EVALUATION OF THE CALLUNA REPORT ON ALA

PM by Fredrik Ronquist, with contributions by David Martin, Anders Telenius and Markus Skyttner
2016-12-13

Background

In October, Ulf Gärdenfors at Artdatabanken, the Swedish University of Agricultural Sciences (SLU), distributed a report on the Atlas of Living Australia portal (ALA) from Oskar Kindvall at Calluna AB. The report evaluated ALA with respect to functionality, performance and costs of implementation in Sweden. It also compared ALA to the Swedish Analysis Portal (SAP) software and the Swedish Lifewatch infrastructure (SLW). Oskar Kindvall is the former SLW and SAP system architect.

This PM summarizes a discussion of the claims in the Calluna report between Fredrik Ronquist at the Swedish Museum of Natural History (NRM) and David Martin, previously chief architect of ALA and now responsible for the UK implementation of ALA. Markus Skyttner, IT architect in the informatics team at NRM, and Anders Telenius, manager of the Swedish GBIF node, also contributed to this PM.

General Concerns

While it is clear that functionality, performance, and fit to national needs are important factors in evaluating competing platforms for a Swedish biodiversity data portal, other factors are likely to be more important in the long term. They include operation costs, size of the developer community supporting the maintenance and development of the code base, and transparency of the development process. Lock-in effects are also important: to what extent is it possible to change the hosting organization or replace the entire system or parts of it if needed?

The ALA code base uses open-source licensing and runs on free and open-source stacks (FOSS). The development process is completely transparent and is based on best practices in distributed open-source development. The ALA project is backed by a team of ten developers in Australia, with stable funding for the coming decade. The GBIF secretariat in Copenhagen and ALA have recently agreed to align their code bases, putting another handful of developers with stable long-term support behind the ALA code base. Countries running biodiversity data portals based on the ALA software are also contributing to the development of the system, as are many independent developers (so far, more than 70 developers have contributed to the ALA system).

The SAP code base does not run on a FOSS stack. The development process is opaque and not based on standard practices used in distributed open-source development. Unlike ALA, it is not possible for outside teams to analyze the development activity behind the SAP code base, or to install, run and test the performance of the SAP system without assistance from Artdatabanken. As far as we know, no developers outside Artdatabanken are contributing to the development of SAP, and the system is not run

outside of Artdatabanken. The Swedish Research Council has clearly signaled that they will no longer fund continued development of the SAP code base.

In conclusion, a Swedish biodiversity data portal based on ALA appears quite attractive from a long-term sustainability perspective. There would have to be significant problems associated with functionality, performance, fit to national needs, or implementation and migration costs to justify continued Swedish investment in a SAP-based solution.

Quality and performance

The report is making several sweeping claims about the quality, maturity and performance of the SAP and ALA software, which are not substantiated by data. The view expressed in the report that the performance of a system running on "bare metal" is vastly different from that of a system running on a virtual machine (VM) is outdated. Almost all modern systems are run in virtualized environments because of the many benefits of VMs from a system operations perspective. While the ALA software is designed to run on VMs, it can also be run on bare metal if the slight performance gain achieved by this would outweigh the disadvantages of running the system in a non-standard environment. If the ALA performance were really a concern, a fair comparison of the performance of ALA and SAP on real-world tasks should be undertaken.

The Calluna report implies that the ALA software is less mature than SAP. However, the ALA software has been in production in some form since 2010, which predates the first release of SAP by several years. It is running now in more than ten instances in eight different countries. The total runtime of ALA systems across the world widely exceeds the runtime of the single instance of SAP. The number of users and records vary between ALA instances, but some instances handle considerably more users and data than SAP. Thus, there is a large community that can attest to the maturity of the ALA code base.

