SEEDING THE AFRICAN DATA INITIATIVE

DAVID STERN

University of Reading

d.a.stern@reading.ac.uk

The African Data Initiative (ADI) is a highly collaborative project that aims to transform statistics education and how people use and understand data, both in Africa and beyond. The first major activity of ADI has been the development of R-Instat, a front-end to R, tailored to African needs and developed largely in Africa. This paper describes the background, initial activities and the principles of ADI. The principles provide structure to guide and communicate thinking behind ADI decision making, for both existing and future activities. The ADI collaboration exists primarily through a common desire to contribute towards Africa's data revolution alongside a collective principal based approach.

INTRODUCTION

The African Data Initiative (ADI) was instigated through crowd sourcing in mid-2015 (AMI, 2015). It succeeded in raising the funds required for the first ADI component, which consists of filling the gap in statistics software identified in the campaign. It also succeeded in that there were over 400 people who contributed to the initiative, over half of whom were from Africa, Fig. 1, where the gap had been identified.

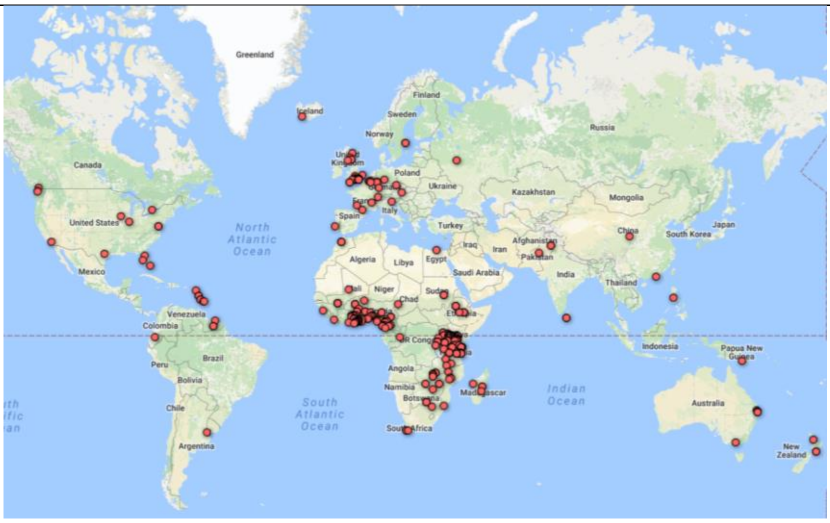

Fig. 1 Contributors to the African Data Initiative

The case was made for the development to be based within Africa and the work has been largely at AMI (African Mathematics Initiative) in Kenya and at AIMS (African Institute for the Mathematical Sciences) Tanzania, with mentoring and support from staff based in Reading and Oxford, UK. The software being produced is only "new" in its appearance. It is essentially a menu-driven front-end to the R statistics software environment.

This paper starts by outlining key aspects of the background to ADI including the motivation. It then presents the underlying principles guiding ADI and what they imply in practice. Finally the key activities of phase one and the transition to subsequent phases are described.

BACKGROUND TO ADI

This project is a collaboration primarily involving AMI, AIMS, SSC (Statistical Services Center), SSD (Statistics for Sustainable Development) and Maseno University. It is centered on solving the problem of statistical literacy and education tailored to African environments and in the context of the substantial progress in statistical methods (Stern, Coe, & Stern, 2017). Direct attempts at solving this problem have failed, which is perhaps because it really is a "wicked problem" (Conklin, 2006).

Aspects of the collaboration have a long history (Musyoka, Stern, & Stern, 2014) and the approach taken was patient with incremental modernization of teaching and curriculum (Stern, Ongati, Agure, & Ogange, 2010). Substantive efforts have been put into reforms in MSc teaching (Stern, Coe, Stern, & McDermott, 2013) but there has not yet been a transformative effect, possibly because the situation was too complex for a single direct solution.

