The ICTD government revenue dataset: still the best option for researchers

Wilson Prichard and Kyle McNabb

In 2010 the ICTD launched efforts to create the ICTD government revenue dataset (GRD), which is increasingly recognized as the best possible source of cross-country revenue data for researchers. An important motivation was concern about the quality and transparency of data available from the IMF: Publicly available data had significant limitations, while the private data used in much IMF research was not available to other researchers, and appeared to contain significant errors.

The immediate goal of the ICTD GRD was thus to provide better and more transparent data to researchers. That goal was achieved with the launch of the data in September 2014. The longer-term goal has been to encourage international organizations, led by the IMF, to invest in higher-quality revenue data and to make that data publicly available. We have recently seen important progress, most recently from the IMF, who in August released the World Revenue Longitudinal Data (WoRLD) dataset, making public the revenue data used internally by IMF researchers – something the ICTD and others have long advocated for. The WoRLD dataset employs a similar methodology to the ICTD GRD, merging data from multiple international sources in order to improve data coverage. Its public release marks an important and necessary step toward greater transparency.

However, as we outline below, the ICTD GRD remains a much more complete and higher quality source of data for most researchers. This brief note summarizes key differences between the two data sets, and the significant advantages of the ICTD GRD.

Sources and Coverage

The IMF WoRLD relies exclusively on data from the IMF World Economic Outlook (WEO), the IMF Government Finance Statistics (GFS), the OECD Revenue Statistics, and the OECD Revenue Statistics in Latin America. Notably, tax data from the WEO is not itself publicly available: As such it is not considered for inclusion in the ICTD GRD. By contrast, the ICTD employs all of the same sources, other than the WEO, but also draws on IMF Article IV Reports and data from CEPAL (*Comisión Económica para América Latina y el Caribe)* in Latin America.

The incorporation of data from IMF Article IV reports in the ICTD GRD has been particularly complex: it has been necessary to manually extract from original reports, while inconsistent categorization in those reports has required extreme care in merging the data. However, the benefits of this effort have been significant, as incorporating data from IMF Article IV reports allows the ICTD GRD to achieve significantly greater data coverage than the IMF WoRLD. This is illustrated in Table 1, which highlights that there is better data coverage for every revenue and tax subcategory in the GRD between 1990-2013.[1] [2]

[1] Notably, the GRD also has very good coverage as far back as 1980 (although for the sake of comparability, this is not included in Table 1).

Table 1. *Data coverage: WoRLD vs. GRD*

	IMF *WoRLD*		**ICTD *GRD***	
	#	% of total	#	% of total
Countries	186	-	192	-
Total Obsv.	4464	-	4632	-
Revenue	3683	83%	3859	83%
Tax	3264	73%	4058	88%
Income Tax	2375	53%	3527	76%
PIT	2326	52%	2790	60%
CIT	2534	57%	2820	61%
Payroll	723	16%	2381	51%
Property	2024	45%	2925	63%
Goods & Services	1327	30%	3517	76%
VAT	1772	40%	*2422	52%
Excises	2506	56%	2859	62%
Trade	1302	29%	3543	76%
Other	?	?	3331	72%
Social	1814	41%	3276	71%
Grants	2102	47%	2975	64%
Nontax	?	?	3836	83%

*Includes both VAT & other Sales tax.

Data Merging, Data Cleaning and Accuracy

A more important difference lies in the approach to merging data from multiple sources. The IMF WoRLD dataset is based on an automatic merging algorithm, which combines data from different sources based, for the time being at least, on a strict hierarchy of preferences, as follows:

1) For all countries the total revenue figure is taken from the WEO
2) For OECD and Latin American countries, all tax data is from the OECD Revenue Statistics
3) For all other countries total tax revenue is from the WEO, while sub-categories of taxation are from the IMF GFS
4) Because the IMF GFS has multiple data series, General Government data is preferred, with Central or Budgetary Central Government data adopted where General Government data is not available.

In principle such an automated algorithm is highly attractive, as it is both quick and objective. However, *in practice* this approach will only yield fully valid data *if* the raw data in different international sources is mutually compatible. Unfortunately, this assumption does not hold in practice, creating problematic breaks in the data, which require manual cleaning by users – and further reduces data coverage.

Figure 1, which displays total revenue and total tax for South Africa in 2012, illustrates that figures reported by the WEO (used in the WoRLD) often do not match those found in the IMF GFS (used in the GRD), with the result that the merging of these two sources for any single country-year can be misleading. In this case, it is clear that the WEO data is significantly lower than that reported in the

[2] NB. Not every subcategory will have the potential for 100% coverage (For example, not every country has a Payroll or Property tax).

