Title: The Evolution of the Community Module: How lessons learned from national, regional, and subject-focused use cases have been used to support inter-institutional collaboration

Type: Business/technical paper

Authors: James Toon[1]

1. Elsevier, Amsterdam, Netherlands

Keywords

Research Information Management Systems (RIM, RIMS, CRIS), Community Repositories, Aggregation, Extract, Transfer and Load (ETL), Data Deduplication, Harvesting, Web Services, Pure, Use Cases, Lessons Learned.

Summary

This abstract provides an update on development and implementation progress of the Pure community module in the 5 years since the initial launch of the service. The paper revisits the conclusions of our 2018 presentation (Toon et al., 2018), providing updates on the 'future developments' and encourages a discussion of the complexities in supporting multi-institutional collaborations. In addition, we will provide an overview of some key challenge areas we have been working on together with community customers and community owners (technical and operational).

Background

The landscape of research funding is changing. Stagnating or declining grant success rates mean that increasing effort is spent securing grants from across a smaller pool of available funds, with the greatest likelihood of success coming from those institutions who have a higher likelihood of being able to guarantee results (Alberts et al., 2014). Changes to the concentration of funding have resulted in an opportunity to develop critical mass in research (Heathwaite, 2019) but which leads to a less subject diversity. These economies of scale in research funding may also be challenging for some institutions who are less well placed than others for success within this landscape, with the concentration of funding resulting in a type of 'Matthew effect' (Bol et al., 2018)

It thus becomes increasingly evident that to be able to compete effectively requires a community effort to maximise success, leading to the need for institutions to work together as a means to pool capability when making strategic decisions. We are moving away from inter-institutional *competition*, toward inter-institutional *collaboration*.

This collaborative, community-based approach is also important in addressing use cases that support the onward dispersal as well as concentration of research funds via inclusive workflows, particularly when discussing national/regional requirements. Communities additionally support showcasing and 'expert finder' functionalities via robust reporting. And through data re-use, they facilitate a reduction to the administrative burden of research by promoting an enter once, use many times policy.

The Community Module is a service offering that provides a technological framework to satisfy the needs of inter-institutional collaboration.

The Pure Community Solution

The Pure Community Module facilitates the management and reporting of multi-institutional research projects. It provides a robust extract, transfer and load process (ETL) together with a multi-institution portal that makes it easy to collate and aggregate the data, showcase the research assets and understand the contributions of each participating institution**.**

There has been significant growth in the use and interest in community-based solutions since the launch of the service in 2017, covering a variety of customer needs. Pure communities are present in 7 countries, comprising 33 participating Pure instances (and 6 aggregating community modules). A further 124 institutions are participating in multi-tenant Pure communities. In our presentation we will provide a more detailed overview of the use cases for the type categories described below, including a specific customer case study focusing on the academic shared services model.