Reflections on why this and other initiatives have not had the desired impact led to the conclusion that a different approach was needed (Stern D. , 2014). This alternative approach is to build change that cuts across academic levels and disciplines while spreading through viral-scaling, described in the section below. The ADI project has grown from these reflections. The first phase creates potential sparks to start the transformation, while subsequent phases will support the spreading of reform.

PRINCIPLES

ADI has been conceived as a "principles-based initiative" (Patton, 2017). This recent approach to evaluation has proven to be invaluable in other projects tackling complex problems. Using principles can help to make coherent justifiable decisions and the communications of principles can help to build communality between project partners. This section describes some of the core ADI principles and explains how they have been used in decision making. For communication purposes we present them using 'catch phrases':

- Viral Scaling,
- Open by Default,
- Making it Easy,
- Options by Context,
- All Academic Levels,
- Incremental Modernisation,
- Good Statistical Practice.

Each of these phrases offers practical insights when taken as a principle.

Viral Scaling

ADI has grown from reflections on past experience (Stern D. , 2014) which proposed the opportunity of 'Viral Scaling' as a route to sustainable impact. When working with potentially dynamic institutions a key point is that achieving sustainability within one can be harder that scaling across multiple similar institutions. So scaling can be a mechanism to increase the sustainability as well as the extent of the impact.

For viral scaling the innovation needs to be able to spread independently, being passed from one user directly to another. An initiative can only go viral if an average user instigates more than one new user. Keeping this in mind has pushed R-Instat towards many of the other principles. It needs to be free to distribute (Open by Default), to be thought of as part of a set of resources appropriate for multiple audiences (Options by Context) which can stand alone and be used without needing specific training (Making it Easy).

Open by Default

The principle of Open by Default is that no justification is needed to go open. Not going open is also fine as long as there is a justifiable reason. The decision for R-Instat to be open source was easy and obvious. Building from R we are "standing on the shoulders of giants" and the intention is for R-Instat to enable more people to make use of R.

Harder decisions concerned choices of development tools and which platform to develop for. There was a strong desire to build from existing open source R interfaces, particularly R-Studio, which would have led to a cross platform solution. However, the aim of having the development done substantially in Africa was important. Development in VB.Net using Visual Studio has enabled us to produce a stable product while building the team. Other alternatives had a learning curve that was too steep.

R-Instat is not intended as an open replacement for the standard commercial statistics packages. In resource rich environments, students should ideally become familiar with more than one package and R-Instat might sometimes be a useful addition to the mix. In resource poor environments R-Instat combined with R-Studio is designed to be sufficient for student needs.

Making it easy

R-Instat is designed to be as easy as possible for users to engage with. It should be accessible for anyone who has experience with the use of a spreadsheet package. This aim has driven our plan to make the individual R-Instat dialogs easy to use. When more flexibility is needed this is usually through a sub-dialog.

Developing the graphics system for R-Instat has been particularly time-consuming. We believe strongly in the grammar of graphics (Wilkinson, 2005) and in its implementation in R, ggplot2 (Wickham, 2010). However, even some experienced R users find ggplot2 difficult to master and shy away from it. Our challenge was to make the graphics dialogs easy and intuitive while also giving access to the immense flexibility of the ggplot2 system.

A second aspect is to make it as easy as possible for educators to use interesting data in their teaching. Having access to a rich library of datasets in R-Instat is critical and is helped considerably by the rich variety of datasets provided in R-packages.

Options by Context

The idea and phrase "Options by Context" has come from research methods work related to agricultural research (SSC, 2015). The phrase has influenced the look and feel of the software substantially as we recognize that not all target audiences will use R-Instat in the same way. This has, in particular, influenced the decision to include tailored menus for specific audiences.

Options by context also applies to our ideas for educational initiatives. This leads to an expectation for educational resources to be used in different ways, at different institutions, by different types of learners. This combines well with the open by default principle and implies that resources created should be editable where possible, so they can be adapted by users to their own contexts.