GFS, while the wider question of why WEO tax and revenue data is preferred to GFS (when the latter reports both very consistently) is unclear.[3]

Figure 1. *South Africa, 2012; WoRLD vs. GRD*

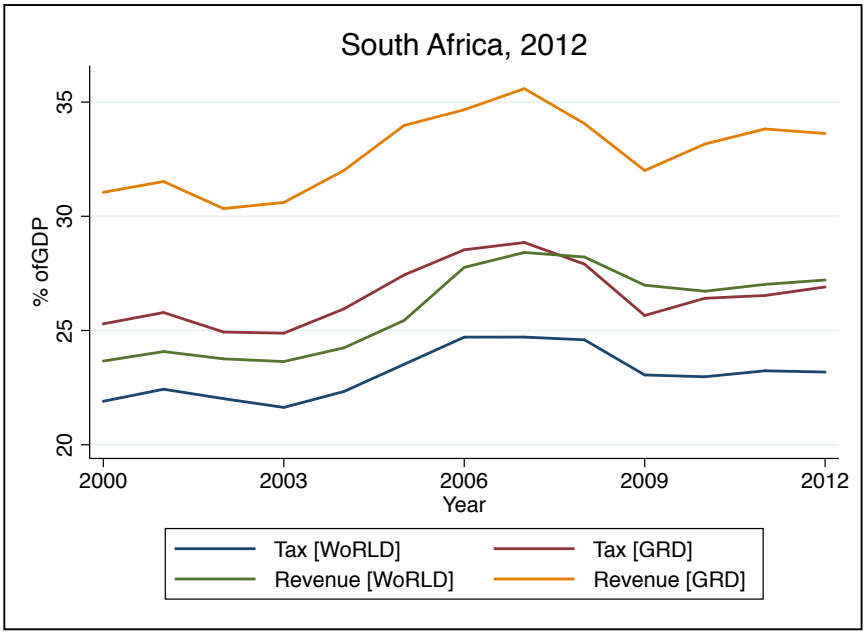


Table 2. *South Africa, 2012; WoRLD vs. GRD*

Category	**WoRLD**	(Source)	**GRD**	(Source)
Revenue	27.2	WEO	33.6	GFS
Tax	**23.2**	WEO	**26.9**	GFS
Income	?	GFS	14.0	GFS
PIT	8.5	GFS	8.5	GFS
CIT	5.5	GFS	5.5	GFS
Payroll	?	GFS	0.3	GFS
Property	?	GFS	1.4	GFS
Goods	?	GFS	10.0	GFS
General	6.6	GFS	6.6	GFS
VAT	6.6	GFS	-	GFS
Excises	2.2	GFS	2.2	GFS
Trade	?	GFS	1.2	GFS
Social Contrib.	0.6	GFS	0.6	GFS
Nontax	?	GFS	6.7	GFS
Grants	0.1	GFS	0.1	GFS
Sum of (available) Tax components:	**22.8**		**26.9**	

NB. Figures displayed are % of GDP

Table 2 takes a closer look at this particular case, and further highlights the problems with taking data from two different sources for the same country-year. The GRD (on the right) takes all its data from the GFS, with the result that the sum of the tax subcomponents is equal to total tax collection. The same cannot, however, be said

