

Drowning in Spreadsheets

Is your business forming a spreadsheet daisy chain?

James Adman discusses some of the common problems of spreadsheet proliferation.



Background

Spreadsheet applications such as Microsoft Excel are widely used within most organisations and with good reason. They're powerful, easy to use, and they allow users to perform calculations quickly and easily. Some users within a business will quickly become adept at manipulating large data extracts in Excel, often deriving new fields, pivoting data, aggregating data, even producing charts and dashboards. "All good", you might say.

Unfortunately this power comes at a cost and it's not always immediately obvious that the cost is there. The benefits are felt straight away, but the disadvantages take time to build up. The aim of this article is not to discourage you from using spreadsheet applications like Excel, but to encourage you to adopt practices that will allow you to use these tools safely.

So let's consider the lifecycle of a typical spreadsheet.

1. Obtain Data

Let's say you need to produce a report which includes some calculated measures. You know the data you need and you find out where you can get a copy. It makes sense for you to keep this source data in the spreadsheet, if only so that someone else can trace back from your calculations to the source figures you used.

However, at that point you have created a new copy of that source data and this must be done with extreme caution. When someone comes back to the report months or years from now, is it clear exactly when that data extract was taken, which criteria were used to extract the data or whether any data values have been "corrected".

At this stage, you have taken ownership of a copy of data and you are responsible for that copy, its security, its accuracy, its metadata and its distribution. Unwittingly, you have made yourself into a data steward.

If you have taken ownership of a copy of data, then you are responsible for that copy, its security, its accuracy, its metadata and its distribution.

2. Perform Calculations

The next stage is the creation of a report. The calculations for the report may be performed within separate worksheets, within additional columns in the source data worksheet, or even applied to the source figures themselves.

The problem here is that it may not be clear which data is derived and which is the source data. If you change the values in the source data sheet, it's not the source data anymore, but it looks like it is! If you add extra columns of derived data to the same worksheet, it could still be confusing as to where that data has come from. If values are pasted instead of the formulae, then the calculations are lost.

3. Publish

The next stage is publication. Often this isn't the spreadsheet itself, it may be charts and tables inserted into a presentation or a website. At this point, the spreadsheet and the report have become disconnected.

At the point of publication, the results of your work are visible and potentially scrutinised. If someone questions your figures, will you be able to trace back to the same version of the spreadsheet that was used to produce the report? Can you be sure it hasn't been changed?

4. Distribute

Publication typically brings additional interest. At this point, other people may want to start using your figures for their own reports. People typically look for the easiest way to get the job done, and if there's potential to re-use your work, people will take it. This

becomes yet another copy of source data, and we return to step 1. Except that it isn't the source data anymore!

This can happen again and again. Before long you have a spreadsheet daisy chain, which is difficult to unravel and would alarm anyone who is interested in maintaining data quality and data lineage.

The Risks

If this example spreadsheet lifecycle sounds familiar, you are definitely not alone. In fact, it is so familiar that many people just try to live with it. The aim of this article is to convince you not to live with it and to make a change.

Managing data in this way will eventually lead to poor quality data in reports. Depending on the audience of the reports, the implication of poor data quality may be: poor business decisions, loss of credibility, legal compliance issues and possible financial or legal penalties for breaching regulations. If these issues lead to an investigation of data management practises, the spreadsheet daisy chain is going to be hard to defend.

The spreadsheet also creates a very inefficient chain of data propagation. Each person in the chain will be dependent upon the previous people. Any issues identified need to be passed back along the chain.

Furthermore, by keeping all the raw data in your spreadsheet, you have far more data stored locally than you need. With the continuous stream of information security failings published in the press, can you defend why your local laptop has a spreadsheet containing all the low level data, when all you needed to publish were some high level KPIs?

Over time, departments become more and more dependent upon spreadsheets. Before long you have little departmental "Cottage Industries" producing spreadsheet applications, often completely outside the governance of corporate application development strategies. These spreadsheet applications will inevitably need support and enhancement. You may end up with applications which have a total cost of ownership that was never budgeted for.

Small Steps

I have seen these kinds of problems in organisations which have very different levels of maturity of their information management strategy. Some may have strategic Business Intelligence systems, but may have trouble with end user adoption. They know the reporting tools are there, but they don't really know how to use them so they just export the data to Excel instead. Some may have no centralised BI system, and the only copy of their data is in spreadsheets. These require different long term approaches, but in the short term both cases could be improved by some simple, quick win strategies. These will help reduce the risks of spreadsheet misuse and are small steps that will simplify the process of tackling a longer term solution.

1. Go back to the source. Identify the most trusted source of the truth, not someone's local copy of that data. If there is no trusted source system, identify the data owner and obtain the data from them.
2. If you have a reporting environment that allows you to write your own reports, make use of it. It will make your work far more visible and reusable and will avoid going down the "Cottage Industry" path. Don't try to take a short cut and just dump the source data to Excel.
3. If you must extract data to Excel, make sure you record the metadata regarding your extract and don't change the values. Make it a protected worksheet and always retain the information regarding the "who", "what" and "when" of the extract. It should be possible for someone else to use your metadata to obtain the same data from the source.
4. If you need to correct data, try to get it corrected at source. If you really need to do any data cleansing, retain the original data extract and always record what you have done.

Spreadsheet applications will inevitably need support and enhancement. You may end up with spreadsheet applications which have a total cost of ownership that was never budgeted for.

5. If you have calculations, make sure they are properly annotated and tested. It will provide a record of what you have done and it will certainly make it easier for either yourself or someone else to understand.
6. If you distribute your reports, make sure you label the distributed version and the spreadsheet that was used to produce the report and keep copies in a safe place. Follow your current document management procedures; if you don't have procedures, ensure you keep a labelled copy somewhere safe. You should always be able to identify which version of the spreadsheet you used to create a report.
7. Don't re-issue your local copy of the source data. Redirect data requests to where you obtained your data.

Long Term

These small steps are a tactical solution to help you to manage the spreadsheet problem in the short term. Longer term, a proper information management strategy needs to be defined. The problems described above lead to multiple version of the truth. A well defined corporate Business Intelligence solution provides a single source of the truth, and exposes this information in a way that allows people to make use of it.

Defining a corporate information management strategy requires more detail than can be covered in this article, especially when

you take into consideration the many different starting points. However, it is important to recognise that, when defining ways to manage the proliferation of spreadsheets and their data, this will not fix the problem on its own. However, it will help you to control the problem.

Conclusions

Spreadsheets are here to stay and rightly so. They offer a variety of data manipulation and presentation tools at your fingertips

and there's no need to stop using them. With well defined procedures, spreadsheets can add considerable value to a business. However, if left uncontrolled, you might end up drowning in a sea of spreadsheets.



James Adman is an Information Management Consultant at IPL, a leading UK IT services company specialising in the delivery of intelligent business solutions. He has over 10 years experience helping a wide range of high profile clients exploit the full potential of their information.

James has significant information management expertise gained across a variety of sectors including manufacturing, finance, government, transport, emergency services and petrochemical. James has a number of ongoing engagements in Business Intelligence including system procurement, specification and design. He can be reached at james.adman@ipl.com.