Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

13th International Conference on Current Research Information Systems, CRIS2016,
9-11 June 2016, Scotland, UK

Data Management Administration Online (DMAOnline)

Masud Khokhar[a,*], Hardy Schwamm[a], John Krug[a], Adrian Albin-Clark[a]

[a] *The Library, Lancaster University, Lancaster, LA1 4YH, United Kingdom*

Abstract

In the uncertain Higher Education environment today, where value for money and financial rigour is more important than ever before, it is vital that institutions create and sustain services that exhibit evidence of impact and provide value for money. In the last two years, external pressures from UK funding councils on complying with their Research Data Management (RDM) policies has caused institutions to develop services and support models in an urgency. These services are usually created for a fixed period, often with a short term investment in staff and/or infrastructure, and primarily because of the lack of clarity in the resultant value for money at an early stage.

Monitoring compliance with funding council requirements is complex. Many institutions use Current Research Information Systems (CRIS) to handle their publication and research data catalogues. However, these systems provide only a basic level of functionality for RDM (e.g. submission of datasets information and linking it with project and publications information). Compliance reporting is not provided out of the box and essential information is usually kept in additional systems or spreadsheets by institutions (e.g. whether a data access statement exists or not). This makes the whole process of RDM compliance monitoring cumbersome and time consuming.

We introduce Data Management Administration Online (DMAOnline)[1], a Jisc Research Data Spring[2] project, which facilitates a novel metric based analysis of an institution's compliance with RDM mandates. DMAOnline brings together key RDM information from a variety of sources and provides a normalised structure for the underlying data. This enables ingest of data from a variety of sources e.g. CRIS, Institutional Repositories or Excel sheets. Currently, DMAOnline has the capability to harvest its information from Elsevier's Pure CRIS and Excel files. It also allows users to add in additional information not available from these sources. A powerful dashboard is created for the user that provides information on compliance with RDM policies, data storage usage, data management plans, DOIs minted, datasets preserved, and basic costing. Other systems that DMAOnline already does or intends to harvest information from include DMPOnline[3], Archivematica[4], DataCite[5], and IRUS-data UK[6].

Keywords:
Research Data Management, RDM, CRIS, Analytics, Metrics, Compliance, Jisc

E-mail address: masud.khokhar@lancaster.ac.uk

[1] www.dmao.info
[2] https://www.jisc.ac.uk/rd/projects/research-data-spring
[3] https://dmponline.dcc.ac.uk
[4] https://www.archivematica.org/en/
[5] https://www.datacite.org
[6] https://www.jisc.ac.uk/rd/projects/research-data-metrics-for-usage
[*] Corresponding author. Tel.: +44-(0)1524-594-236.

1. Introduction

In this paper, we introduce DMAOnline, an online aggregator of data from various RDM systems to provide key compliance and business case development information. DMAOnline allow relevant stakeholders to view the state of RDM at their institution quickly, including compliance statistics with funding councils as well as further evidence to develop new and ongoing, data-driven, business case(s). It provides a visually appealing and functionally powerful dashboard that provides this information in an easily digestible fashion. This allows institutions to get key engagement from senior leadership, a problem which is still being faced by several institutions of all sizes.

During the course of development of Lancaster University's RDM services, we were surprised to hear similar questions arising from many institutions. At a senior level, the leadership want to ensure compliance, minimise risk, and insist on value for money. At a service level, managers want to provide reporting on compliance, develop business cases for senior leadership, and construct a cost effective support model. At a technical level, IT managers are interested in forecasting storage needs and cost, handling of sensitive data and interoperability between systems. All of this information is kept in disparate non-related systems and institutions find it really difficult to provide a complete, near real-time picture of their RDM activities.

For this purpose, we decided to develop DMAOnline, to cover the needs of these diverse stakeholders by bringing research data information together from a variety of sources and systems. It provides relevant data to the right set of stakeholders, providing them with evidence on compliance monitoring, key intervention points to enhance compliance, new and on-going business case development (e.g. procurement of storage infrastructure), and for support services that can be provided at the point of need.