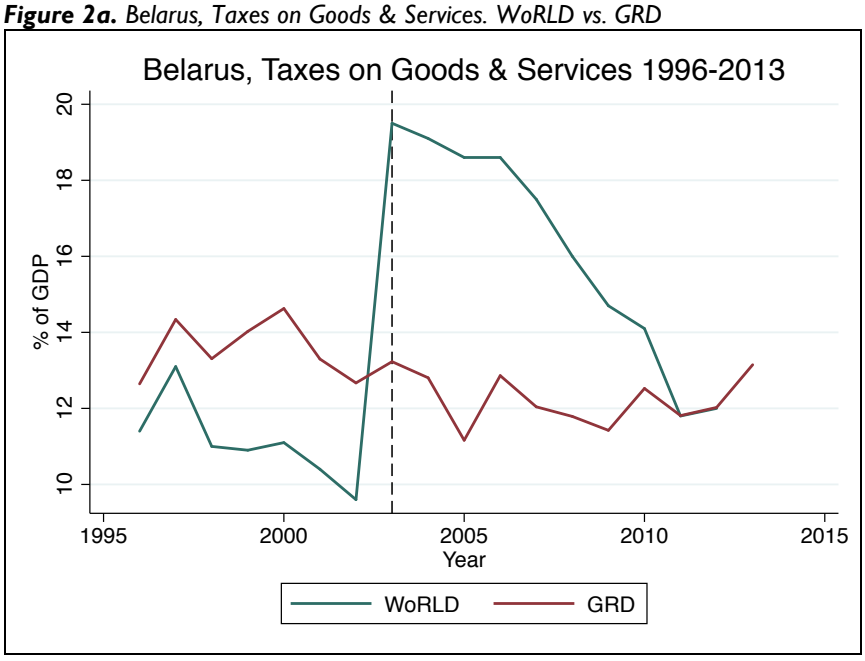
[3] A similar problem emerges with the fact that the WoRLD does not report total revenue excluding grants (and, indeed, that the grants figure comes from a different source). The total revenue figure given is thus potentially problematic because grants, which are hugely variable across countries in how they are recorded, can be hugely volatile and are thus almost certainly misleading at times.

for the WoRLD. Indeed, we cannot hope to understand how the WEO total tax figure has been computed without details on each of the subcomponents.

This is, of course, only one example, and whilst there are a good number of cases where the WEO and GFS tax figures line up almost exactly, it appears more common that the two sources are out of line. For the subset of country-year observations where both datasets had figures, the average difference in the total revenue figure was 3.4% of GDP.

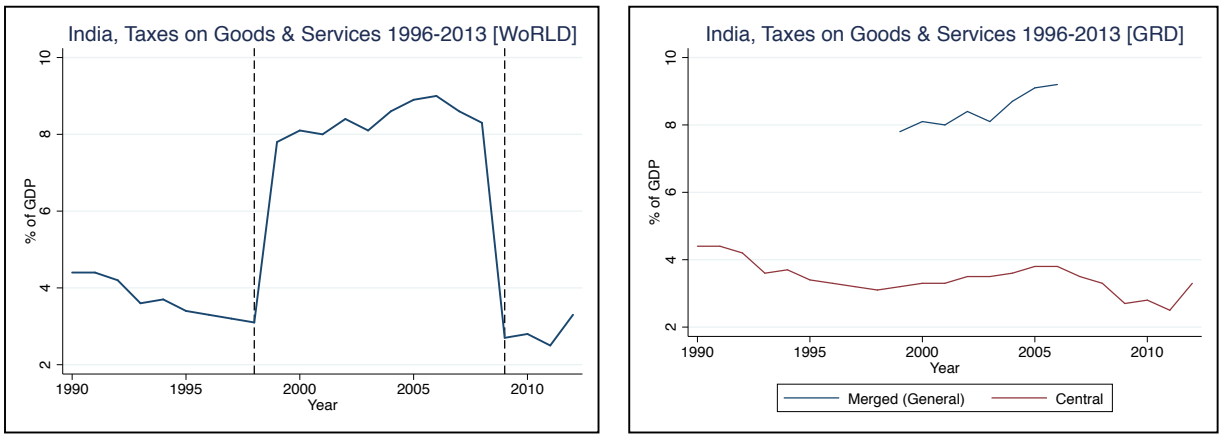
Figures 2a and 2b illustrate a larger problem, where data from the IMF GFS shifts from central to general government, again creating a misleading break in the data. That is, the change in data source creates the appearance of a major change in tax collection, whereas in practice no such change has occurred. The example considered in figure 2a is for taxes on goods and services in Belarus. The GRD employs IMF Article IV data until 2011, at which point it switches, seamlessly, to GFS data resulting in a fairly consistent series. The WoRLD, however, uses GFS data throughout. At 2003 (marked) it switches from central to general, and the resulting problem is clear; to imply an increase of over 100% in taxes on goods and services is not only wildly unbelievable, but also carries serious implications for any econometric research that might wish to employ this data.

