Research projects as a driving force for open source development and a fast route to market

RODA, SCAPE and E-ARK - a case study

Hélder Silva, Miguel Ferreira, Luís Faria
KEEP SOLUTIONS, LDA
R. Rosalvo de Almeida, n°5
Braga, Portugal
{hsilva, mferreira, lfaria}@keep.pt

ABSTRACT

Research projects, specially in the computer science domain, have consistently provided outputs as open source products or updates to long-standing open source projects. This occurs due to the shared openness nature of both research and open source, that enables re-use by the community spanning new developments in both research and open source products. But when an open source project serves a community and a real-world problem, the impetuousity of research can clash with the inertia of real-world application. Nevertheless, research projects can bring the much needed innovation to open source projects, and open source projects can bring the much needed route to market that research funders look for the outputs of the research they fund, ensuring the budget spent in research actually reaches the community and improves the world.

This paper presents an analysis of this dynamic with a case study about RODA, an open source repository for digital preservation, used in memory institutions such as archives, and two research projects, SCAPE, focused on digital preservation scalable services, and E-ARK, focused on standardization of information packages, integration with real-world applications, and database preservation.

The paper further tries to identify good practices for using existing open source projects in research and assure that research outputs are further carried into main versions of open source projects and find their way to the final user.

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Information Storage and Retrieval—*Digital Libraries*

Keywords

Preservation, Repository, Research, Open Source, Integration

1. INTRODUCTION

RODA[1] is an open source digital repository specially designed for archives, with long-term preservation and authenticity as its primary objectives. Created in 2006 on a 2 year project lead by the Portuguese National Archives in partnership with the University of Minho, it would later lead to the creation of the KEEP SOLUTIONS company[2] which up to now continues to develop RODA, foster its open source community, and provide commercial services for maintenance, support and on-demand feature development.

SCAPE[3] was a co-funded project by the European Commission under the Seventh Framework Programme. It ran from 2011 up to 2014 and aimed to develop scalable services for planning and execution of institutional preservation strategies on an open source platform that orchestrates semi-automated workflows for large-scale, heterogeneous collections of complex digital objects.

E-ARK[4] is an ongoing project co-funded by the European Commission under the Competitiveness and Innovation Framework Programme. It will run from 2014 to 2017 and it aims to develop a pan-European methodology for electronic document archiving, synthesising existing national and international best practices, that will keep records and databases authentic and usable over time.

This paper provides an overview of how research projects, namely SCAPE and E-ARK, were included in the RODA project, how their results were incorporated in the main features, how the roadmap is aligned with future developments, and how the output of research reaches the end users.

2. RODA

RODA is a complete digital repository system that provides functionality for all of the main units that compose the OAIS reference model. RODA fully implements an Ingest workflow that validates SIPs and migrates digital objects to preservation friendly formats, and provides Access by delivering different ways to search and navigate over available data as well as visualising and downloading stored digital material. Data Management functionalities allow archivists

[1] http://www.roda-community.org
[2] http://www.keep.pt
[3] http://www.scape-project.eu
[4] http://eark-project.com

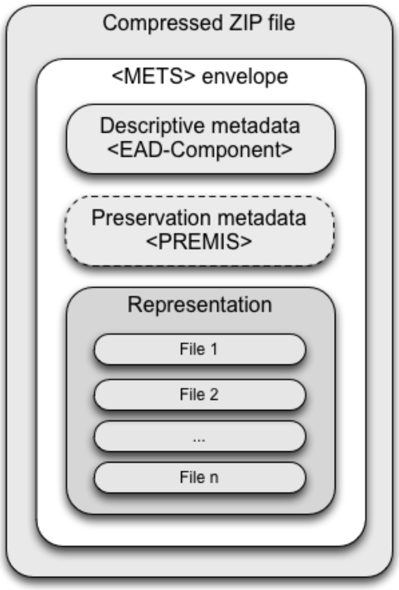

Figure 1: RODA SIP structure.

to create and modify descriptive metadata and define rules for preservation actions, e.g. scheduling integrity checks on stored digital objects or initiate a migration process. Administration procedures allow the definition of access rights to data and operational permissions for each user or group.

