UK COMPUTING RESEARCH COMMITTEE

RCUK Capital Investment

UKCRC Response to the RCUK Capital Investment Consultation

The UK Computing Research Committee (UKCRC), an Expert Panel of the British Computer Society, the Institution of Engineering and Technology and the Council of Professors and Heads of Computing, was formed in November 2000 as a policy committee for computing research in the UK. Its members are leading computing researchers from UK academia and industry. Our evidence reflects the experience of researchers who each have an established international reputation in computing.

Q1: What do you think are the significant capital investments needed in the next 10 years to maintain future sustainable national capability in your area of research?

The most important investment in computing research is the human capital of UK researchers and we recommend this as the first priority for investment. Where capital investment is needed, however, an important priority is data. Many researchers across all RCUK funding areas now rely on shared data and computational tools that access these data. Huge investment has, of course, been made by many organisations at many levels in data storage, curation, and computation facilities but the rapid and sustained increase in scale of use of data and software tools now merits a concerted national effort in this area. This issue spans all the research councils.

The problems of dealing with the scientific data deluge are widely debated in the international science community. A recent special issue of Science (February 2011, www.sciencemag.org/site/special/data/ overviews the threats and opportunities. As part of their review, Science ran a survey of their peer reviewers (of whom 1700 responded). In summary www.sciencemag.org/content/331/6018/692.short they report:

> *"About 20% of the respondents regularly use or analyze data sets exceeding 100 gigabytes, and 7% use data sets exceeding 1 terabyte. About half of those polled store their data only in their laboratories - not an ideal long-term solution. Many bemoaned the lack of common metadata and archives as a main impediment to using and storing data, and most of the respondents have no funding to support archiving."*

The Science review also asked whether expertise was available in the lab to analyse their data in the way they want - only 27% said they had that expertise.

The current position in the UK is essentially laissez faire. Data and computational tools produced by RCUK funded research are mostly retained by the institutions that were funded to produce them. The research councils attempt to promote curation of these assets through a combination of encouragement (e.g. to make data from projects "open" after projects' end) and obligation (e.g. to make published papers available open access). Many institutions also, independently, have developed policies and initiatives for data preservation and software availability and some domains have developed repositories (although long-term funding is variable). Though well intentioned, this is ineffective. Most data and software produced by RCUK research is stored in locations that are difficult or impossible to identify other than by the researchers that created them (and eventually, perhaps not even by those researchers) and key data curation decisions often are taken at the end of a project when there is no time left to implement them properly. Many research institutions are addressing this problem but, at best, this will yield good data/software packages that are still hard to find. Google search is not

sufficient to gain the effect of a data/software management capability for the UK research community from a widely distributed collection of data sets and software packages.

A RCUK capability would address this problem by:

- Ensuring sufficient persistent data storage for the UK research base to cope with the physical demands of data repositories and access bandwidth for UK science. This need not be a dedicated facility and might not be supported entirely by RCUK funding (see Q2 below).
- Recommending standards on formats for data, software and metadata with a view to re-use of these across research projects.
- Maintaining a strategic approach to the retention of scientific data and software tools (including decommissioning).
- Providing a (virtual) platform and portals for experimental access to data and software tools, so that shared benchmark data/tools could more readily be shared, tested and built into future funded research.
- Providing a common policy that can be used for ethical review and hence ensure effective data management

Although this proposal comes from UKCRC, the computer science community would not be the only, or even a principal, beneficiary of a national capability in research data and software management. Although many computer science activities require benchmark data, the main generators of data are the traditional sciences (and, increasingly, the social sciences). Many software tools are also produced and maintained by computing-savvy researchers outside computer science. An improved data and software base is likely to stimulate significant progress in algorithm, processing and data mining that feeds into this national knowledge economy.

Q2: What are the key challenges, if any, in ensuring this capital investment?

The primary challenge is organisational because a RCUK capability for research data and software management must work in concert with the capabilities being developed locally at universities and other related research institutions. A UK capability should not attempt to solve all the problems of scientific data management so the investment needs to be targeted. One target is cost reduction. This can be achieved in several ways:

- Reducing the up-front cost to RCUK of new research. Currently the cost of creating data sets and software tools plus the investment necessary to ensure that the data/software generated via RCUK research is preserved for future use beyond the originating project is not obvious from grant costings. It must, however, be a significant cost on any grant that constructs, from scratch, data or tools that exist as products of earlier grants, and the evidence suggests that many such grants exist.
- Reducing the cost to universities, enabling them to support more research at grassroots. Every major university is wrestling with this problem and, although part of the solution may be local, there are likely to be economies of scale from a national capability. For example, consensus on standardised data formats is hard to achieve and maintain from a local base.
- Cultural change in the use of data and software tools. Although great progress has been made in computational thinking in the physical and social sciences, we are a long way from fully exploiting the potential of all the data currently held in UK laboratories. Meanwhile, the technology available to acquire data continues to increase in its effectiveness so the gap between potential and actual use will grow. There is an opportunity to gain research value from the existing resource, not only within scientific communities but across communities (surely a prerequisite to the aim of the research councils to work together across traditional disciplinary divides). Combining datasets, and investigating them for new insights other than those for which they are originally collected, requires a carefully crafted ethical policy and data protection/privacy management approach.

A second, related challenge is consensus. A capability in this area needs buy-in from grassroots researchers to be effective. This requires systems that are easy to understand and obviously beneficial - not easy but easier now than hitherto because the sheer scale of the problem is now itself a driver of consensus.

A third challenge is choice of system architecture. A range of possible solutions are possible, from (at the “lightweight” end of the range) a loose confederation of local data services to (at the “heavyweight” end) a bespoke data facility. Between these extremes are various mixtures of third-party “cloud” data services combined with local “walled gardens” and centralised services. Achieving the right balance of these requires discussion between academics, funders and industry. It also requires care in the arrangements for access, standardisation and maintenance so that the rights of individuals and organisations are well balanced against the obligations to the public good of RCUK funded research.