Figure 2a. *Belarus, Taxes on Goods & Services. WoRLD vs. GRD*



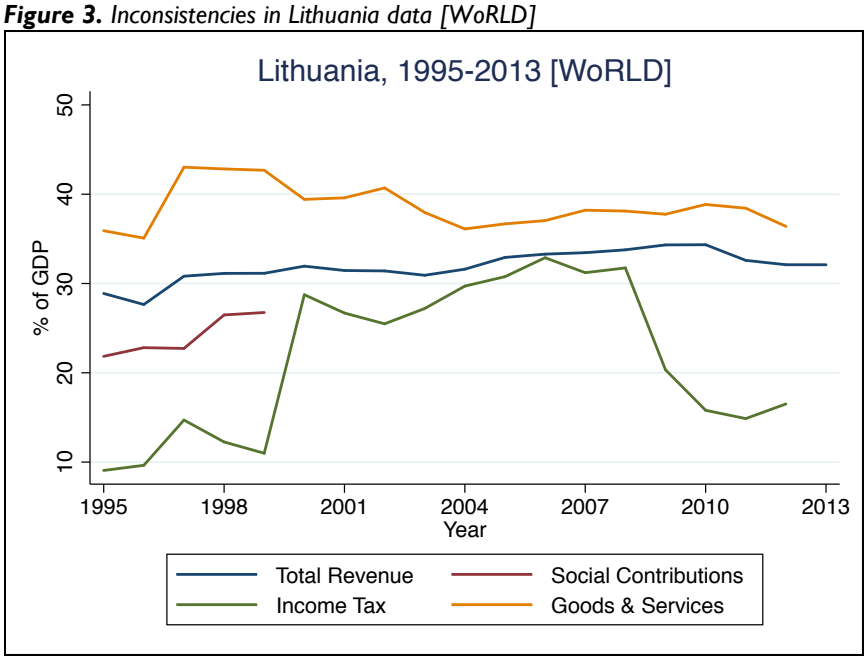
Turning to Figure 2b, we see a different kind of example. The case is again for taxes on goods and services, this time for India. The left panel shows the approach taken in the WoRLD, which shows a precipitous jump and subsequent fall for the years 1998 - 2009, when General data is used. The GRD offers two solutions here. Its 'merged' dataset (where general is usually preferred, if available) includes only those years for which general data is available. Alternatively, a separate Central dataset is available, which includes a consistent series for central government data. Both GRD options avoid the massive jump that is imposed in the WoRLD due to the automated approach of data merging.

Figure 2b. *Comparison of GRD & WoRLD approaches*



Finally, Figure 3 illustrates the largest and simplest problem: In a few cases the data entered is simply incorrect, but the absence of fine toothed manual cleaning of the data has allowed the errors to persist. The figures for Lithuania show, amongst other errors, taxes on goods and services to be greater than total revenue. Such simple problems are not unheard of for new datasets, and will undoubtedly be corrected in subsequent rounds. However, the existence of the error highlights the importance of transparency in weeding out errors.

Figure 3. *Inconsistencies in Lithuania data [WoRLD]*



What solution does the ICTD GRD offer to these issues? There are three key elements:

1) It selects a single source for every country year combination, thus reducing potential incompatibility between total revenue figures, total tax figures and more disaggregated tax figures.

2) The data merging for the ICTD GRD has been done *manually*, with every data point reviewed *at least twice*, in order to identify both inherently problematic data, and discontinuities between data sources.
3) This approach is what allows the ICTD GRD to make use of IMF Article IV reports, which were all entered manually.

In the presence of imperfect underlying sources manual data cleaning is absolutely imperative to data quality, and only the ICTD GRD has pursued such a process completely and transparently for all countries. The above examples illustrate that without such due care and attention, major inconsistencies and problems can persist in the data.

Non-Tax Revenue

The IMF WoRLD dataset is also characterized by one key omission, as it does not include a category for total non-tax revenue. Conceptually, this is potentially problematic: Understanding the fiscal reality of a country requires an understanding of *all* revenue sources, as they are likely to influence each other. Most obviously, access to non-tax revenue is likely to influence choices about how much tax revenue to collect. Recent studies have made clear that it is essential to include a measure of total non-tax revenue in any econometric analysis that seeks to explain total tax collection. Yet the absence of a non-tax revenue variable in the IMF WoRLD makes this impossible.

Intuitively, this should be easy to overcome: Subtracting Total Tax, Grants and Social Contributions from Total Revenue would, normally, leave us with total non-tax revenue. However, this simple operation is only strictly valid if all figures come from the same source, or sources that are mutually compatible. As discussed above, this is not always the case for the IMF WoRLD. Where the total revenue variable is taken from the WEO, but the total tax variables are taken from an alternative source, there is risk of significant inaccuracy if non-tax revenue is simply calculated as the difference between total revenue and total tax revenue. The resultant value *may* accurately reflect total non-tax revenue, but equally may reflect differences between the two sources, thus driving misleading analysis.

