

Summary

1. Software is often an important output of a research project. Sometimes a piece of software is the primary output of the research, sometimes it is a tool developed for a specific purpose, sometimes it is a proof of concept. Often research data can be of little use without the specific software developed to process it.
2. It is widely accepted that the preservation of software in working form is particularly difficult to ensure over the long term. Software depends critically on many aspects of its environment, such as the hardware and software platform, compilers, libraries, and other associated software components. Thus digital decay can affect software far more than other digital artefacts because changes in any one of these related technologies can render a piece of software inoperable. Furthermore, even when the software in question is actively maintained, its interfaces and functionality may evolve so that it becomes impossible to validate results by rerunning a particular analysis *as it was* in some previous version of the software.
3. This project will investigate issues around the deposition and preservation of software artefacts in repositories. It will develop guidelines for the preservation of software research outputs, in particular, considering the use of existing software repositories such as sourceForge. It will develop these guidelines from the partners' experience in running a number of software repositories but in particular by monitoring and analysing in detail a particular case study which is a thematic software repository in the software engineering domain.

Scope

4. The project will focus particularly on:

The classification of software artefacts to enable their retrieval and use in new contexts. The classification of software is multi-dimensional. It can be classified by function, by language, by platform, by maturity, by interface, by license conditions etc. To enable retrieval and use of software in stand alone form requires at least searching by function and platform, to enable its incorporation into other software requires further criteria such as searching by language, compiler version, etc.

The preservation of software artefacts as platforms and associated components evolve. Software depends critically on its environment. The hardware platform, operating system, compiler, libraries, are all likely to evolve between deposition and retrieval. Each of these potentially make it highly non trivial to be able to rebuild and rerun software as was when it was stored. Different approaches to this exist, from the preservation of all associated components as they were at the time of deposition, to the proactive adaptive maintenance of software with different approach being appropriate in different situations.

The utility of standards as a means to protect software from changes in its environment. Several relevant metadata standards exist notably, the ISO standard OAIS which provides a reference model for Open Archival Information Systems and the Library of Congress maintained PREMIS standard (PREservation Metadata: Implementation Strategies) which defines core set of preservation metadata elements

applicable to a wide range of digital preservation activities. Standards for licensing and Rights (e.g. Creative Commons) are also relevant. We will investigate these and other standards and assess the costs and benefits of their use for software repositories.

Process issues for preservation of software as a research output. A major hurdle to widespread deposition of software is clearly the significant effort involved in doing so effectively. Furthermore, the value added through wider accessibility may well accrue to others than the original depositors. One way to ameliorate this effort and to make managed deposition more attractive to developers is to build the process into the development lifecycle and so make deposition yield benefits immediately during the development process rather than an add-on burden at the end of the project. The goal here is to make preservation benefits a low cost addition to the management of software required during its development.

5. More generic issues related to preservation of digital objects in general will not be considered here as they are well covered by other projects in this and other programmes.

The Case study

6. The partners will draw their on experience of running a number of software suites on behalf of the academic community. An example of a set of ongoing software suites run by the CCLRC is the Collaborative Computational Projects¹. An example of a large software suite developed over 20 years which is no longer under development but still in use is the Starlink suite². The detailed work of this project however will be undertaken in relation to a *new* software repository from the software engineering domain. The case study chosen for work will be the Verified Software Repository³ which is being run jointly by the University of York and the CCLRC. This case study is chosen because of its long term nature as part of one of the Grand Challenge in Dependable Systems Evolution⁴ and because of the breadth of the types of its content. In being new, furthermore, it is still of a scale where it is feasible to experiment with a variety of tools and processes to an extent which would not be possible in a larger, more established project. The particular software engineering community which the repository supports are also highly knowledgeable in the management and reuse of software and so would be well placed to assess the impact of the tools and techniques investigated.
7. In common with many other software development projects, the Verified Software Repository is currently hosted on the sourceForge repository of open source software. The use of sourceForge is currently widespread and it hosts many examples of community based open source software development projects of major importance as well as a very large number of smaller projects because of its convenience and free availability. Because of its genericity, however, sourceForge provides only some general

¹ <http://www.cse.clrc.ac.uk/ccp/index.shtml>

² <http://dev.starlink.ac.uk/statcvs>

³ <http://epubs.cclrc.ac.uk/work-details?w=33971>

⁴ http://www.ukcrc.org.uk/grand_challenges/index.cfm

support for the variety of software associated artefacts and processes and this is in a highly non-prescriptive “laissez-faire” form. It does not provide guidance or support specifically for long term preservation nor does it offer any guarantee of longevity. This project will in particular investigate and report on the ways to use sourceForge in conjunction with archival techniques to maximise the long-term benefit of preservation of the content.