2. Use cases

DMAOnline resolves real issues that institutions face when dealing with the complex, changing and expanding world of RDM. To identify these issues, we gathered feedback from institutions[4], looked at good practices[1 2] and dissected funder policies to identify critical compliance use cases. During our initial work, these use cases were refined further and new use cases were added for business case development and data preservation. These use cases are briefly described below and are categorised under Compliance, Business Case Development, and Preservation. For obvious reasons, most of the work done so far is concentrated on compliance.

2.1. Compliance use cases

2.1.1. Understanding of research data production

It is important that institutions and its constituents (e.g. faculties, schools, departments) can understand their research data production to assess compliance and determine advocacy needs. This applies to the research data that underpins publications and is immutable in its nature.

2.1.2. Data Management Plans production

In the UK, some institutions and most funding councils have mandated the production of a Data Management Plan (DMP) at grant application stage[2]. It is important to keep track of how many projects have been awarded to an institution in comparison to the number of DMPs that were written against these projects, and to intervene to provide support and guidance where a DMP is missing.

2.1.3. Data access statements and DOIs

One of the key requirements for EPSRC RDM policy[3] compliance is citation of data via published research using persistent links such as DOIs. This is usually accomplished by adding a short data access statement in the research publication. Determination of the presence of this statement is a very challenging task and records are usually kept in Excel sheets manually outside of research systems. There is a need to add this information in the same place as the rest of the RDM information to view a complete picture of RDM compliance.

2.2. Business case development

2.2.1. Infrastructure needs and costing model

In addition to staff costs, RDM also brings with it infrastructure related costs, often for procurement of storage (disk space). Without a clear understanding of the current storage requirements of research data, it is difficult to understand how storage needs to grow in the upcoming years. For procurement, recovery of unused disk space , and recovery of costs from funding councils, a basic costing model with the usage statistics of disk space against projects is of high importance.

2.3. Preservation

2.3.1. Archival activities

Research data needs to be preserved for a long time[7] for the outputs to remain usable and valuable. There is a strong requirement to understand how many research datasets are preserved (Archival Information Packages (AIPs)), their storage usage, the file formats and the number of Dissemination Information Packages (DIPs) have been created. This use case fulfils a key requirement of Filling the Digital Preservation Gap[7] project for Archivematica and is developed by Universities of Lancaster and York in conjunction with Artefactual.

3. Key features and technical architecture

3.1. Dashboard

The premise behind DMAOnline is to bring RDM related information from a variety of disparate systems and spreadsheets together to inform the institutional RDM practices. We take a dashboard approach towards providing key information at the top level in the form of tiles (Fig. 1.) each of which can further be expanded for detailed views. We believe that this provides an easy and visually appealing way to present RDM data for stakeholders who don't engage with RDM activities on a day-to-day basis. One important reason for using the dashboard approach is to provide integrated flexibility and capability to view and update relevant information with ease. We want users to engage with a single system to keep RDM information aggregated together. This is done intelligently so that information that is ingested from source systems is kept read-only and information that can only be added manually is updated straight from the interface. One example of this is the presence of data access statements, which is an important piece of information for compliance, but cannot be automated or ingested from a source system like Pure. Another example is whether a DMP has been reviewed or not. This is important for institutions like Lancaster who provide DMP review services. The dashboard provides a very simple integrated way to update this information.