Resource Revenues

A closely related concern, and perhaps the single most important difference between the two data sets, is that the ICTD GRD seeks systematically to distinguish between natural resource revenues (oil, mining) and all other tax revenues – primarily by drawing on IMF Article IV reports – while the IMF WoRLD does not do so. This sometimes has major consequences, as international sources are inconsistent in whether natural resource revenues are recorded as tax or nontax revenue for major resource producing countries. While this categorization is often accurate in strictly accounting terms, it can generate highly misleading data for analytical purposes.

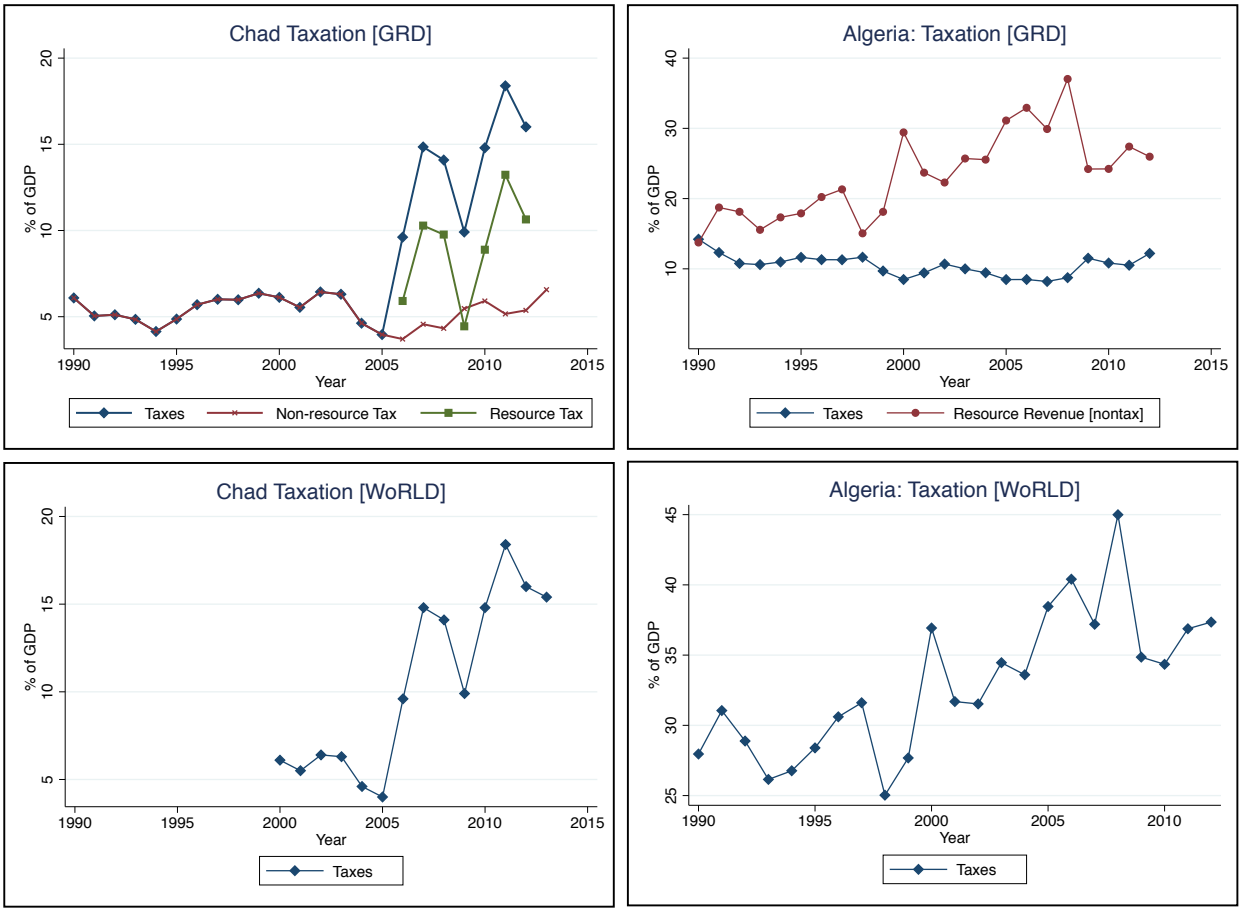
Within the IMF WoRLD data some major oil producers appear among the highest tax collectors in the world (e.g. Angola, Algeria), whilst others report exceptionally low tax collection, with very high non-tax revenue (e.g. Bahrain, Iraq). For many other countries, tax collection varies significantly depending on how resource

revenues are categorized. This likely invalidates any econometric research using the full IMF WoRLD data. Inconsistency in treatment across countries implies dramatic measurement error. Meanwhile, researchers are generally most interested in understanding either non-resource tax revenue *or* resource revenue, but the IMF data prevents a clear and consistent distinction.