Methodology

Case Study based analysis

8. The project will be based primarily on experiences from a particular case but will also draw on the previous experience of the partners in running other software provision services. The case study is a recently initiated thematic software repository being developed by an international collaboration and supported by several nationally funded projects⁵. The target repository has a long term goal and will therefore be addressing issues surrounding long term preservation and access to software research outputs in a particular field of software engineering.

Reporting Observations

9. From the setting up and running of this repository the proposers we will gain significant experience in the practicalities of setting up and running a thematic software repository and this project will record and report on that experience. In particular, we will assess technological options and measure and report on experiences with those chosen. Where possible within limits imposed by limited resources, we may undertake comparative pilot studies where more than one option is developed in parallel.

Data Analysis

10. Throughout the project, but particularly in the later stages, we will analyse the data collected on resource levels and report on this

Recommendations and Guideline

11. The primary output of the project will be reports on our experiences. We will also make available the data collected and the methodology employed to analyse it. From the analysis of this data, we will develop and disseminate recommendations and guidelines for the provision of repositories of this nature.

Schedule

12. The project will run in 4 cycles of 6 months. In each cycle will collect data, analyse and report on our experiences in conducting the case study and make these reports available to the funders. In agreement with the funders we will develop an appropriate means to disseminate these results in a wider forum. The end of each cycle will also provide a milestone for interchange of outputs between the four workpackages.

⁵ See the background resources described at the end of this proposal.

Dissemination

13. At the end of each project cycle, that is at project months 6, 12, 18 and 24, we will report to the funders on the progress of this work and agree with the funders the dissemination plan for the next cycle. This dissemination is expected to take the form of papers and presentations at appropriate events and, additionally, the production of a detailed technical report on the work of each of the 4 workpackages:

The classification of software artefacts to enable their retrieval and use in new contexts.

The preservation of software artefacts as platforms and associated components evolve.

The utility of standards as a means to protect software from changes in its environment.

Process issues for preservation of software as a research output.

The conclusions of these reports, that is the validation of the design decisions taken, will come with hindsight from the experiences of the work. For this reason, we will undertake the major dissemination activities during the fourth cycle. In particular, although drafts will be developed throughout the project, we would expect to publish and publicise the reports towards the end of the project. We will agree with the funders on the most effective way of disseminating these reports.

Programme of work*WP1. Classification of Software artefacts*

14. Effective retrieval of software artefacts requires classification in several dimensions depending on the purpose behind the retrieval request.
- Retrieval for reuse in new context
 - Retrieval for revalidation of previous run (for example to revalidate previous experimental analysis)
 - Retrieval for adaptation in the same context
 - Retrieval for regression testing/debugging.
15. Each of these will require classification by a variety of variables. For example retrieval for stand alone use is likely to require at least classification by function and platform; retrieval for adaptation and incorporation into other software may require for example classification by function, language, platform, licensing conditions. Classification could be undertaken under for example the following schemes:
- Classification by function,
 - Classification by language,
 - Classification by platform,
 - Classification by maturity,
 - Classification by interface,

- Classification by license conditions
16. This workpackage will consider the artefacts in the case study repository and use this analysis to develop and refine a suitable set of classifications.

WP2. Preservation of software

17. The evolution of software as its associated environment changes is recognised as a major challenge. Software reuse, adaptive maintenance, version control and configuration management are issues in their own right which have been extensively researched within the software engineering community.
18. This workpackage will identify the major issues associated with the preservation of software and, taking into account the existing body of work in the above areas as well as work in the preservation community (e.g. the inSPECT project), will make recommendations for preservation policy and processes for software, with reference to the VSR as an indicative case study.