The report also implies that the ALA software is poorly designed because it does not separate static and dynamic modules, which is simply not true. The report refers to time-outs that occurred in the Australian portal while it was being tested. No system is perfect, but time-outs are generally rare in the Australian instance of ALA, and more details are needed to analyze the causes of these particular problems. One possibility is that the time-outs are linked to the request for computationally demanding plots, as indicated in another part of the Calluna report. This is functionality that is not supported in SAP.

Functionality

We are pleased to see the Calluna report concluding that SAP and ALA are quite similar with respect to functionality. Of the 39 analyzed features, about half are judged to be equivalent and 70 % are addressed in some form in both systems. Seven features are pointed out as missing in ALA; they are mostly minor features related to download formats. On closer examination by us, it turns out that four of them are actually supported in ALA but were simply missed in the study. For instance, ALA *does* support alternative grid systems; this is used in the UK portal but not in the studied Australian instance of ALA. Thus, it should be straightforward to support RT90 and SWEREF 99 TM

in a Swedish ALA portal. The only features that are actually missing in ALA involve support for Excel-format downloads, and support for an extended ontology for occurrences.

ALA supports a number of download formats including csv, which is usually adequate for Excel users besides being more generally suited for data wrangling tasks. If Excel support were deemed important for Swedish users, it would be straightforward to add it to ALA.

ALA supports presence and absence occurrence records but SAP extends this basic ontology by supporting a couple of different types of presences and absences, namely "natural occurrence", "non-spontaneous occurrence", "absence in habitat where the species is expected to occur", and "absence at a location where the species has been found previously". Both ALA and GBIF are interested in supporting such a richer ontology for occurrence records based on an internationally accepted standard, so we can expect support for this to be added to ALA in the not-too-distant future.

The four ALA features that are missing from SAP are apps that cover major functionality areas. They include an R package (ALA4R), support for phylogenetically structured analyses (PhyloLink), crowd-source digitization of museum records (DigiVol), and a mobile app for reporting biodiversity data (BioCollect). Of these, the Calluna report claims that neither the R package nor phylogenetically structured analyses are needed in Sweden. This claim is somewhat surprising to us, since we see an R package as critical functionality, and we are excited about the new possibilities opened up by the support for phylogenetically structured analyses of Swedish biodiversity data. We also note that NRM and SLU, with other partners, applied in May 2015 to the Swedish Research Council for support of a national biodiversity informatics infrastructure ("SeIBER") including both of these features.

The BioCollect app is dismissed in the Calluna report as being "extremely primitive" and "lacking good taxonomy support". While there is certainly much to be said about BioCollect, it is difficult to describe it as primitive. It supports more than 40 different survey types, and has been used to collect data from more than 1,000 M AUD worth of biodiversity projects in Australia. The taxonomy support in ALA is, arguably at least, more sophisticated than that in SAP. While some ALA apps force adherence to a taxonomic backbone, like SAP, other tools allow data owners to upload and analyze data that include taxon names outside the backbone, while flagging the non-matching names as potential data quality issues. In contrast to the SAP solution, this allows users to handle data associated with names that are not in the backbone. This is important for some users, for instance in handling manuscript names (species that are not formally described yet) or names of taxa that are missing in the taxonomic backbone because they occur naturally outside of the focal area of the portal.

Implementing ALA in Sweden

The Calluna report implies that the cost of handling the content behind the Swedish web services, and adapting ALA to fit Swedish needs, is such that a switch from SAP to ALA would be extremely costly. We would like to see a more detailed specification of those costs. We note that the work needed to feed data from Swedish systems like Artportalen and DynTaxa into a Swedish ALA portal is the same work needed to deliver those data to

GBIF. As long as Sweden is a member of GBIF, we need to address these are tasks anyway, and there should be no extra cost involved.