Before RODA is able to ingest Information Packages, which in OAIS are called Submission Information Packages (SIP), a formal or informal agreement between the Producer and the Repository must be made in order to specify the contents of the Information Packages (e.g. specifying required sets of information and in which standards they should be encoded) and any timeframe applicable. As RODA has its own SIP format (see Figure 1), anyone who wants to deposit into the repository has to produce RODA compliant SIPs. To accomplish that, one can use the desktop tool (RODA-in) which allows the creation and upload of RODA SIPs, or directly use RODA Web User Interface (RODA-WUI).

Both these approaches have limitations as they don't scale and they use RODA own niche SIP format. This may become a problem when massive creation of SIPs is required and existing systems do not produce RODA's SIP format. A way around it is to develop programs that integrate existing systems with RODA, by producing SIPs in RODA's format, but there are too many systems and not all institutions have the resources to develop their own ad-hoc integrations. If no mandates for having a same way to build and share Information Packages are in place, being them recommendations or legal impositions, integration and sharing of information between systems or entities of the same country is hard and becomes even harder on a broader international context.

After ingest, OAIS recommends that actions for ensuring that the information available in the repository keeps accessible and understandable to its Consumers must be put in place. These actions are defined and monitored by a Preservation Planning process, which can be as simple as file corruption detection using checksums or as complex as file format migration and quality assurance. From a technical point of view, RODA and its Preservation actions (set of plug-ins which can be manually executed or scheduled for later execution) make it easy to perform either of these preservation tasks. But from a management point of view, a well founded decision for selecting optimal preservation task to ensure the continuous access to the information available at the repository must be made.

Preservation Planning is defined as the task responsible for *monitoring the environment of the OAIS and which provides recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, and sufficiently usable by, the Designated Community over the Long Term, even if the original computing environment becomes obsolete* [1]. This process can be done by the repository manager, periodically, in a manual or semi-automated fashion. But as the information in the repository grows and becomes more diverse, manually monitoring all risks that might afflict file formats and plan the proper action to perform can become infeasible. Automation of some of the steps in preservation planning becomes, therefore, crucial to maintain a trustworthy repository and the authenticity of the curated digital objects.

These are just some of the constraints observed on the RODA implementation before the research projects and that were an object of research on the same projects. On the next sections the research project results will be presented and the process of integration of these research results into the open source project will be described.

3. RESEARCH INITIATIVES & RESULTS

In the SCAPE project, RODA was used as a reference implementation by integrating with Scout[5], the preservation monitoring tool, Plato[6], the preservation planning tool, and Taverna[7], a domain independent workflow management system used to run preservation tasks. These integrations allow RODA to enact a preservation lifecycle, that continuously monitors the existence of preservation risks on the content, devises a preservation plan to mitigate them, executes the plan transforming the content, and monitors back again to verify if the problems were solved.

3.1 Scout

Scout, a preservation monitoring tool, supports the scalable preservation planning process by implementing an automated service for collecting and analysing information on the preservation environment [6].

Scout works by configuring *source adapters*, that obtain and normalize information from different sources in order to save that information in a knowledge base. Those sources, as illustrated in Figure 2, can be content (e.g. file-system usage), organization policies, format and tool registries, the Web (e.g. using Natural Language Processing to extract knowledge from websites), and even human knowledge. In conjunction with the information collected, Scout allows the

[5] http://openplanets.github.io/scout

[6] http://www.ifs.tuwien.ac.at/dp/plato/intro

[7] http://www.taverna.org.uk

creation of queries, a mechanism to allow reasoning in the information gathered to detect changes. On top of the queries, triggers (a watch condition) can be created to periodically evaluate them. And when that condition is not met, Scout allows, for example, an e-mail notification. Upon notification additional actions like Preservation Planning can be initiated.