WP3. Standards

19. Standards and their supporting software enable interoperability and form a key component in the effective operation of the case study repository. With respect to software, this can be regarded in OAIS in two aspects. Firstly, the role of software as part of the representation information for the preservation of other digital objects (e.g. science data). Secondly, the use of the OAIS conceptual architecture for the preservation of software in its own right.
20. The case study project will be undertaking an active role in monitoring and development of relevant standards, both directly through the participation in working groups on such relevant standards bodies as ISO, BSi, W3C, and OMG, and through active collaboration with other UK initiatives such as the Digital Curation Centre and the Digital Preservation Coalition. This project will review and report on that experience and make it available to the wider community.

WP4. Process

21. The case study will be undertaking a number of activities and managing a suite of resources on behalf of the software engineering community. Co-ordinated management of these activities and resources will add value to the outputs of research in the area, and so improve the efficiency of the overall research programme by facilitating experimentation, encouraging standardisation and employing economies of scale.
22. This Workpackage will review and report on experience gained from this work in the case study project to provide a series of practical guidelines on how to build the use a software repository into the lifecycle of projects delivering software as part of their research outputs.

Dissemination

23. The case study project will be using a variety of mechanisms for publicity and dissemination including activities such as summer schools, workshops, competitions, newsletter, websites, wiki, and an international journal to record the contributions made. This project will assess and report on the effectiveness of these methods as well as disseminating its own work in appropriate fora. (See also Paragraph 13.)

Community Impact

24. Current initiatives to encourage better management of research outputs are focused largely on publications and data outputs. It seems that the preservation of software, which is another major research output, is being largely ignored in this context. Often software research outputs provide an essential link between the published results and data behind them. Without the preservation of that software, or software with equivalent or superior functionality, the link between the data and research results can be lost and with it of much of the value to be gained from their preservation.
25. It is widely acknowledged that the preservation of software is an extremely difficult task and perhaps it is for this reason that the problem is sometimes being “swept under the carpet”. The high level rationale behind this project is that by demonstrating the possibility of preservation of software through a particular case study, we will encourage the community to face up to the challenge and promote the preservation of software as a feasible, worthwhile activity.
26. The project will liaise with other related projects. For example, the CCLRC are participants in the DCC and the CCP SourceForge-like Facility project and the team have established links with both these projects. We also work closely with the OMII in other areas, particularly through the NGS project. Links would also be established with OSS Watch.

Benefit to participating organisations

27. The case study activities are funded by other means and so the experience gained from this will be available to the participating organisations in any case. The JISC funding will allow the experience and knowledge gained in that work to be made available to a wider community and to influence other engaged or about to engage in similar endeavours in order to make their own work more effective and efficient. Thus the benefit to the participating organisation will be in terms of the reputation gained from being seen to be leading in this area.

Resources

28. The resources employed on this project will leverage a substantial effort in the international collaboration in order to disseminate the experiences in software preservation gained by this initiative to a wider audience.

29. The principle investigator at York, Professor Jim Woodcock, will devote 10% of his time to the project providing the overall strategic direction and bringing to the project very significant expertise in software engineering. He will make major technical contributions to the project related to specification of software functionality, reuse, modularisation, and software engineering lifecycles.
30. The resources required for this project at CCLRC will be 0.25 FTE of a senior software engineer to undertake the day-to-day management of the work, plan the activities in detail and liaise closely with the case study teams and other stakeholders as well as leading the definition of the methodological approach. A further 0.25 FTE of Research Associate effort will be required to undertake evaluations, collect, monitor analyse and report on the observations undertaken. Local management of these staff, under the overall responsibility of Dr. Juan Bicarregui, will be provided by the host institution at no cost to the project.
31. In addition to this dedicated effort, we would have access to the case study team in terms of experience in running and using the particular repository and other expertise at the participating institutions from other related activities (see appendix A).
32. The budget includes an element for travel between sites and for dissemination at appropriate events. Due to funding from other sources for the case study, we will be able to run meetings for this project alongside others at relatively low cost. Similarly we will not require any other support costs for the staff involved.