During the construction phase of SLW, Artdatabanken has insisted that data owners deliver their data to SAP using specialized Swedish protocols, and that GBIF fetch the data indirectly through SAP. Regardless of whether SAP or ALA is used for the Swedish biodiversity data portal, this is not the optimal infrastructure design in our opinion. Data should be fetched directly from the data owners using internationally established protocols, if at all possible. The Swedish Environmental Protection Agency could help by making this clear to Swedish data providers funded by them, as some data providers now are confused by conflicting requests from the Swedish GBIF node and Artdatabanken.

According to our analysis, the adaptations of ALA needed to run the system successfully in Sweden are minor. The Swedish GBIF node started serious work on a Swedish ALA portal in August, and we expect to have a full ALA instance in operation with Swedish data by the end of 2017. We still think this plan is realistic using a resource allocation of one system administrator and one extra data manager that we will bring in during 2017. This is possible thanks to the support from ALA and the GBIF secretariat, and the assistance of Swedish data providers.

During the fall of 2016, we have worked with ALA to put the core of the system into Docker containers. Docker is cutting-edge technology for systems integration and deployment, the adoption of which will further simplify the task of installing and running ALA portals. The “dockerization” will continue with the remaining ALA modules in the spring of 2017. At NRM, we are pleased to feed this work back into the ALA collaboration as a Swedish contribution towards the ALA mission.

We also note that it has been possible with reasonable efforts to adapt ALA to the national needs in seven countries outside of Australia so far. Among other things, this work is facilitated by the support in the ALA infrastructure for translation of the portal to new languages. Of course, it is possible that we have missed some special difficulties associated with the implementation of ALA in Sweden, so it would definitely be valuable with a more detailed specification of the problems alluded to in the Calluna report.

A switch from SAP to ALA in Sweden will clearly require additional user training. However, there is rich online material documenting ALA tools in multiple languages, facilitating the adoption by Swedish users. Whether or not Sweden decides to run an ALA-based portal, ALA skills will be essential for Swedish biologists studying the flora and fauna of other countries. Popular study areas for Swedish biologists include Australia, Spain, the UK, Costa Rica and Brazil, all of which are running or in the process of deploying ALA portals.

The Informatics Team at NRM

Even though we have become accustomed to a somewhat different tone in our conversations with Artdatabanken than with other partners, we are still surprised by the attack on our group in the Calluna report. With respect to the claim that NRM has failed to live up to SLW expectations, we would like to point out that we were the first group outside of Artdatabanken to set up an approved data delivery service according to SLW requirements. What happened after that appears to be the result of poor communication between the teams, something that is difficult to blame on one part only.

When analyzing the success of the DINA project, it should be born in mind that the complexity of this project is similar to that of the Artportalen 2.0 endeavor. The aim is to build a modern, open-source web-based collection management system (CMS) supported by a large developer community and being used at many institutions around the world, much like ALA. Unlike ALA, we have not had access to substantial startup funds (around 300 M SEK for ALA); we are building the initial code base through international collaboration instead. It is true that this has turned out to be more complicated than initially anticipated. However, in view of the recent interest in DINA from some of the largest natural history museums in the world, including museums with both commercial CMSs and CMSs developed in-house, it would seem somewhat premature to call the project a failure. Those interested in learning more about the DINA project are welcome to attend the symposium "Open Source Systems in the Public Sector", arranged by NRM on January 23, 2017, where our partners at Agriculture and Agri-Food Canada give their view on the DINA collaboration.

Conclusion

The Calluna report concludes that many of the smart ALA solutions ought to be used as inspiration in the further development of SAP. However, this means spending limited Swedish resources on further development of a system that, as far as we can judge, will only be used in Sweden, a system that can only with difficulty be hosted by organizations other than Artdatabanken, a system that is not supported by developers outside Artdatabanken, and a system with operation and maintenance costs that appear to be more than twice as high as those of ALA (based on Artdatabanken estimates of SLW costs from January this year). To us, at least, it appears to be a better choice for Sweden to instead work with the rest of the world within the ALA collaboration, contributing to and benefiting from the efforts of the growing global community backing this platform.