Scout is currently integrated with RODA, and to perform this integration changes had to be made in order to be able to configure Scout to monitor the several aspects of the repository. RODA already exposes APIs to allow integration with others systems, but not always those APIs expose the information required for all purposes. In RODA particular case, we wanted to monitor both the content (i.e. files) as well as repository events (e.g. ingest finished). To make this integration possible, the following features were needed:

- Report API[8]: OAI-PMH [7] interface that exposes repository events information like ingest started, ingest finished, etc.
- FITS plug-in: RODA plug-in responsible for doing characterization on every file stored in RODA using FITS tool[9].

The first one is directly integrable with Scout (using source adaptor for repository Report API), whereas the second one needs an extra tool: C3PO - Clever, Crafty, Content Profiling of Objects[10], which analyses the technical properties of large sets of objects based on metadata generated by characterisation tools such as FITS and Apache Tika[11] and provides aggregated information of those technical properties (e.g. file size, MIME type, compression scheme). The FITS plug-in output is fed into C3PO which in its turn is configured in Scout to be a source of information (using source adapter for C3PO).

By formalizing in Scout a set of conditions that must be met in order to state that no preservations risks exist, the mandatory responsibility of an OAIS compliant repository of ”Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies” [1] is addressed.

3.2 Plato

Plato, a preservation planning tool, implements a well documented and validated preservation planning methodology and integrates registries and services for preservation action and characterisation [2, 10].

Plato provides a Web interface that allows preservation plans to be built interactively, guiding the planner through a well defined decision making process:

1. Definition of high-level requirements and break down to measurable criteria thus creating an objective tree;

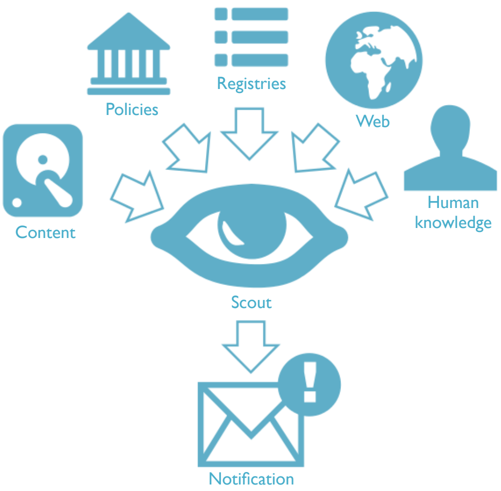

Figure 2: Scout sources diversity.

2. Evaluation of potential preservation strategies by applying selected tools to a manageable sub-set of objects that should cover the essential characteristics of the collection being analysed;
3. Analysis of the results and decision taking of whether any of the strategies should be applied, and if affirmative, an executable preservation plan can also be created.

At the very end of the process of creating a preservation plan, the output is the plan in the form of an XML file. To deploy the plan into RODA, a service named Plan Management API[12] was developed, which allows the creation, retrieval, update and deletion of Preservation Plans from the repository. Alongside with CRUD[13] operations, it also allows search using SRU [14] as the search protocol and CQL [11] as the syntax for representing the queries. Also, for management purposes, this API allows the monitoring of Preservation Plans in the repository (i.e. if they are active, if they are being executed in a certain moment in time, if they executed with success, etc.).

As soon as a plan is deployed into RODA, a unique identifier is associated to that particular plan. This way, when a plan is executed and changes are performed in an intellectual entity, RODA is able to relate each other. It does that by creating a PREMIS [9] event (per intellectual entity) that connects, for preservation purposes, a plan and the representation files of that intellectual entity. This way, when browsing the preservation timeline of a particular intellectual entity in RODA, if any preservation event was created due to preservation actions executed on the context of a preservation plan, this plan can be immediately consulted, describing why the action was executed and detailing all the decision-making process for selecting the exact action that was executed, including the tested alternatives, used samples, experiment results, final decision and execution details.