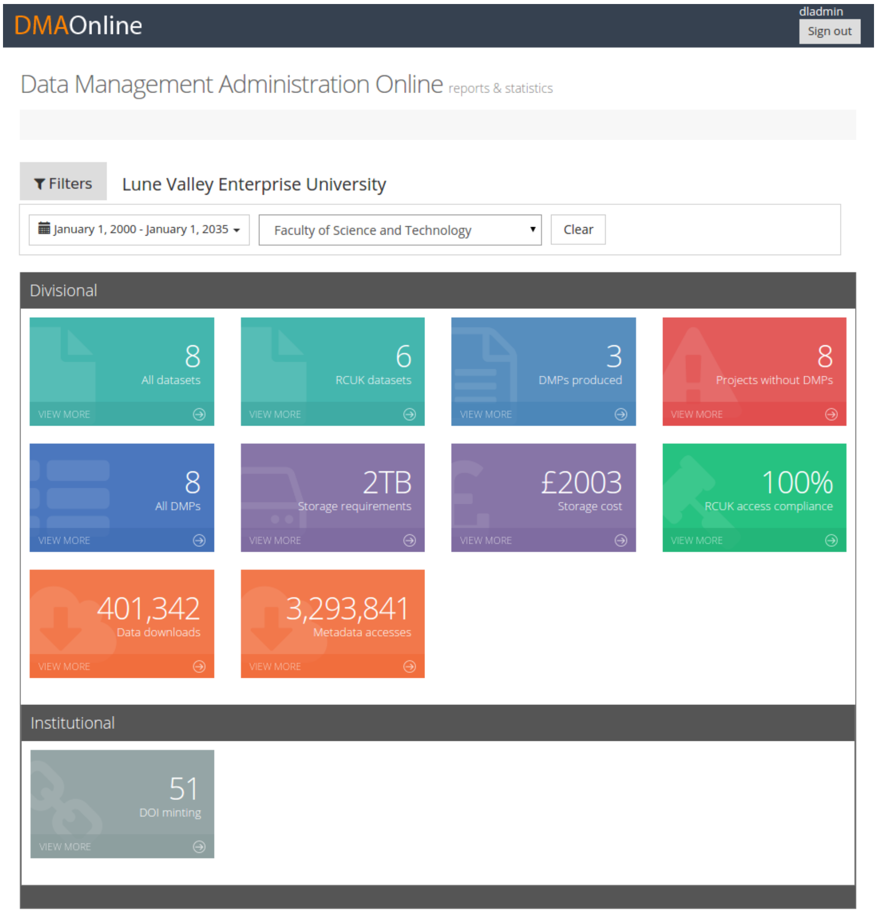
3.2. Source systems and normalisation

DMAOnline is built from the ground up to have the capability to ingest data from a variety of source systems. By building a system agnostic source schema, we concentrate on what is needed for RDM compliance and business case development from an end-user point of view. The schema design is flexible enough to accommodate new fields and rigid enough to provide key compliance information out of the box. Because of time constraints of the project, we have concentrated our efforts initially on automating ingestion of RDM information from Pure, a CRIS from Elsevier. This still equates to a substantial impact as 28 HE institutions currently use Pure in the UK (over 200 in the world[8]) with several others coming on board. We have also been in touch with University of London Computing Centre (ULCC)[9] who are working on integrating DMAOnline with EPrints hosted services. We believe that interoperability is crucial to the success of any project that tries to work with more than one source system, including for DMAOnline,

[7] This is considered to be at least 10 years and usually around 25 years

[8] https://www.elsevier.com/solutions/pure/who-uses-pure/clients

[9] http://ulcc.ac.uk



Lune Valley Enterprise University

January 1, 2000 - January 1, 2035

Faculty of Science and Technology

Fig. 1. DMAOnline institutional dashboard

and we have been and will continue to work with Elsevier to get this right for Pure. The normalised schema, tables and their relationships, can all be viewed here[10].

3.3. Technology considerations and architecture

A lot of attention has been paid to the technical architecture of DMAOnline, keeping in mind the scalability and sustainability aspects of the product. We completely separated the front-end (FE) and the back-end (BE) layers to allow a wider adoption of DMAOnline by institutions, either as a full stand-alone product or as an integrated product in their existing Business Intelligence (BI) products. All communication between the FE and BE is done via RESTful API calls. This allows consumers of DMAOnline to either use the complete solution (with both FE and BE components) or just the BE component using their own institutional FE solution (such as Tableau). This provides powerful flexibility to the institutions and expands the potential of DMAOnline to be adopted by a wide variety of institutions.