All Academic Levels

The idea that educational reform can be more effective when considered across all academic levels resulted from work in post graduate education which led to opportunities for that reform to spread to other levels (Stern D. , 2013). From basic statistical literacy at school to PhD students, the aim, partly through ADI, is to spark reforms which lead to substantial impact on the understanding and use of data.

The first advantage from this approach is because many of the people trained in African postgraduate education are involved in education at undergraduate or school-level. Hence if the reforms they experience can translate into the classes they teach they are more likely to become part of the reform process. The second advantage is that reform across academic levels encourages the cross-fertilization of materials and pedagogies. This is particularly important in making the teaching of basic material more exciting and interesting.

Incremental Modernisation

In many universities statistics courses need to change fundamentally to produce genuinely useful graduates. However, wholesale changes are often slow and contentious to implement. There is a risk that attempting large changes might meet with sufficient resistance that nothing changes.

In contrast, small changes, starting within a single course, can be useful as well as being a precursor to more substantive change. These small changes are often possible within the current syllabus, so changes in the curriculum may follow initial changes in the content or in the style of teaching. Some could even be started within a mathematics or statistics "club", as an extra-curricular activity, rather than in a formal course.

The process of change through incremental improvements is potentially a cultural shift within African universities which could be a 'win-win' for both staff and students. If staff find ways to improve student educational outcomes, document their innovations and publish the results, then students have a better education, while the staff can use the resulting publications for their reputation and promotion. This could lead to a cultural shift whereby educational innovation is institutionally supported and valued.

Good Statistical Practice

At the heart of ADI is the desire to promote and encourage good statistical practice. Within R-Instat, decisions ranging from the menu structure to the choice of R packages are designed with this principle in mind. Although there is agreement on aspects of good statistical practice, such as the inclusion of descriptive comments in scripts or moving beyond averages (Ross, 2016), it is not always well defined and even within the development team this had led to rich and interesting discussions.

The launch of R-Instat is designed to open up the discussion on what constitutes good statistical practice in different contexts and how this might translate into the software. R is already excellent at encouraging such rich discussions and the intention with R-Instat is to conclude the discussions with an implementation of the 'good current practice'.

PROGRESS SO FAR

The first phase of ADI started with the formation of a team responsible for development of R-Instat. The core team consisted of bright graduates rather than experienced "programmers". This was based on the hypothesis that with the right skills and guidance, mathematical science graduates can do things that will substantively impact development. This team has working together for almost two years. In the first year some members of the team balanced their role for ADI with being part-time AIMS tutors while others in the team were full time on ADI work. In the second year the core team shrunk and all members now worked full-time. New members were added opportunistically to get back to a similar size of team.

Throughout the process the core team was based in Africa, with support from a small group based in the UK. The UK support was largely funded through other sources and has provided much of the experience and mentorship that has guided the project. After the first year the main team could contribute dialogs and code to the software on the front end interface. The second year has been one of substantial progress where the core team has contributed more deeply to the code, for example in their use of R and the construction of common user controls. The training of the team through this mentorship has been an important deliverable.

In 2016 two small additional grants were obtained which has enabled the continued development through the second year. The first was for climatic components to be included as a ‘tailored menu’ in the software. This was appropriate, because the philosophy of the software, built on an existing small package, called Instat. This was originally developed in the 1980s, with a Windows version appearing in 2001. It was designed to be simple-to-use while encouraging good statistical practice. It also had an extra customized menu specifically for the analyses of historical climatic data. Although no line of code from Instat was used the parallels are sufficient that the new package is called R-Instat.

The second addition related to an objective methodology to assess the degree of corruption risks in government contracting in countries in the EU (Fazekas, 2016). A grant through the University of Oxford was provided to assess the extent to which this methodology could be used in Africa. The analyses by (Fazekas, 2016) made use of Stata, and part of the grant was to investigate the extent to which the analyses could also be implemented as a ‘tailored menu’ in R-Instat.