[8] https://github.com/openplanets/scape-apis
[9] http://projects.iq.harvard.edu/fits
[10] https://github.com/peshkira/c3po
[11] http://tika.apache.org
[12] https://github.com/openplanets/scape-apis
[13] CRUD stands for Create, Retrieve, Update and Delete

Having RODA already the capability of performing preservation tasks (through Preservation actions), adding an external tool with given proofs on building preservation plans that formally describe requirements, analyses alternative solutions to mitigate preservation risks and allows well-founded decision, allows RODA to support even more of the digital preservation processes defined on the OAIS and ISO 16363 [5] as mandatory for digital preservation and repository trustworthiness.

3.3 Taverna

Taverna is a domain independent workflow management system, i.e. a suite of tools used to design and execute scientific workflows [16]. It includes Taverna Engine (used for enacting workflows) that powers both Taverna Workbench (the desktop client application) and Taverna Server (which executes remote workflows). Taverna is also available as a Command Line Tool for faster execution of workflows from a terminal without the overhead of a GUI.

Using Taverna Workbench, one can interactively create workflows. A workflow is defined as "the automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules"[14]. In practice, and in Taverna's particular case, it can be seen as the way to describe, manage, and share complex scientific analyses. This can be achieved by combining several components called Services, either sequentially or in parallel, that can be of several types such as:

- Web services (local or remote in either REST or WSDL format);
- Local scripts (Bash scripts, R scripts);
- Beanshell (Java code snippets);
- Local services (pre-defined Beanshells for specific tasks such as file/XML/text manipulation, database connectivity through JDBC, etc.);
- Sub-workflows.

After finishing the workflow design, it can be run immediately in Taverna Workbench as well as in Taverna Server or Taverna Command-line.

As RODA originally does not provide functionalities for managing and running Preservation Plans (available in the repository), which in this case contain Taverna workflows, changes had to be made. This way, two new APIs were created:

- Data Connector - REST API for manipulating intellectual entities and associated representations in the repository;
- Plan Management - REST API to retrieve available preservation plans from the repository, to manage their state (enable/disable) and the status of a particular execution (being executed, execution successful or execution failed).

Having mechanism to manipulate data in the repository as well as to manage preservation plans is not enough as a mechanism is needed to process the preservation plan and run it in Taverna Suite. For this, a tool called Plan Management Webapp was created. Besides managing preservation plans available at the repository (creation, edition and deletion), it allows to execute a preservation plan.

When executing a preservation plan, the Plan Management Webapp retrieves the entire plan from the repository (as initially only metadata is retrieved for listing purposes), sets the execution status to ”being executed", identifies the objects that must be changed and retrieves them from the repository. Then, it isolates the executable plan (i.e. Taverna workflow), executes it in Taverna Suite providing appropriated objects as input and collects the results. Then, if everything goes as expected and if those results need to be sent back to the repository, it does so by using the Data Connector API. Also, to wrap up, it sets that plan execution status to either ”success" or ”failure".

On the one hand, having two new APIs makes it easier to integrate RODA with third-party tools/systems. On the other hand, having the possibility of running preservation tasks with Taverna workflows (which are tightly connected to a Preservation Plan) increases compliance to ISO 16363 as it better fulfils the Preservation Planning requirements.

A report on the compliance of the system presented above, named SCAPE Preservation and Watch Suite or SCAPE Preservation Environment, that brings together RODA, Scout, Plato and Taverna, with ISO 16363, assessed solely from a software technology perspective (ignoring therefore organizational, financial or physical infrastructure requirements), show that 69 of the requirements are fully supported, 2 are partially supported, 6 are not supported, and 31 are out of scope (ignored) [4]. Almost all of the requirements are supported solely by this software suite, and the rest can be supported by manual procedures, which is a vast improvement from previous versions.

4. FUTURE RESEARCH

RODA is used on a pilot of the E-ARK project, which will develop a total of six different pilots. This full scale pilot will be conducted jointly with the Portuguese Agency for Public Services Reform (AMA)[15] and the Instituto Superior Técnico[16] as RODA will be the long-term archival solution and these two entities the data providers.

