Daniel Finn
10/1/13
Digital Preservation
Kara Van Malssen

The SCAPE Project

Scalable Preservation Environments (SCAPE) is an ongoing research and development project funded by the European Union that aims to improve institutions' ability to manage and preserve large-scale, complex digital collections. The 42-month endeavor launched in February 2011 with a budget of 11.5 million Euros, and will end in July 2014.[1] According to the project's literature, the problem at hand is that, "the volume of digital content worldwide is increasing exponentially. This fact demands that preservation activities become more scalable and automated. However, current solutions for the preservation of digital object collections fail when applied to very large or complex sets of data."[2] SCAPE is attempting to combat this problem through increasing the role of automation in quality assurance and analysis of digital collections, as well as automating key preservation planning functions.

SCAPE is a very large undertaking comprised of partners from all over the European Union, such as Ex Libris, The British Library, the Internet Memory Foundation, Microsoft Research, Open Planets Foundation, Brno University of Technology, and nearly a dozen others.[3] The organizations being used to research and implement the resulting software tools are defined as "testbeds."[4] The testbeds are the engine of the project because they will serve as the exemplars for the organizations for which the SCAPE platform tools are designed, and hence they will define the problems, workflows and types of collections to which the tools will be most applicable. The project defines three types of testbeds, or organizations; large-scale digital repositories, research datasets, and web content. As such the project has the potential to provide solutions to a vast number of organizations interested in the preservation of digital information.

The key objectives the project laid out for itself are 1) Scalability, 2) Automation, 3) Planning, and 4) Integration.[5] The first objective refers to the ability of the SCAPE platform to handle differences in size in four areas; number of objects, size of objects, complexity of objects, and heterogeneity of collections. All of the tools resulting from the project have the aim of addressing these needs that vary from institution to institution, and even collection to collection within one organization. Automation refers to the goal for the tools to be able to handle complex workflows through computational clusters and decrease human intervention in the areas of

---

[1] Rainer Schmidt, "SCAPE Preservation Platform: Design and Deployment," presentation given at iPres 2012 Conference, 1-5 October 2012, accessed 9/30/13 at http://www.scape-project.eu/publication/an-architectural-overview-of-the-scape-preservation-platform.

[2] Open Planets Foundation, "The SCAPE Project: A Brief Introduction," accessed 9/28/13, ttp://www.openplanetsfoundation.org/blogs/2012-12-10-scape-project-%E2%80%93-brief-introduction..

[3] SCAPE Partners, accessed 9/30/13, at http://www.scape-project.eu/partners

[4] SCAPE Project, accessed 9/30/13 at http://www.scape-project.eu/about/project.

[5] SCAPE, accessed 9/30/13 at http://www.scape-project.eu/.

preservation planning, collection monitoring, and quality assurance. The planning objective refers to SCAPE expanding the capabilities of a service called PLATO.[6] PLATO is designed to help preservationists assess digital collections and select the best options for the preservation tasks at hand with reference to institutional policies. One major area of the research is continuing to develop machine-readable institutional preservation policies in order for the automated planning components to make the best suggestions possible. Finally, integration refers to the goal that all of the tools developed under the SCAPE project will be able to be offered under one open-source software platform.

The SCAPE project is designed to investigate these objectives in a number of research and development sub-projects.[7] The testbed group is included in the list of sub-projects, and most likely is intended to drive progress on the integration objective, as SCAPE project developers use their tools to create workflows to solve real world problems. Another sub-project is the SCAPE Preservation Platform, which investigates the combination of open-source software (SCAPE platform) and hardware components used to implement the various functions of the platform. One of the advantages here is that the model for implementing the platform is very flexible given the scalability of hardware options.[8] This is due to the fact that SCAPE's software architecture uses Apache hadoop, which allows for scaling computational functions from one to thousands of computers.[9]

The third sub-project is Preservation Components, which is devoted to developing the tools that will be compiled in the platform. Many of these are available already in beginning stages. Examples include Jpylyzer, a program for JPEG2000 files that can validate files as well as extract properties, and xCorrsound, a suite of audio analysis tools that can perform various tasks comparing sound files within a collection, mostly based around discovering redundant holdings.[10]