3.3.1. Back-end layer

For the BE technology, we have chosen open source and tried-and-tested technologies. For the database layer, we use PostgreSQL. For the RESTful APIs, we use OpenResty. OpenResty is a full-fledged web application server which

[10] http://dmao.info/assets/dbschema/ss.out/index.html

extends Nginx core by adding 3rd part Nginx modules and their dependencies. One of these modules is Lua scripting language which is what we use to develop the database integration and query construction interface between the web application server and the PostgreSQL database layer. For data ingest code, we use Python. Python is used primarily for its powerful XML processing capabilities for the data exported from various source systems. E.g. Pure dataset APIs only export their data in XML format currently.

3.3.2. Front-end layer

For the FE technology, we have chosen a mixture of commercial and open source technologies. The primary reason for this was to avoid re-inventing the wheel and focus on development of key functionality rather than visual aspects of the dashboard. We use the commercial Metronic responsive admin dashboard template[11] which provides the visual features out of the box for the DMAOnline dashboard. It also provides base level AngularJS bindings, which is a technology we use to extend and further develop the FE for DMAOnline. One example of this is integrating Angular UI Grid, a core component of AngularUI suite into the FE for a flexible and detailed view of tiles. This allows users of DMAOnline to easily select columns to display, perform column sorting and searching, in place editing of columns, define selectable rows to create sets and export them in CSV or PDF for simple reporting and analysis elsewhere. We have also developed an API service in JavaScript which abstracts all the necessary BE API calls into a single location for general use within the dashboard application. We also ensure that all communication with the back-end including data updates is done using the standard DMAOnline API calls, thus maintaining the de-coupling of front and back-ends. Further developments were done to allow for filtering of displayed results by faculty, department and date ranges. The filters use two way data binding provided by AngularJS for a dynamic update of displayed data when a filter is selected. We invite the readers of this report to also see our blog post on Technical choices for DMAOnline[6] for further details.

3.4. Multi-institution capability, scalability and performance

From the beginning of DMAOnline, we have envisaged that it has the potential to become a shared national service for institutions tackling RDM. For this reason, the normalised schema design incorporated institution specific identifiers to allow separation of data at an institution level. The front-end was extended to allow multiple institutions to login with their own credentials, and the API design requires an institution specific API key to get further information. We anticipate that we can develop this functionality further by incorporating SAML compatibility, allowing institutions to use their Shibboleth or OpenAthens credentials to login to DMAOnline rather than a separate set of credentials. To provide proof-of-concept for this multi-institution capability, we have been working with Universities of Birmingham, St. Andrews and York to be our early adopters and test the system for us. We are currently going through data sharing agreements to harvest Pure data from these institutions. It is however crucial to mention that the system can only work with the quality of data that we have in our institutional systems. If the quality of data and the internal relationships are not developed and maintained, the output of DMAOnline will neither be comprehensive nor accurate. In terms of performance, the APIs are throttled at an institution level and allows paging capabilities. The Angular UI Grid component can handle ten of thousands of rows without any performance issues, reducing the need for paging as research data grows in the future per institution. Nginx is an asynchronous server, which provides massive performance and scalability advantages over process based servers. As an example, handling 10,000 simultaneous connections would require Nginx to consume only a few MBs of RAM[12].