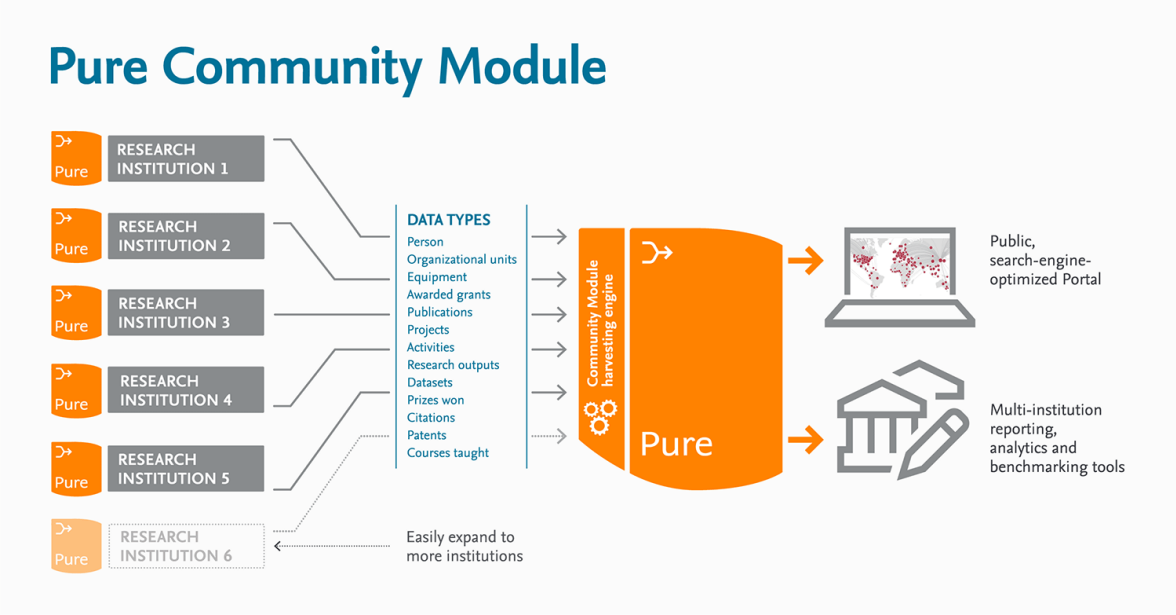

Figure 1 - High level overview of Pure Community Architecture

Categorisation of user needs

The core architecture described in figure 1 is largely consistent across all community service users, however the functionality and needs of users are not always the same. This is a highly crucial factor in both the development of features and in the operational approaches taken.

We have categorised the Pure communities into different sub-types based on their underlying objectives.

- **Type A** – Multiple institutions with private CRIS/RIMS, combined with an aggregating, shared central Community instance, allowing comparative evaluation and showcasing of community expertise via the community portal

- **Type B** - National performance management via a transparent, inclusive digital workflow, or in support of a transparent, inclusive national evaluation via data collection
- **Type C** – State/regional infrastructure to develop economic & collaboration impact, demonstrate expertise to foster collaboration & economic development, and Identify expertise within a region

Lessons learned

The solution has developed since our prior presentation to the EuroCRIS community. We have gained significant insights and lessons learned into the challenges faced when managing the technical infrastructures that support the extract, transfer and load processes (ETL) required, and in managing the creation, deduplication and modification of community content. Our presentation will provide a more detailed overview of these lessons, and a summary of how the Pure community service has been adapted in response.

Our experience to date suggests that it is helpful to separate these insights into the following technical and non-technical categories.

Technical Examples	**Summary**
Handling of large data volumes	Data volumes in an aggregate community system are considerable compared to individual institutional systems. This has a direct impact on the performance and sensitivity of processes required to maintain the transfer of data between local RIM systems and the community instance.
Vendor agnostic support for the community	The aggregation of data into the community is required to be vendor agnostic as much as possible. This can be realized through use of XML integrations that work via the Pure synchronisation framework
Identification and deduplication of content from multiple Pure instances	The aggregation of content sources creates a challenge in the identification, validation and merging processes across all content. A parameter driven approach to deduplication is proven, however the process needs to be respectful to local policies
Data quality management within the aggregation instance	Data must meet the needs of the community use cases to provide maximum value. The community service needs to consider the extent to which data can be curated within the aggregation instance for the overall betterment of the community.
Completeness of data	Within individual records, data extracted from institutions may vary considerably. Where the community use case demands a certain level of coverage/completeness of records this may have to be managed carefully.
Shared classifications and taxonomy within the aggregation instance	Multiple participating institutions can develop and implement their own approaches to managing taxonomy supporting their own research information. Where the community supports an aggregate use case, support is required to improve how such classifications and taxonomy are normalized into the aggregation instance.
Context specific filtering for use in showcasing	Community organisational structures are unlikely to be a direct representation of the combined organisational structures of the participating instances. This requires a means to make available

	alternative structures and/or a subset of the overall community dataset for purposes of showcasing