One of the goals is to support a pan-European SIP format which will make it easier to create Information Packages to be transferred and ingested into archives in a way that is efficient, reliable and applicable across all European countries. Another goal is to enhance RODA ingest process, to be more flexible and customizable, thus making it easier to integrate with third-party systems without the need of human intervention, making the system more scalable.

The pilot will demonstrate that the pan-European SIP structure designed in E-ARK is adequate to support the con-

[14] Quoted from http://www.taverna.org.uk/introduction/why-use-workflows

[15] http://www.ama.pt

[16] http://tecnico.ulisboa.pt

tent types currently supported by RODA (i.e. relational databases, text documents, video, audio and images) and, provide a framework for automatic SIP creation by Document Management Systems.

This project will also focus on access to the content, specially complex content such as relational databases, finding scalable methods to provide access to archived databases and also providing methods to allow an advanced analysis and reuse of database content by using, for example, data warehousing and OLAP technologies [3].

5. ROUTE TO MARKET

As any open source project, RODA has its source code freely available. Also, as it is a good practice in software development, RODA source code is versioned. RODA uses Git [15] as a versioning system, and publishes its source code on its main repository[17] in GitHub.

When a research project starts, a fork [8] of the main source code repository is created enabling a separated trend of developments to be carried out. Then, in the end of a particular project, all developments made are analyzed in order to decide which ones should be integrated with the next official version. This analysis needs to be done as not all of the developments may be widely applicable and therefore may not be suitable for a broader audience or might have a severe impact on other features previously developed by the community.

All development trends, main and alternative ones created for the research projects, are published in GitHub and available for the community to try and develop upon them. But only the main version is continuously maintained by the core developers and used as a base for new research projects. This ensures that the development work is focused and the project doesn't become too fragmented to be maintained.

On the research initiatives and results presented in section 3, that mainly relate to the SCAPE project, Scout, Plato and Taverna are services external to RODA that integrate with it using 3 APIs and one plug-in. These APIs were developed to be repository system independent, and some of them are implemented for other repository systems[18]. The APIs themselves were developed to follow standards (like OAI-PMH, PREMIS, Dublin Core, METS), to be flexible (i.e. minimal mandatory information) and to have the least possible impact on the underlying data models. All of these characteristics allowed for the APIs to be merged into the main source code and shipped on the next version of RODA. On the plug-in, a software logic that was deemed necessary to be added to RODA itself, the fact that it was implemented as a plug-in, i.e. a modular and contained software component that adds a specific feature, allows it to be easily merged and shipped with the next version.

The future research presented in section 4, that mainly relates to the E-ARK project, describes future developments that have a much deeper impact on RODA data model and business logic. A change of the SIP format might introduce information restrictions or flexibility that can have an impact on how ingest is done and what information can, or must, be kept on the system, introducing changes on the data model. This is an accepted risk on the capability to merge these changes into the main source code, bringing the results of the project to the users. The risk is mitigated by aligning the objectives of the research project with the roadmap of the open source project itself, assuring that the deep changes are profitable for the whole community, and testing the changes with real world cases, keeping in close contact with the target community. The latter is given by the nature of the E-ARK project, that is funded by a competitiveness and innovation framework programme, which shapes the project objectives not in "blue sky" research, but in the creation of a favourable ecosystem and market growth. This is materialized in the E-ARK project by the focus on pilots, that drive and test the developments on real world cases, integrating the systems with reference institutions in the European context, ensuring the alignment with the community and testing in real-world scenarios.

6. CONCLUSIONS

RODA is a complete digital repository that delivers functionality for all the main units of the OAIS reference model. Even so, and as any software that wants to be successful, it needs to be open for further improvements. Those improvements can be triggered by its own community needs as well as for changes in the Digital Preservation community, which need to be continuously monitored in order to keep RODA up-to-date with the best practices of this field. Being RODA based on open source technologies and well established standards such as METS [13], EAD [12] and PREMIS, makes it easier to improve. This is shown by the improvements made in the SCAPE project as well as the improvements that will be made in the E-ARK project. Also, another great advantage of being open source is the fact that these improvements are freely and immediately available.