A sub-project called Planning and Watch is developing the tools that will assist in automating the process of triaging and developing preservation strategies for large digital collections. The planning component is called PLATO, mentioned above. The Watch component is called SCOUT and is designed to help gather knowledge about digital collections (e.g. monitor registries such as PRONOM, and incorporate the machine-readable policies) and automates monitoring according to this information. It would provide notifications based on this information, such as notifying regarding the obsolescence of a format, informing that a file does not fit established policies, or that new software is available to help analyze certain types of collections.[11] These components are extremely exciting in that they are designed to assist in

---

[6] PLATO Website, accessed 9/27/13, http://www.ifs.tuwien.ac.at/dp/plato/intro.html.
[7] Open Planets Foundation.
[8] Rainer Schmidt, "SCAPE Preservation Platform."
[9] Apache Hadoop website, accessed 9/29/13 at http://hadoop.apache.org/#What+Is+Apache+Hadoop%3F
[10] SCAPE Project Tools, accessed 9/30/13 at http://www.scape-project.eu/tools.
[11] Github, "About SCOUT" accessed on 9/29/13 at https://github.com/openplanets/scout

decision-making processes that are always sources of frustration and delay in large-scale collections.

The last subprojects are Coordination and Takeup. Coordination is designed to interact with the other sub-projects to ensure the integration objective is achieved, and the takeup sub-project is based on publishing research and results so that the project remains open.

With about 10 months left in the project, progress seems to be going according to expectations. An evaluation report was released in January 2013 stating as much, citing some successes and available deliverables as well as ceding expected shortcomings and areas needing improvement.[12] Areas that were seen to be successful included runtime stability, correctness, and organizational fit. Correctness was tested through two of the quality assurance tools; the xCorrsound audio analysis and identification. False positives for redundant findings had been reduced to 0.4%, and accurate identification reached 96%. The ultimate goals for these figures are 0.1% and 98%, respectively.[13] Organizational fit is an exciting success in that the software is achieving what the testbed organizations need it to. The most important area in which improvement and further testing was acknowledged is in the area of extending the computational processes for many of the components across multiple computers using the hadoop framework.[14] A number of goals were not deemed developed enough for evaluation, so hopefully these have seen improvement and will continue to see improvement through the rest of the project's timeline.

Another very helpful way for individuals and organizations to see the progress of the project is through the SCAPE Project Wiki.[15] The Wiki provides a glossary of identified problems and the workflows to solve these problems using various components of the SCAPE platform. Additionally it provides a forum for organizations to post their problems and for developers and users to discuss the results. It also serves as a general hub for information regarding the project, including general introduction and description, lists of project staff, upcoming events, and other information.

The SCAPE project is a massive undertaking that has already seen some successful results, and should continue to develop exciting applications. Due to the scalability and scope of the platform and its various components, there should be developments that everyone involved in the world of digital preservation will be able to take away to improve the effectiveness of their efforts in the face of increasingly large and complex digital collections.

---

[12] Bjarne Andersen, et al, "SCAPE: First Evaluation Report" January 2013, accessed 9/30/13, at http://www.scape-project.eu/wp-content/uploads/2013/08/SCAPE_D18.1_SB_V1.0.pdf.
[13] Ibid, 24.
[14] Ibid.
[15] SCAPE Project wiki, accessed 9/30/13, at http://wiki.opf-labs.org/display/SP/Home.

Webography

Apache Hadoop website, accessed 9/29/13 at http://hadoop.apache.org/#What+Is+Apache+Hadoop%3F

Bjarne Andersen, Catherine Jones, Kresimir Duretec, Peter May, and Yair Brama. "SCAPE: First Evaluation Report" January 2013, accessed 9/30/13, at http://www.scape-project.eu/wp-content/uploads/2013/08/SCAPE_D18.1_SB_V1.0.pdf.

Github, "About SCOUT," accessed 9/29/13 at https://github.com/openplanets/scout.

PLATO Website, accessed 9/27/13 at http://www.ifs.tuwien.ac.at/dp/plato/intro.html.

Rainer Schmidt, "SCAPE Preservation Platform: Design and Deployment," presentation given at iPres 2012 Conference, 1-5 October 2012, accessed 9/30/13 at http://www.scape-project.eu/publication/an-architectural-overview-of-the-scape-preservation-platform.

SCAPE Project Website, accessed 9/30/13 at http://www.scape-project.eu/.

SCAPE Project blogs, accessed 9/28/13, at http://www.openplanetsfoundation.org/blogs/2012-12-10-scape-project-%E2%80%93-brief-introduction.

SCAPE Project wiki, accessed 9/30/13, at http://wiki.opf-labs.org/display/SP/Home.