Timescale

33. Much of the value of a repository comes from its breadth and scale therefore longevity is key to the success of the programme. For this reason we propose to run the project for as long as possible within the scope of this call, and would hope to be able to continue the work beyond that timeframe.

Funding

34. The JISC are asked to provide 80% of the full economic cost for the dedicated staff time described above and the partners will contribute the remaining 20%. In addition to this, the international collaboration undertaking the case study will provide significant gearing to the project resources. The cost of additional meetings to pursue the objectives of this project are included in the request of funding from the JISC.

Budget

Directly Incurred Staff	April 07– March 08	April 08– March 09	TOTAL £
Total Directly Incurred Staff (A)	£	£	£
Non-Staff	April 07– March 08	April 08– March 09	TOTAL £
Travel and expenses	£1,549	£1504	£3053
Hardware/software	£	£	£
Dissemination	£	£	£
Evaluation	£	£	£
Other	£	£	£
Total Directly Incurred Non-Staff (B)	£1549	£1504	£3053
Directly Incurred Total (A+B=C) (C)	£1549	£1504	£3053
Directly Allocated	April 07– March 08	April 08– March 09	TOTAL £
Other	£	£	£
Directly Allocated Total (D)	£37592	£36525	£74117
Indirect Costs (York)	£4407	£4281	£8688
Indirect Costs (CCLRC)	£19714	£19155	£38869
Indirect Costs (E)	£24121	£23436	£47557
Total Project Cost (C+D+E)	£63262	£61466	£124728
Amount Requested from JISC	£50610	£49172	£99782
Institutional Contributions	£12652	£12294	£24946
Percentage Contributions over the life of the project	JISC 80 %	Partners 20 %	Total 100%

Risks

35. The following risks have been identified and considered:
- Risk:* Case study not typical
Consequence: Invalid generalisation from single example
Mitigation: We will have project conclusions reviewed by staff experienced from other software repository projects.
- Risk:* Case study not pursued for the length of this project.
Consequence: Premature close to the project or move to another case study.
Mitigation: We consider this highly unlikely as the case study is a major international collaboration.
- Risk:* Scope of project too large for resourcing available
Consequence: Some objectives not pursued or not adequately undertaken.
Mitigation: If this was found to be the case, we would work with the funders to prioritise the activities.
- Risk:* Reorganisation of the participating organisations.
Consequence: Delays or premature cessation of the project.
Mitigation: Current plans to merge CCLRC with PPARC are unlikely to impact on this project.
- Risk:* Availability of appropriately skilled staff.
Consequence: delay in delivery or reduction in quality of outcome.
Mitigation: Suitably skilled staff are available within the current teams and alternatives are also available.
36. Process for ongoing risk management. We will reconsider and review the risk register for the project at each of the project milestones.

Intellectual property

37. A fundamental issue in any repository is the ownership of the resources it contains. The situation with respect to access of research outputs is currently under debate in the research councils and elsewhere. The Open Access movement argues that public funding prescribes public access and for published output the of the research councils have recently reiterated their support for this principle. However, some IPR clearly also rests with the investigator and/or their Institution. Overall co-ordination of policy towards ownership of public sector research has been the subject of increased attention since the Baker Report.
38. For software the issues are particularly complex, for example, a large variety of open source license formulations exist. For the case study the issue is unlikely to be particularly high priority as the content of repository will be open for public browsing. Whilst recognising that these issues are critical to the usefulness of software repositories, due to necessarily limited resources, we do not intend to focus on issues around intellectual property and digital rights management which is being addressed by other projects.

Sustainability

39. The case study is underway and will be sustained from other sources of funding. The proposed project will leverage this effort to deliver the project outputs a time limited basis. It will consider as far as possible within its lifespan, all issues related preservation. However, clearly many issues related to long term preservation of software will not emerge in this timescale and may well be the subject of follow-up projects at a later date.