The R-Instat system has been designed to be as modular and hence as consistent as possible. This adds complexity to the core code but provides stability to dialog functionality which enables automated creation of R code by the front end. We have also created substantive R code, particularly related to the management and linking of data frames (Parsons, Stern, & Stern, 2017).

R-Instat version 0.9 (July 2017) is the first distributed release. The resulting software, even with its remaining imperfections, should enable change for our primary African education audiences, for whom R-Instat has been designed. At the same time initial testers suggest that the software may have broader appeal both within Africa and internationally.

If R-Instat proves its potential it may transition into a further phase of development that is more aligned with existing open source communities. The new phase could build from RStudio’s existing code base to re-implement R-Instat ideas as a stable multi-platform application. This would, in particular, allow the use of RStudio’s link to R which is more stable and flexible than the current system. Ideally this development process would continue to be a genuine collaboration in which African partners play a substantial role.

CONCLUSIONS

The first phase of ADI has been about filling a gap in resources. Alongside the efforts to fill the gap there have been ongoing efforts to find effective uses of the resources across academic levels, from schools (Mbasu, Mawora, & Stern, 2017) to Universities (Musyoka, Stern, & Stern, 2017). These initiatives build from the growing body of experience and are contributing towards creating easier mechanisms of implementation.

This creates a position where in the next phase it may be possible to take the current open resources to scale through small interventions across large numbers of African institutions. This led to the title of the paper and the belief that with the current open educational resource in place, reforms could flourish, given just a shower of support.

ADI is transitioning from a small well defined collaborative project to an ambitious collection of activities across multiple partners. The principles presented in this paper are the fibers which keep the set of activities connected with the expectation that the whole can achieve more than the sum of its parts.

REFERENCES

AMI. (2015). *African Data Initiative Campaign*. Retrieved from Chuffed: https://chuffed.org/project/africandatainitiative

Conklin, J. (2006). *Dialogue mapping : building shared understanding of wicked problems.* Chichester, England: Wiley Publishing.

Fazekas, M. T. (2016). An Objective Corruption Risk Index Using Public Procurement Data. *European Journal on Criminal Policy and Research*, 369-397.

Musyoka, J., Stern, D., & Stern, R. (2014). Building strength from compromise: a case study of five year collaboration between the Statistical Services Centre of the University of Reading, UK, and Maseno University, Kenya. *ICOTS 9.* Flagstaff: IASE.

Parsons, D., Stern, D., & Stern, R. (2017). Making Multilevel Data Ideas More Accessible. *Proceedings of the IASE Satellite Conference "Teaching Statistics in a Data Rich World".* Rabat, Morocco.

Patton, M. (2017). *Principles-Focused Evaluation.* New York: Guilford Press.

Ross, T. (2016). *The End of Average: How we Succeed in a World that Values Sameness.* San Francisco, : HarperOne.

SSC. (2015). *Options by Context playlist.* Retrieved from YouTube: https://www.youtube.com/playlist?list=PLg1i766TIIKakjZ30fuRJ_vxY9lhCfcs4

Stern, D. (2013). Developing Statistics Education in Kenya Through Technological Innovations at all Academic Levels. *Technology Innovations in Statistics Education, 7*(2).

Stern, D. (2014). Reflections on using technology to teach statistics in Kenya. *ICOTS 9.* Flagstaff. Retrieved from http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_3D2_STERN.pdf

Stern, D., Ongati, O., Agure, j., & Ogange, B. (2010). Incremental modernisation of statistics teaching and curriculum at Maseno University, Kenya. *ICOTS 8.* Ljubljana.

Stern, R., Coe, R., & Stern, D. (2017). Still coming down from the mountains. *IASE Proceedings.* Rabat: IASE.

Stern, R., Coe, R., Stern, D., & McDermott, B. (2013). MSc Training in Research Methods Support. *Technology Innovations in Statistics Education, 7*(2).

Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*.

Wilkinson, L. (2005). *The Grammar of Graphics (2nd Edition).* Springer.