The approach to this problem adopted by the ICTD is two fold. First, where possible, it distinguishes between the resource and non-resource components of both tax revenue and non-tax revenue. This allows users to measure total tax collection as recorded by the underlying international sources, but also, more critically, to measure total *non-resource tax revenue*, which is both consistent across countries and the variable most commonly of interest to tax researchers. Second, where necessary the ICTD adopts a second best approach, recording only *non-resource tax revenue* under the tax revenue categories, while recording all resource revenues as non-tax revenue. This is not strictly accurate from an accounting perspective but, critically, is much more accurate analytically for most types of tax research, as it allows consistent comparison of non-resource tax revenue across countries.

Whilst the outcome is thus not perfect – and is limited by the quality of underlying data from IMF country reports – it seems clearly preferable to not addressing the problem at all. Figure 4 illustrates two examples. For Chad, which first reports resource tax revenues in 2006, we see a large jump in the WoRLD figure for tax revenue. The GRD, however, is able to break resource tax revenues out of total tax and provide a consistent non-resource tax series. Algeria represents a different manifestation of the problem. Resource revenues are treated as tax in the WoRLD, but as nontax in the GRD. The result is that Algerian tax figures appear heavily inflated and volatile, when in truth this is due almost exclusively to hydrocarbon revenues. By contrast, the ICTD is able to construct a consistent series for non-resource tax, which is more consistent and credible over time for most research purposes.

Figure 4. *Classification of Resource Revenues: WoRLD vs GRD*



Clearer documentation

A key goal of the ICTD GRD has been to ensure a high degree of transparency and documentation about the construction and use of the data. Every data point is attributed to a specific source, all data choices are carefully documented, potentially problematic data are explicitly flagged and a detailed Working Paper was released with the data, including discussion of limitations and appropriate use.

In some respects, the IMF WoRLD dataset follows the same path, to its significant credit. Every data point is attributed explicitly to its underlying source, thus allowing users to understand the origin of the data and to check for potential problems. Meanwhile, although the automated merging of data has important weaknesses, it does minimally ensure transparency about data choices.

However, there remain two major concerns. First, the creators of the data set have yet to release an accompanying working paper and user guide. Most critically, such a document should transparently describe the limitations of the data – documented here - and offer guidelines about how it can, but also cannot, be used for research purposes. Without such guidance the likelihood that researchers will simply download the data uncritically, without accounting for the limitations described here, seems very large. The result would be misleading research findings.

Second, and more curiously, the WEO tax data that is widely used in the WoRLD dataset is *not* publicly available through the WEO. The creators of the WoRLD dataset should be congratulated for making this data public for the first time, but still better would be to also have access to the original source and information surrounding its construction.

Summing Up

Despite these weaknesses, the IMF WoRLD dataset has a specific niche: it is an effort to create a composite dataset based on the primary international databases that is transparent, full automated and free of discretion. It has resulted in tax data from the WEO being made public for the first time. Used carefully, the WoRLD dataset may be useful in producing country specific, regional and global trend data about revenue collection, and for some limited types of econometric analysis.

However, it comes with very significant limitations. The data will only be truly reliable for any kind of analysis if researchers deal carefully with resource producing states, and if researchers carefully, and manually, clean the data of existing breaks and errors. Even then the failure to address the question of resource revenues, along with continued missing data, imposes major limits. Meanwhile, the absence of clear documentation and guidance for using the data suggests a tremendously high risk that the data will be misused.

As such, the ICTD GRD remains by a significant margin the best available resource for almost all types of tax and revenue research and analysis. There are, naturally, some risks to the manual merging and cleaning of data behind the ICTD GRD, while data on resource revenues remains imperfect. However, dealing imperfectly with the weaknesses of international data remains far preferable to any other alternative. The continued transparency of the ICTD GRD, and its increasingly wide adoption, has served as an invaluable confirmation of its broad quality and accuracy.

That said, there remains much scope for collaboration moving forward. Over the long-term, the IMF remains uniquely positioned to collect high-quality government revenue data. Most importantly, there is an urgent need for high quality and publicly available data on resource revenues. The release of the WoRLD dataset appears to signal an increased commitment to research data and to data transparency, and is hopefully only the first step in this process.