	Music & VP arts	12,830	4,972	21,779	11,168	1,732	3,080	1,682	6,682
3413	Music & VP arts	4,551	1,542	11,138	3,771	2,103	1,951	3,492	15,560
3414	Music & VP arts	350	341	928	511	-	159	-	421
3415	Music & VP arts	11,794	1,348	11,750	2,783	538	3,259	2,762	12,552
3416	Music & VP arts	14,584	2,374	9,266	3,427	1,099	1,291	387	1,843
5233	Non-DCMS	-	934	198	7,792	-	-	-	-
3431	Publishing	21,828	5,786	27,131	4,419	919	2,601	1,658	3,582
5421	Publishing	482	718	2,717	8,377	-	-	-	551
5422	Publishing	1,475	5,753	4,459	28,506	-	-	-	334
5423	Publishing	629	1,680	3,948	17,235	-	126	-	191
5424	Publishing	-	185	-	5,391	-	-	-	-
3432	Radio and TV	24,659	1,375	16,382	1,201	468	-	1,021	997
5244	Radio and TV	777	557	2,574	7,508	-	-	140	198

Source: GLA Economics

Table 9: Where creative workers are employed

	Working in London	Working outside London
Advertising	72,209	127,889
Architecture	19,478	55,259
Crafts	26,184	287,219
Fashion	46,724	194,777
Leisure software	80,558	232,721
Film & video	18,447	34,830
Radio and TV	27,834	30,021
Music & VP arts	48,216	88,579
Publishing	42,183	108,499

Table 10: Who employs London's creative and non-creative workers?

	Working in Creative industries	Working outside Creative industries
Creative workers	204,015	149,916
Not creative workers	623,585	2,797,419

Tin men, scarecrows, and lions: will shared data be better data?

These initial considerations allow us to think about what is probably the most important reason for a system of collaborative data improvement: namely, in field after field it is being discovered that the sum is greater than the parts: when a group of people work together, if the ground rules are set correctly, they can produce something better than a single individual.

Rules are to be taken seriously. Larry Sangler, a founder, set up *Citizendium* as a 'Wikipedia with editors and real names' for only two reasons: Wikipedia's editorial procedure or lack of it, and the unaccountability of anonymous contributors.

In establishing rules, however, we need to be mindful of two things. First, how do the existing rules work, and what is wrong with them? Second, how can we devise rules that will make things work better? I would be cautious of a sweepingly enthusiastic vision which simply assumes that, because a new form of collaboration is introduced, the result will necessarily and under all circumstances be better. It certainly has the capability to be better; I suspect, however, that we will need to pay attention to the rules.

Datapedia will need rules. The basis for a rule is its outcome. How should collaboration improve data quality? The following list suggests itself:

- *pool* data,
- *compare* data,
- *scrutinise* data,
- share *functionality.*
- *revise* data,

Pooling data aggregates data from different sources to create a larger sample or wider coverage in time, territory or other variables: for example, calculating North American GDP by adding together Canada, USA and Mexican data.

Comparisons examine the effect of differences in time, space or other variables, for example, comparing the proportion of employment in the creative industries in Sydney, London and Paris.

Data can be *scrutinised* by comparing the reporting of the same indicator from different sources, different revisions, or using different methodologies, assumptions and classifications, and interrogating the reasons for the differences.

Functionality can be added to data by means of server-side bolt-on 'gadgets' such as graphic display software, mapping software, or statistical processing and mathematical transform software. For example, a provider of mapping or visualisation services could offer a webservice whose input is geo-coded time-series data for cities and the output would be a chloropleth map, a geographic or time-series trend line, or a visualisation such as that provided, for example, by Gapminder.

Data can be *revised* and improved by a similar process to the text Wiki: users can suggest localised revisions, and the impact of these revisions on final results can be seen without discarding the originals. A consensus can be reached, through discussion, on the best revisions.

At the end of the yellow brick road lies the very last of these objectives.

Dimensional localisation and computational integrity

How can we arrange for collective data revision? In my view this is the central issue.

The way that Datawiki can approach this is my final point. I am going to introduce two concepts termed *dimensional localisation* and *computational integrity*, at least until somebody suggests more usable and less grandiose terms.

