End-of-Life Scenarios for Repositories of Virtual Organisations

Leslie Carr, University of Southampton. lac@ecs.soton.ac.uk
Chris Gutteridge, University of Southampton. cjg@ecs.soton.ac.uk

Abstract: A repository may be a useful information gathering solution for a virtual organisation, enabling knowledge sharing across the boundaries of its component organisations and fulfilling the virtual organisation's external dissemination objectives. However, virtual organisations have a pre-determined lifespan, and this paper outlines an option for the repository once its hosting organisation is disbanded.

Keywords: repository obsolescence, sustainability.

Introduction

Virtual organisations are fluid and cross boundaries of real organisations. They come together over a shared set of goals or a joint activity, often a funded project or a large-scale scientific experiment. Virtual organisations have been a particular focus of the e-science or cyber-infrastructure domain, as they feature the high profile extended, large-scale data gathering activities seen in the hard sciences inevitably involve collaborations between many dozens of partners over many countries. At a more modest scale, collaborations between partners in EU Framework projects, or between interdisciplinary activities on a national scale have many of the qualities of virtual organisations.

A repository may be a useful component of the infrastructure of a virtual organisation. As an organisation in its own right, with its own intellectual outputs and its own defining literature, a repository can provide (and make explicit) the organisational domain, memory and experience. Although relevant items could be deposited into any or all of the partners' institutional repositories, it is desirable for the virtual organisation to have its own identity and knowledge store under its own control, either for political or pragmatic reasons. Each partner organisation may have a deposit policy that admits only items generated by its own members into its own repository (leading to fragmentation of the VO's knowledge) or each participant may be reluctant to put their items into someone else's repository.

Although repositories can be as helpful to virtual organisations as they are to real ones, there is one core attribute of a virtual organisation that makes it a poor host for a repository – a virtual organisation is fundamentally limited in time. Any funded activity (a project, an experiment) has a fixed duration that is known at the start of the project. Although that time – six years for a EU project or 15 years for the Large Hadron Collider project – may be significant in terms of a researcher's career it is still tiny when compared to the lifetime of a real institution[1]. Relatively few repositories are started with the exclusive aim of long-term preservation, but a fly-by-night attitude to information collection is antithetical to the principle of information persistence that a repository embodies.

[1] The oldest universities in Europe (Bologna, Paris, Oxford) are around 900 years old, many times older than even the nations in which they are now situated.

The repository community has already addressed the problem of disappearing hosts. The adoption of persistent URLs, either at a technical or a policy level is well documented. If the host organisation for a repository disappears, then the URLs can be diverted to an alternative host and the permanence of the information, and the long-term efficacy of its identifiers guaranteed.

However, the mechanism by which this information transfer is effected has not yet been worked out. It may be the case that an alternative organisation will simply host the entire operating environment under a different IP address. Alternatively, the information content of the repository may be imported into a separate area of another repository. This solution may require some level of metadata and data translation, depending on the compatibility of the two systems. These technical solutions ignore the policy differences that created the initial repository as a separate entity in the first place. It may not be appropriate to 'sublet' an area in an alternative repository, or an acceptable repository may not be available.

An alternative approach is to reduce the impact of a repository so that it can be hosted as a static web site at very low cost. A repository is a complex piece of software that models a persistent object store with all kinds of local services related to ingest and reuse of the deposited material for various purposes that are relevant to the community. If the community itself is disbanded, then there is no longer any requirement to support ingest or any of the extended services for the local community. Preservation and other curation services can be managed on the versions of the objects that are exported from the repository in a suitable DIP (or exposed AIP), but the need for a durable public access service to back up "the public record" remains.

A Repository or a Web Site?

The OAIS repository model emphasises the role of ingest, dissemination and (internal) archival processes. Figure 1 shows this from the perspective of the repository users – the external and user-facing processes are emphasised and the repository internal processes hidden within the 'well-maintained object store'. When a repository is mothballed and its community of registered users disappear, the ingest processes will cease to be used (fig 2a). This leaves only the archival processes of the object store and the dissemination processes that need to be maintained.

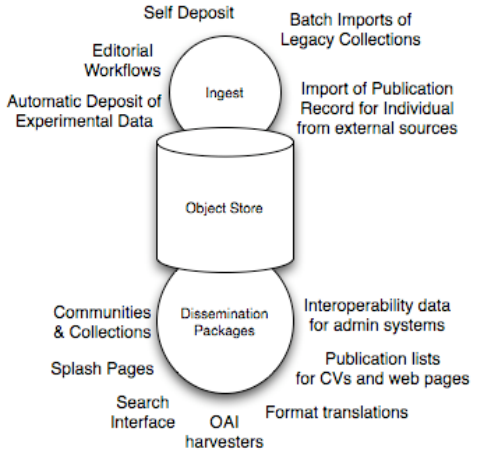

Figure 1: Repository from the perspective of a User Community

If the community is very small or very focused then the need for dissemination may also disappear (fig 2b) – the repository may have been a mainly internal device for the organisation and there may be no external parties interested in accessing the objects that embody the organisation's knowledge store. In that case, the repository could be simply closed down and removed from the information space. However, from a preservation perspective, it might well be the case that the repository or its contents should be preserved *even if no-one currently wants to use or read any of them.*