4. Limitations and intended future work

In a short span of time, we have been able to take a concept of 'all your RDM information available in a single place' to a reality with DMAOnline as a near-production ready product. In our view, DMAOnline is the only dashboard solution that aggregates data from various sources to provide a complete picture of RDM readiness for an

[11] http://themeforest.net/item/metronic-responsive-admin-dashboard-template/4021469

[12] https://www.nginx.com/resources/wiki/community/faq/

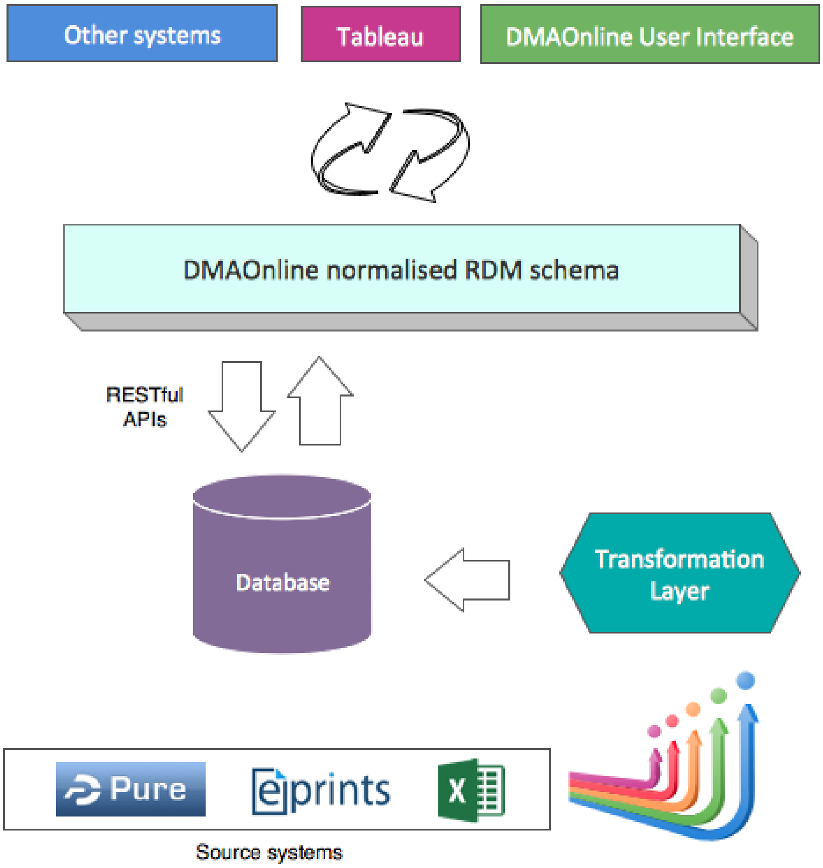


Fig. 2. DMAOnline's modular architecture

institution. With a project of this nature, Interoperability between systems is critical. A lot of effort has been gone into understanding and improving the web services for Pure to incorporate all data fields that are necessary for DMAOnline to function properly. Further work is needed on Pure web services as well as APIs for many other services which is currently being conducted. We have also been working with the international Pure community to help provide design guidelines back to Elsevier for their API developments. We have received very positive feedback so far from the UK, Finnish, German, Dutch, Danish and Belgian user groups in this regard and we will continue working with them and Elsevier to further improve Pure APIs.

In addition to Pure APIs, we have been working with University of York and Artefactual on development of Archivematica APIs that would support DMAOnline. We are in discussions with DCC to get DMPOnline APIs developed and with Jisc to discuss IRUS-data UK API developments. Vendor dependency also slows down development and we have learnt to get them on board from the beginning. However, while the response has been slow, it has been generally positive about all the developments. The area of further developments are highlighted in the table below along with how they support the RDM community.