Table 1 - Examples of technical insights gained from communities

Non-Technical Examples	Summary of Issue
Community and participant stakeholders may have different objectives	A crucial issue is to recognize that aggregation community stakeholders may have different objectives to the community participants. Expectations may differ in key areas and need to be managed carefully, for example in different metadata policies or the criteria for data quality.
Required quality, focus and breadth of scope for community data	Data retrieved from participating communities includes a 'core' set that can be extended to include additional information in support of additional use cases, such as for knowledge exchange and impact. A comprehensive picture of research activity requires a joint agreement on collection policies from participating institutions to ensure a complete outcome
Agreements on standardisation and common formats	Shared standards in taxonomy and classification significantly reduces complication during the ETL process. This requires a joint agreement approach and appropriate data governance across all participating community members
Legal/Security issues across institutions/boundaries	Controls over data ETL may differ based on a geographic case and depend on different support/infrastructural models.
Establishing recommendations for governance	Communities require a more comprehensive and hands on support/consultancy model

Table 2- Examples of Non-technical insights gained from communities

Future directions

We continue to adapt and develop the Pure Community solution according to our customer needs, and to ensure a cost effective and value driven approach. Our roadmaps are an indication of the direction of travel and based on the assumption that we will need to continually revisit and adapt. Working together with customers, account and product teams we have established three areas of investment.

New feature development in support of community needs

- Adapting the community solution to accommodate developments around data quality management within local participating instances. For example, recent improvements in Pure to support better representation of Pure entities (for example, the addition of hierarchies for external organisations).
- The development of feedback loops, supporting the central curation of data within the community and the option to re-distribute information back to participating instances
- Extending support for non-Pure participation – ensuring that community participants can be system agnostic
- Working with third parties to secure data in support of the community goals – for example, for the inclusion of data related to research funding, or for metrics data that can be used in conjunction with data feedback loops to the benefit of the whole community

- Frameworks for community reporting/funder outcomes management and/or assessment

Continuous development of reliant infrastructures around data quality

- Pure has a well-established approach to data quality known as the 5Cs of data quality – (Complete, Correct, Connected, Current and Compliant). Working together with the custodians of this model, the product team intend to extend and enhance these concepts around data quality in the context of the aggregation, supporting trust in data at the community level.
- Implementation of processes to support optimal and performant architecture for community implementations, such as continual improvements in the underlying codebase, logic used in support of data deduplication and use of novel approaches to ETL.
- Value in high quality data diagnostics - a key objective for the community solution is to build transparency into the data transfer process so that it is possible to interrogate the system to trace the route and progress of any given piece of content from end to end in its process journey.

Consultancy/Support operations for communities

- The governance and organisation of communities are crucial to their success. Our implementation and product teams work closely with customers to support and guide where appropriate. We have learned that community support requires a more involved and hands on approach and we plan to develop a defined model for community engagement.

Conclusion

The Pure Community solution is now a well-established approach to supporting the inter-institutional use case of CRIS/RIM systems on a national, regional or subject basis. This presentation aims to provide an update for the EuroCRIS community on the approaches taken, highlight the common use cases encountered, provide insights into the lessons learned since the service was established and to highlight a number of key areas of development as we evolve the system further.

We conclude that the future is bright for such collaborative, inter-institutional solutions. Our approach to the development of the service will continue to focus on the changing needs of community in order to deliver the best possible value for the investment made.

References

1. Alberts, B., Kirschner, M.W., Tilghman, S., Varmus, H., 2014. Rescuing US biomedical research from its systemic flaws. Proceedings of the National Academy of Sciences 111, 5773–5777. https://doi.org/10.1073/pnas.1404402111
2. Bol, T., Vaan, M. de, Rijt, A. van de, 2018. The Matthew effect in science funding. Proceedings of the National Academy of Sciences 115, 4887–4890. https://doi.org/10.1073/PNAS.1719557115
3. Heathwaite, L., 2019. Independent review of SFC's research pooling initiative.
4. Toon, J., Kujath, A., Khazzam, E., 2018. Developing a model approach for community led CRIS aggregations.