Blockers for Reproducible Analytical Pipelines (RAP) projects

Findings from the RAP champions meetup at the Office for National Statistics, London, 16 October 2018

Summary

At the inaugural meeting of RAP champions, Government Digital Service colleagues (GDS) facilitated a discussion with around 30 cross-government Reproducible Analytical Pipelines (RAP) champions to explore what prevents them from doing RAP projects. This short note sets out the main issues that were identified by the group. Participants were asked to discuss their blockers and record each under the headings of 'organisation and culture', 'data' and 'technology'. The findings are summarised below.

Interestingly, most blockers related to organisational or cultural issues, but the group also raised some issues relating to technology and data.

Themes

Organisation and culture

The champions reported that:

- There can be lack of senior support due to perceived risks of the RAP approach. Some specific concerns were raised:
    1. RAP risks simply creating a new set of "black boxes" and a niche group of "high priests" who can operate them, leading to undue reliance on a small number of experts to build and sustain the RAP deployment. This risk was echoed by champions, who told us that RAP projects are not always sustainable because expertise is often held by a few individuals. Newcomers or other analysts already in the department may lack the technical expertise to pick things up as experienced staff move posts.

2. Some people perceive that the process of automating statistical production will remove analyst involvement entirely, with the corresponding risk that it might become *less* statistically rigorous and *more* error-prone. Champions also reported that some analysts felt that their roles or jobs might be threatened by the automation of workflows.
3. There is a lack of time for instigating and building RAP workflows because of conflicts with business-as-usual work.
4. There is a misconception that RAP requires that teams use R or Python for implementation. These tools have advantages (and a lot of infrastructure already exists here) but the approach is platform-agnostic.

Data

The champions reported that:

- data storage and access procedures are suboptimal, including examples where data is stored in disparate spreadsheets or convoluted databases
- data isn't always machine readable and in a form that's ready for use, which can slow the process
- There's a lack of clarity around data sensitivity and handling

Technology

The champions reported that:

- There is a lack of access to – and support for – the tools and environments needed to deliver RAP projects. The picture varies widely across departments.
- Champions were unsure of deployment methods that are both secure and long-term. We will probably want to work towards "containerised" BAU platforms in the longer term as RAP projects mature. For now, this is a medium-term problem as most departments are just starting out with RAP deployments.
- Control of each step of a RAP pipeline may be distributed across multiple teams or individuals, slowing the process down.

Duncan Garmonsway (GDS)
Matthew Dray (GDS) and
Martin Ralphs (GSS Good Practice Team)
3rd December 2018