But some planning and design is needed to ease the effort needed to merge and publish the outputs of research on open source products. Research outputs are many times not production ready, are domain specific, and can have an impact on the platform that break existing functionality. Easy extensibility of the open source application, using e.g. plug-ins, is an important characteristic to enable a fast inclusion of new research outputs into main versions, especially if the features are domain-specific. Also, the use of APIs for integration, that use of standards, are flexible and are designed for least impact on the data model, can be of paramount importance to enable publishing of the features to the end user.

If impact of the platform is unavoidable, identifying the risk in early stages, accepting it exists, and aligning the roadmap of the open source project with the research objectives is important. In these cases, ensuring the developments follow the community interests and keeping a close connection with the community is needed to ensure developments fit the community real world cases.

In software development, open source and even more so in research, change is inevitable and even necessary, but planning, design and communication are very important to keep

[17] https://github.com/keeps/roda

[18] http://wiki.opf-labs.org/display/SP/Repository+APIs

project in the right path and maintain community adoption.

7. ACKNOWLEDGMENTS

This work was partially supported by the E-ARK project. The E-ARK project is co-funded by the European Commission under CIP-ICT-PSP-2013-7 (Grant Agreement number 620998).

8. REFERENCES

[1] Reference Model for an Open Archival Information System (OAIS). Technical report, Consultative Committee for Space Data Systems (CCSDS), 2002.

[2] C. Becker, H. Kulovits, A. Rauber, and H. Hofman. Plato: A service oriented decision support system for preservation planning. In *Proceedings of the 8th ACM IEEE Joint Conference on Digital Libraries (JCDL 2008)*, 2008.

[3] S. Chaudhuri and U. Dayal. An overview of data warehousing and olap technology. *SIGMOD Rec.*, 26(1):65–74, Mar. 1997.

[4] M. Ferreira, L. Faria, M. Hahn, and K. Duretec. Report on compliance validation. Technical Report MS63, SCAPE project, 2014.

[5] ISO. Space Data and Information Transfer Systems—Audit and Certification of Trustworthy Digital Repositories. ISO 16363:2012, International Organization for Standardization, Geneva, Switzerland, 2012.

[6] M. Kraxner, M. Plangg, K. Duretec, C. Becker, and L. Faria. The SCAPE planning and watch suite: supporting the preservation lifecycle in repositories. In *iPRES 2013 - 10th International Conference on Preservation of Digital Objects*, 2013.

[7] C. Lagoze and H. V. de Sompel. The open archives initiative: Building a low-barrier interoperability framework. *Digital Libraries, Joint Conference on*, 0:54–62, 2001.

[8] J. Loeliger and M. McCullough. *Version Control with Git: Powerful tools and techniques for collaborative software development.* " O'Reilly Media, Inc.", 2012.

[9] PREMIS Editorial Committee. Data Dictionary for Preservation Metadata: PREMIS version 2.0. Technical report, Mar. 2008.

[10] S. Strodl, C. Becker, R. Neumayer, and A. Rauber. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL'07)*, pages 29–38, New York, NY, USA, June 18-23 2007. ACM Press.

[11] The Library of Congress. Common Query Language. http://www.loc.gov/standards/sru/cql/ [Online; accessed 21-October-2014].

[12] The Library of Congress. Encoded Archival Description. http://www.loc.gov/ead/ [Online; accessed 21-October-2014].

[13] The Library of Congress. Metadata Encoding & Transmission Standard. http://www.loc.gov/mets/ [Online; accessed 21-October-2014].

[14] The Library of Congress. Search/retrieve via url. http://www.loc.gov/standards/sru/ [Online; accessed 21-October-2014].

[15] L. Torvalds and J. Hamano. Git: Fast version control system. *URL http://git-scm. com*, 2010.

[16] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. Nieva de la Hidalga, M. P. Balcazar Vargas, S. Sufi, and C. Goble. The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. *Nucleic Acids Research*, 41(W1):W557–W561, 2013.