Provenance Tracking in CXXR

Chris A. Silles[1,*] **, Andrew R. Runnalls**[1,**]

1. University of Kent, Canterbury, Kent, CT2 7NF, UK

* C.A.Silles@kent.ac.uk ** A.R.Runnalls@kent.ac.uk

Keywords: Provenance, Lineage, Auditing, S AUDIT, CXXR

Provenance is a record of lineage of a data object, and describes such things as what source data the data object was derived from, and the sequence of commands which was applied to generate the data object. Information systems are now ubiquitous in application domains where identifying the provenance of data is a critical ability, such as ensuring reproducibility of scientific research. This has led to the establishment of the field of *Provenance-Aware Computing*.

One of the pioneering papers within the provenance-aware computing literature is *Auditing of Data Analyses*, published by Becker and Chambers in 1988. In this article, the authors describe an *S AUDIT* facility which featured in *New S*. When *New S* was released in 1988, it signified a milestone in provenance-awareness. When a user issued a command within a *New S* session it maintained a record of the command, as well as which objects were read from or written to during the course of its execution. *R* currently has no support for an auditing facility such as *S AUDIT*.

Recently, the development of the methods employed within the field of provenance-aware computing for collecting and querying provenance has reflected the growing demand for more detailed information about the origins of data. Questions being asked of provenance information are typically complex, and it would be impossible to answer them using only a facility such as *S AUDIT*. Further advances have been made in the area of interoperability. The *Open Provenance Model* describes a method for the representation of provenance information so that, among other things, it may be exchanged between systems.

This paper describes how we have so far introduced facilities for provenance tracking into *CXXR*. *CXXR* is a project to refactorise the R interpreter into C++ while retaining as far as possible full functionality. The goal of *CXXR* is to allow for easier creation of experimental variants of the R interpreter.

In this paper we will describe and demonstrate the features we have introduced, such as the following facilities for inspecting the provenance of a given object,

- The sequence of operations performed on it (i.e. how it came to be);
- Which objects were used in its creation (i.e. its *ancestors*);
- Which objects used it for their creation (i.e. its *descendents*).

We will also discuss issues surrounding provenance collection in the context of a statistical environment, such as

- At what granularity provenance should be collected, and users be able to query it;
- How interoperability with other provenance-aware systems can be achieved.

Finally the paper will reflect on issues we have encountered while conducting this research, and outline its future directions.

References

Chambers, J.M, Becker, R.A (1988). "Auditing of Data Analyses", *SIAM J. Sci. Stat. Comput.* 9, 4, 747–760

Moreau, L., et al. (2008). "The Open Provenance Model", *Technical Report, University of Southampton*, `http://twiki.ipaw.info/bin/view/Challenge/OPM`

Moreau, L., Groth, P., Miles, S., Vazquez, J., Ibbotson, J., Jiang, S., Munroe, S., Rana, O., Scheiber, A., Tan, V., Varga, L. (2008). "The Provenance of Electronic Data", *Communications of the ACM* 51, 4, 52–58

Runnalls, A.R. (2009). CXXR: Refactorising R into C++, `http://www.cs.kent.ac.uk/projects/cxxr`