Anyone who has used a pivot table or supercube – or for that matter, an accounting package – should be able to see that there is a crucial difference between the way such programmes apply repetitive calculations, and the way that a spreadsheet user does it. If we make table 8 into a pivot table, the user has no choice but to accept that, for example, all the creative and non-creative jobs add up to the total of all jobs, or that London jobs added to non-London jobs add up to British jobs. A specific function – addition – is applied to all items within the dimension [London/non-London].

Similarly, in an accounting package, all items on an invoice add up to the invoice total; all items on an account add up to the account balance, and so on. These appear to be properties of the programming package we are using, but actually they are 'metadata rules' – it is in the nature of the London/non-London split that the components must add up to Britain. *Computational integrity* means imposing this restriction and guaranteeing it.

In a spreadsheet, computational integrity is the responsibility of the user. This is even the case for data provision and it is, unfortunately, more common than the public imagines that a respected agency will publish data in which the columns or the rows just don't add up. If, however, computational dependencies are included in the metadata specification, such errors will simply be impossible, or rather, they will occur only if the application itself is faulty.

The key point about computational dependencies is that they apply, not to individual data elements, but to all the elements of a dimension. In the early days of computing languages, there were many weird and wonderful attempts to express this, most notably APL, arguably the most obscure language in the history of the universe. The basic idea is, however, not a difficult one: certain mathematical relations are, in Aristotelian terms, *necessary* properties of the data. We expect a restaurant bill to add up, not just in Joe's diner last Friday, but always and in all restaurants. It's part of the nature of a bill. Computational integrity, in plain English, ensures that numeric objects conform to their nature.

Now here's the basic point. If we can embed computational integrity in the metadata, first of all we have a much superior guarantee of data integrity. But second, we get dimensional localisation. Suppose we want to compare the effect of substituting the World Bank as a data supplier for the United Nations; or of studying New York instead of London for the distribution of creative employment. We want to make this *one* change, and be assured that everything else will change in comformity with that change.

This is the 'one change' affects an entire dimension – which, I hypothesis, could be recorded in a Datawiki list of revisions and comparisons. It changes 'the dimension, the whole dimension, and nothing but the dimension'.

This, I think, is what can bring Text Wiki practices within the grasp of users. Localisation is a key property of the Wiki. If you revise a Wiki article, your changes are confined to a particular place in the document. 'Search and replace' revisions are,

indeed, surprisingly difficult to specify or implement. This ensures that, when somebody compares two versions of a text, they can *see* where the changes took place.

Conclusion

'If you build it', says the voice in the cult movie *Field of Dreams*, 'They will come'. This paper is not a proposal to issue an enormous contract to a software company to 'build datapedia'. It isn't a proposal for an august international conference to draw up a report on '2020 prospects for Datapedia', or a new WW3 committee or even a new opensource collaboration between a bunch of eager future-builders . It is simply a proposal to start doing certain standard things differently, above all, transmitting data differently.

To this extent it is counter-intuitive. It appears to be asking for *more work* in order to produce *less work*. The reason the QWERTY keyboard will probably never be reformed is, curiously, not the volume of material capital bound up with it – it's easy as pie to produce non-QWERTY keyboards – or even the software investment – again, it's the easiest thing in the world to change a programme. The real obstacle is sheer human investment, the 'work' which would be involved in a thousand million white collar writers learning to type differently. In the same way, a significant community of data providers and users are basically stuck in a groove of exchanging tabular data with missing metadata, and simply 'coping' with the problems this creates, much as typists 'cope' with using their left ring and little fingers for the everyday word 'as'.

There is also institutional resistance. Data providers acquire, from the lack of metadata and explanatory background, a mystique bound up with authenticity; two generations of journalists, for example, whilst they may question the World Bank's judgement, have never thought to examine its numbers, not least because it is simply too difficult. When the IMF revised its estimate of China's GDP downwards, in April 2008, none but the most informed and thoughtful commentators even passed comment.