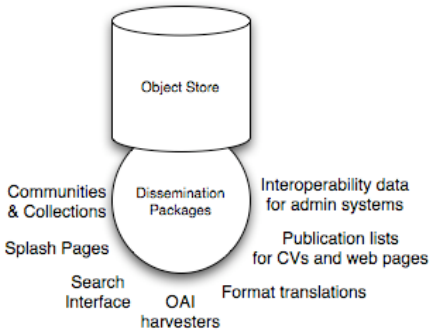

Fig 2a: Repository Without a Depositing Community

Figure 2b: Repository Without Any Users

The preservation policies of various stakeholders may cause some aspect of the repository to be retained – some or all of its data items may be copied to another repository, or the repository as a whole may be kept in a dormant state. Funding agencies may have such an interest in order to account for the investment that was made in the project. (In the UK, the ESRC requires deposit of all 'non-academic' outputs from a project, which could well include web sites and repositories.)

There are a number of ultimate scenarios for an 'abandoned' repository:

1. The repository is maintained unchanged by the same host organisation that maintained it during its active life
2. The responsibility for the repository is transferred to a different organisation, but it is maintained as-is
3. The contents of the repository are migrated to a different (actively maintained) repository through the normal interoperability channels
4. The contents of the repository are migrated to a number of different (actively maintained) repositories through the normal interoperability channels
5. The repository is itself deposited into a preservation environment as a complex digital object
6. The contents of the repository are deposited into various preservation environments

Although they appear at first sight to be most desirable, options #1 and #2 are likely to be 'delaying tactics'. If the underlying purpose and rationale for the repository has finished, it is unlikely that the repository itself can be maintained and migrated through software versions and hardware upgrades on an indefinite basis without active champions and ongoing business cases. Repositories are complex software environments with non-trivial overheads in terms of daily management and maintenance. This leads to the above alternatives for continued support. However, the change in usage patterns may allow the software itself to be sufficiently simplified that new options become viable. Maintaining a simple web site is a very low cost

activity that may be taken on for an indefinite basis with little additional impact on a well managed ICT infrastructure for a medium- to large-sized organisation. So the more that a 'repository' can be made to operate like a 'web site' then the easier it is to find a long-term home for it beyond its immediate life.

In the EPrints software platform for example, the dissemination facilities are provided as a mixture of dynamic services and static web pages. Indeed, the original design objective for EPrints was to avoid as much unpredictable (user-generated) system load as possible by maximising the number of static pages served by the repository.

Static	Potentially Static	Dynamic
• Communities and Collections • Splash Pages • Publication lists for CVs and web portals	• Format translations • OAI harvests • Interoperability data for admin systems	• Search

Figure 3: Static and dynamic disseminators in EPrints

Figure 3 shows those dissemination packages listed in Figure 2 that are pre-generated by EPrints as static web pages and those that are provided by the invocation of dynamic services in the normal course of events, but could be pre-generated. The remaining are dynamic by nature.

AKTprints Case Study

AKTprints (eprints.aktors.org) is a repository that was used by the EPSRC AKT Interdisciplinary Research Collaboration (www.aktors.org) of five universities that lasted for six years (2001 – 2006). The repository collected papers written by the partners that were to be considered as outputs of the project, and fulfilled dissemination and recordkeeping purposes, allowing the project director to keep track of the project outputs and to report them appropriately. Although the project officially ended in 2006, the long-term future of the repository is only being discussed now (end 2007).

The repository is to be maintained as described above, by converting it into a web site, in order to retain a historic view of the project. As a first step, the repository's Web template has been rewritten to remove all references to dynamic services (see figure 3). The search page has been replaced with a link to a Google site search. This apparently leaves the repository as a static web site as far as external users are concerned, but there is a relatively complex apache internal configuration that hides the reality of a complex internal file system directory structure. To simplify this a further stage, a web mirror is made of the (now) static repository using the `wget` command line tool. This mirrored content is suitable for replacing the live repository as it maintains all the persistently identified web pages under their original URLs. It can also be used to provide a DVD-ROM distribution of the repository contents, if necessary.

A complicating factor for this particular repository is that one third of the papers are not available for public view; the project partners are being asked to review the

visibility of their deposited materials in order to determine their long-term availability.

At the time of writing this article, this process is part-way through completion, so further complications may yet have to be dealt with. No provision has been made for specific preservation activities; all of the full-text holdings are in PDF format and are judged appropriate for the purpose of this repository. The responsibility for complex management is left to the partners' institutions and to any copies of these works that may be copied in their home repositories. Although not mentioned above, complete backups of the repository are made from the database tables (as SQL dumps), from the file system data contents, the repository configuration files and from the interoperability formats supported by the repository. This provides multiple pathways for reviving the live repository in the future, if such a course of action is deemed appropriate.

Concluding Remarks

Not all repositories are associated with well-found organisations with a long-term future. Those repositories attached to virtual organisations need to have an exit strategy to be actioned when the hosting organisation winds down. One of the possible courses of action is to refactor the repository as a static website with minimal (close-to-zero) ongoing maintenance costs. Although such a strategy is quite a drastic curtailing of the repository functionality, it may potentially offer the best hope for long term support for the repository's holdings.