The team

40. *The University of York.* York's Computer Science Department is internationally leading in many areas of software engineering. In 2003, it was assessed by the University Funding Council at grade 6*, the highest possible result. British Aerospace has designated the department as a Centre of Excellence, and Rolls-Royce plc has named it a University Technology Centre. Major research groups in diverse areas of Computer Science have attracted considerable funding from industry and the research councils to support their work. The department has a distinguished record upon which to base the evaluations of techniques employed by the case study repository.
41. *Jim Woodcock* holds the Anniversary Chair in Software Engineering at the University of York. Previously, he was Professor of Software Engineering at the University of Oxford, where he founded the Centre of Excellence in Software Engineering and directed its academic Programme. In 1992, his research team won the *Queen's Award for Technological Achievement* for its work with IBM. He was the academic consultant for the first product certified to *ITSEC Level E6* in 1998, and served for over ten years as an advisor on secure systems to the British and US governments. In 2002, he won the *Rudolf Christian Carl Diesel Prize* from the Society for Design and Process Science. He is a co-director of the *United Nations Summer Schools on Theoretical Computer Science*. He serves on the editorial boards for four international journals; he has served on over sixty international conference programme committees, and has chaired fifteen of them. He has given invited papers and keynote speeches at over thirty conferences, and is the author of nearly 200 scientific papers and books. He is currently a Visiting Professor at Trinity College Dublin, Shanghai Jiao Tong University, and the Federal University of Pernambuco, and a Visiting Fellow at the University of Oxford and the United Nations University. He is a past-chairman of the *OCR A-level Computer Science Panel* (for which he was elected a Fellow of the RSA). He is a Chartered Fellow of the BCS.
42. Professor Woodcock is the moderator for UKCRC's *Grand Challenge 6* and principle investigator of the EPSRC project VSR-net which is supporting the case study of this project. He is actively involved in the international activities around this initiative.
43. *CCLRC.* The CCLRC provides facilities for the UK and international scientific community including numerous software and data repositories supporting the research councils' programmes. CCLRC has the infrastructure and experience to support the scale and longevity of the proposed facility and organisational structure and culture to

undertake a role supporting the scientific community. Some examples of software and data repositories hosted at CCLRC are described below. CCLRC is a major partner in the UK Digital Curation Centre and led the RCUK group on research output information (ROIG) which recently developed the RCUK policy on Open Access to publications and the management of research outputs.

44. **Juan Bicarregui** is the head of the e-Science Applications Support Division in the e-Science Department at CCLRC. The sector comprises six groups totalling approximately 60 people and includes groups which support some significant services for JISC, specifically the Digital Curation Centre (DCC), The National Grid Service (NGS) and The National Academic Mailing list service (JISCmail). Juan Bicarregui's research background lies in the advancement and technology transfer of formal software engineering techniques. Juan Bicarregui is the author of over 50 scientific papers and books has served on numerous international conference programme committees.
45. Dr Bicarregui is actively involved in the activities around the case study and is on the board of the VSR-net project.

Scientific Repositories at CCLRC

46. The CCLRC exists to promote and support scientific and engineering research by developing and providing facilities in support of research programmes. With this purpose, it serves the whole academic community and the provision of scientific repositories is central to its mission. It has extensive expertise and the necessary infrastructure to manage national hardware and software computing resources on behalf of the UK research community. Some current software, data and computational facilities provided by CCLRC are listed in appendix A.

Background resources related to the case study

47. The case study project, The Verified Software Repository is described in *The Verified Software Repository: A Step Towards the Verifying Compiler*, Formal Aspects of Computing (2006) [DOI: 10.1007/s00165-005-0079-4] accessible at <http://epubs.cclrc.ac.uk/work-details?w=33971> and further information about the VSR-net project can be accessed at <http://www.fmnet.info/vsr-net/>. This activity is part of the GC6 of the UK Grand Challenges described at http://www.ukcrc.org.uk/grand_challenges/index.cfm. In the US, the VSR initiative is described at <http://qpq.csl.sri.com/vsr> and internationally at <http://www.dagstuhl.de/de/program/calendar/semhp/?semnr=06281> and <http://vstte.ethz.ch/>. The nascent VSR repository is at vsr.sourceforge.net and there is a wiki at <http://www.gc6.cclrc.ac.uk/gc6wiki/VerifiedSoftwareRepository>.