At the end, however, the argument for change is the overwhelming benefits for everyone that arise. And in fact, it is not so difficult to envisage how change can begin, through the work of 'intermediaries' – data users who will take it on themselves to rewrite data in a standard format, just as a generation of Web users have taken it on themselves to place vast amounts of information in text form onto the web. Once a community is established of users that expect data in wikidata formats, it will not be long before providers recognise the value of exporting in this format. In addition much of the work is mechanisable.

The key task is, therefore, not whether habits can be changed but what they should be changed *to*: what metadata is required, how computational and dimensional integrity should be expressed, and how this can be expressed in a manner compatible with emerging XML and ODF standards. With a little bit of good will and understanding, I'm sure this will be resolvable.

Alan Freeman 07/06/2008

Bibliography

In keeping with the subject matter, all references in this document are hyperlinked. The following is an exhaustive list of sources thus referenced.

Alan Freeman. 2007. “London's Creative Sector: 2007 update.” Available at: http://www.london.gov.uk/mayor/economic_unit/docs/wp_22_creative.pdf [Accessed June 7, 2008].

Citizendium. Available at: http://en.citizendium.org/wiki/Main_Page [Accessed June 7, 2008].

Dublin Core Metadata Initiative (DCMI). Available at: http://dublincore.org/ [Accessed June 7, 2008].

GeoDa - An Introduction to Spatial Data Analysis — GeoDa. Available at: https://www.geoda.uiuc.edu/ [Accessed June 7, 2008].

Google Docs Tour. Available at: http://www.google.com/google-d-s/tour1.html [Accessed June 7, 2008].

Hans Rosling. “Gapminder.” Available at: http://www.gapminder.org/ [Accessed June 7, 2008].

MySQL :: The world's most popular open source database. Available at: http://www.mysql.com/ [Accessed June 7, 2008].

OASIS Open Document Format for Office Applications (OpenDocument) TC. Available at: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office [Accessed June 7, 2008].

Online Office, Word Processor, Spreadsheet, Presentation, CRM and more. Available at: http://www.zoho.com/ [Accessed June 7, 2008].

Open Document Format published as ISO standard. Available at: http://arstechnica.com/news.ars/post/20061204-8349.html [Accessed June 7, 2008].

Open-Source File Format Is to Be a Part of Microsoft Office - NYTimes.com. Available at: http://www.nytimes.com/2008/05/22/technology/22format.html?_r=2&ex=1212120000&en=55fce7c06b0a4fa4&ei=5070&emc=eta1&oref=slogin&oref=slogin [Accessed June 7, 2008].

Resource Description Framework (RDF) / W3C Semantic Web Activity. Available at: http://www.w3.org/RDF/ [Accessed June 7, 2008].

SQLite Home Page. Available at: http://www.sqlite.org/ [Accessed June 7, 2008].

The APL Programming Language. Available at: http://www.engin.umd.umich.edu/CIS/course.des/cis400/apl/apl.html [Accessed June 7, 2008].

United Nations Statistics Division - National Accounts. Available at: http://unstats.un.org/unsd/snaama/dnllist.asp [Accessed June 7, 2008].

What Is Wiki. Available at: http://www.wiki.org/wiki.cgi?WhatIsWiki [Accessed June 7, 2008].

Wikipedia, the free encyclopedia. Available at: http://en.wikipedia.org/wiki/Main_Page [Accessed June 7, 2008].

World Bank. “World Bank Group World Economic Outlook Query Page.” Available at: http://ddp-ext.worldbank.org/ext/DDPQQ/member.do?method=getMembers&userid=1&queryId=135 [Accessed June 7, 2008].

XML Files - Introduction to DTD. Available at: http://www.xmlfiles.com/dtd/dtd_intro.asp [Accessed June 7, 2008].

XML Files - XML Tutorial - XML Usage. Available at: http://www.xmlfiles.com/xml/xml_usedfor.asp [Accessed June 7, 2008].

XML Schema Part 0: Primer Second Edition. Available at: http://www.w3.org/TR/2004/REC-xmlschema-0-20041028/ [Accessed June 7, 2008].

Zotero: The Next-Generation Research Tool. Available at: http://www.zotero.org/ [Accessed June 7, 2008].