Table 1. Table 1. Intended future integrations for DMAOnline

Priority	*Current state*	*Benefit to RDM community*
Enhancement of Pure dataset APIs	Issues have been identified and raised with Elsevier. Further work on testing and completeness of APIs for datasets is required.	Pure dataset information is needed to provide dataset and associated publication/project data to DMAOnline. This forms the core of our normalised schema and is used to provide key compliance information.
Integration with DMPOnline APIs	API design documentation is currently being developed. We will be working with DCC to get the new API developed and integrated	DMPOnline is the de facto tool used for data management planning. Integration here will provide key compliance and business case development data.
Integration with Archivematica APIs	Universities of York and Lancaster, along with Artefactual are working together to develop the relevant APIs for Archivematica which will allow harvesting of key information	The information provided to RDM community will include datasets preserved, datasets file formats, datasets archival storage usage amongst other key metrics.
Integration with IRUS-data UK APIs	IRUS-data UK will provide download statistics for research datasets. The API is currently exists in a basic state and further development is needed.	This development will allow RDM professionals to view download data for their datasets straight within DMAOnline.

5. Conclusions and future directions

As research data management becomes a core service amongst institutions, value for money and monitoring compliance become important. DMAOnline offers a practical solution for reporting RDM compliance for institutions. In fact, currently it is the only solution that offers reporting on existing data as well as addition of custom data for further reporting. It plays a crucial role in Lancaster University's RDM infrastructure[8] and our work so far strongly suggests that DMAOnline can become the de-facto RDM reporting tool for Pure customers across the UK and beyond that. We now want to move DMAOnline to a national level service in the UK, get additional Pure customers on board, and add additional source systems.

We are also investigating the long-term sustainability of DMAOnline, especially in the light of the upcoming Jisc Research Data Shared Service. Last but not least, we are eager to work with institutions who may benefit from DMAOnline. If you are interested in working with us, please get in touch with us at rdm@lancaster.ac.uk

Acknowledgements

The Authors would like to thank Jisc for their continued support for the project under Research Data Spring funding reference 3993. The authors would also like to thank Rachel MacGregor (Digital Archivist, Lancaster University), Stephen Robinson (Ruby Developer, Lancaster University), Sarah Jones (DCC) and the early adopters of DMAOnline, especially Karen Clews (University of Birmingham), Jenny Mitcham (University of York) and Anna Clements (University of St. Andrews). This work was derived from Phase I[4] and Phase II[5] reports of DMAOnline project.

References

1. Pryor Graham, Jones Sarah, Whyte Angus. *The Facet Scholarly Communication Collection: Delivering Research Data Management Services: Fundamentals of Good Practice*. Facet Publishing. 2013.
2. Corti Louise. *Managing and Sharing Research Data*. Sage publications. 2014.
3. EPSRC. *Clarifications of EPSRC expectations on research data management*. 9th October 2014. Available at: `https://www.epsrc.ac.uk/files/aboutus/standards/clarificationsofexpectationsresearchdatamanagement/`
4. Khokhar Masud, Schwamm Hardy, Krug John, Albin-Clark Adrian. *Data Management Administration Online (DMAOnline) Phase I report*. version 1.2. figshare. July 2015. Available at: `http://dx.doi.org/10.6084/m9.figshare.1482039`
5. Khokhar Masud, Schwamm Hardy, Krug John. *DMAOnline - Phase two report.* figshare. 2015. Available at: `https://dx.doi.org/10.6084/m9.figshare.2007597.v1` Retrieved: May 20, 2016 (GMT)
6. Krug John, Albin-Clark Adrian. *Technical choices for DMAOnline*. November 2015. Available at: `http://dmao.info/blog/2015/11/13/Technical-Choices-DMAOnline.html`
7. Mitcham Jenny, Awre Chris, Allinson Julie, Green Richard, Wilson Simon. *Filling the Digital Preservation Gap. A Jisc Research Data Spring project. Phase One report*, July 2015. figshare. Available at: `http://dx.doi.org/10.6084/m9.figshare.1481170`
8. Khokhar Masud. *Lancaster University Research Data Management Architecture*. figshare. 2016. Available at: `https://dx.doi.org/10.6084/m9.figshare.3383131.v4` Retrieved: May 19, 2016 (GMT)